Towards Optimal Offline Reinforcement Learning

Mengmeng Li MENGMENG.LI@EPFL.CH
Daniel Kuhn DANIEL.KUHN@EPFL.CH
Risk Analytics and Optimization Chair
EPFL, Switzerland

Tobias Sutter TOBIAS.SUTTER@UNI-KONSTANZ.DE
Department of Computer Science
University of Konstanz, Germany

Abstract

We study offline reinforcement learning problems with a long-run average reward objective. The state-action pairs generated by any fixed behavioral policy thus follow a Markov chain, and the *empirical* state-action-next-state distribution satisfies a large deviations principle. We use the rate function of this large deviations principle to construct an uncertainty set for the unknown *true* state-action-next-state distribution. We also construct a distribution shift transformation that maps any distribution in this uncertainty set to a state-action-next-state distribution of the Markov chain generated by a fixed evaluation policy, which may differ from the unknown behavioral policy. We prove that the worst-case average reward of the evaluation policy with respect to all distributions in the shifted uncertainty set provides, in a rigorous statistical sense, the least conservative estimator for the average reward under the unknown true distribution. This guarantee is available even if one has only access to one single trajectory of serially correlated state-action pairs. The emerging robust optimization problem can be viewed as a robust Markov decision process with a non-rectangular uncertainty set. We adapt an efficient policy gradient algorithm to solve this problem. Numerical experiments show that our methods compare favorably against state-of-the-art methods.

Keywords: Offline Reinforcement Learning, Off-Policy Evaluation, Large Deviations Theory, Markov Decision Processes, Distributionally Robust Optimization

1 Introduction

Recent advances in reinforcement learning have led to remarkable performance improvements in sequential decision-making across various domains, including strategic gameplay (Silver et al., 2016; OpenAI et al., 2019), robotic control (Andrychowicz et al., 2020), autonomous teaching (Mandel et al., 2014), or online recommendation (Liu et al., 2018a) among others. Reinforcement learning is particularly successful whenever online data can be acquired through repeated, low-cost interactions with the environment (Sutton and Barto, 2018; Bertsekas, 2023). In many applications, however, continuous and/or inexpensive interaction with the system is not feasible, limiting the applicability of traditional reinforcement learning methods. In such cases, one must rely on *offline reinforcement learning*, which learns an optimal policy from pre-collected data without the opportunity for further exploration (Levine et al., 2020). Offline reinforcement learning is attractive when active experimentation is prohibitively costly or unethical. It is widely used, for instance, in education (Mandel et al., 2014), healthcare (Oberst and Sontag, 2019) or marketing (Gottesman et al., 2019).

Throughout this paper we study offline reinforcement learning problems that seek a stationary policy with maximum average reward for an infinite-horizon tabular Markov decision process (MDP). While the transition kernel of the MDP is assumed to be unknown, the agent has access to a single

finite state-action trajectory generated under a fixed behavioral policy, which may or may not be known. In offline reinforcement learning, it is common to distinguish *off-policy evaluation* and *offline policy optimization* problems. Off-policy evaluation seeks an accurate estimate for the average reward of a fixed evaluation policy (which typically differs from the behavioral policy that generates the data). Offline policy optimization, on the other hand, uses an off-policy evaluation oracle in order to find a policy that maximizes long-run average reward.

The effectiveness of an offline reinforcement learning algorithm heavily relies on the reliability of the underlying off-policy evaluation oracle. To our best knowledge, the vast majority of commonly used oracles are designed for the discounted reward criterion, with few exceptions such as (Zhang et al., 2021; Saxena et al., 2023). In addition, they typically require access to multiple independent state-action trajectories generated under the same behavioral policy. Arguably the simplest approach to off-policy evaluation is the *direct method*, which uses the available data trajectories to estimate a parametric model for the discounted reward of the evaluation policy. For example, the discounted reward is representable as a function of the transition kernel or the action-value function, which can be inferred via maximum likelihood estimation (Mannor et al., 2007) or linear regression (Ernst et al., 2005; Le et al., 2019), respectively, see also (Antos et al., 2008; Lagoudakis and Parr, 2003). However, the resulting reward estimators are prone to significant bias for several reasons. First, the discounted reward depends nonlinearly on the transition kernel. Thus, unbiased estimators for the transition kernel result in biased reward estimators. This bias is most pronounced if the likelihood ratio between the evaluation policy and the behavioral policy is high, that is, if some actions are chosen much more often under the evaluation policy than under the behavioral policy. Second, parametric estimators for the action-value function are typically biased due to model misspecification. Importance sampling methods reweight the observed rewards using the likelihood ratio between the evaluation and behavioral policies (Precup et al., 2000; Hirano et al., 2003; Swaminathan and Joachims, 2015). This approach obviates an explicit parameterization of the value function but suffers from high variance, which grows rapidly with the length of the observed state-action trajectories. To mitigate these limitations, doubly robust methods combine direct estimation with an importance sampling correction (Jiang and Li, 2016; Thomas and Brunskill, 2016). These approaches use the direct method to construct a baseline estimator and correct residual errors with importance weighting. The resulting estimators remain unbiased if either the baseline estimator or the importance weights are accurate, while generally reducing variance compared to pure importance sampling and mitigating bias compared to the direct method—hence the attribute "*doubly robust*." These methods are known to achieve the lowest asymptotic variance among unbiased estimators (Kallus and Uehara, 2022).

All off-policy evaluation methods discussed so far assume that the testing data is generated under the same transition kernel as the training data. This assumption is inappropriate if the transition kernel of the MDP may change over time because distribution shifts can significantly distort reward estimates (Wang et al., 2024). Distributionally robust off-policy evaluation methods compute the worst-case average reward of the evaluation policy over an ambiguity set of plausible transition kernels. This approach mitigates risks from model misspecification and sampling bias while yielding confidence bounds on the true average reward (Shi and Chi, 2024; Ma et al., 2022; Panaganti et al., 2022; Bhardwaj et al., 2024; Ramesh et al., 2024; Si et al., 2023; Kallus et al., 2022). Notably, state-of-the-art offline reinforcement learning methods often build on pessimistic estimators that underestimate the average reward (Yin and Wang, 2021; Yan et al., 2023; Uehara

et al., 2023; Hu et al., 2025) or work with regularized value functions (Xie et al., 2021; Zhan et al., 2022; Fakoor et al., 2021).

Recall that offline policy optimization estimates the highest achievable reward by maximizing the reward estimate provided by an off-policy evaluation oracle over all admissible policies. It is well known that—for any fixed training sample size—this estimated maximum is *optimistically* biased, even when the underlying off-policy evaluation oracle is unbiased. This is simply a manifestation of the notorious optimizer's curse (Smith and Winkler, 2006). In finite samples, the optimistic bias in the estimated maximum can only be reduced (let alone eliminated) by employing a *pessimistically* biased off-policy evaluation oracle. Such oracles are designed to produce lower confidence bounds on the true maximum reward with a small significance level $\beta \in (0, 1)$. In data-driven optimization, β is sometimes referred to as the *out-of-sample disappointment* (Van Parys et al., 2021).

It is natural to call an off-policy evaluation oracle *efficient* if it is the least conservative among all oracles whose out-of-sample disappointment does not exceed a prescribed threshold. Thus, an efficient oracle strikes an optimal trade-off between the reward it predicts (which should be as *high* as possible) and the out-of-sample disappointment it incurs (which should be as *low* as possible). We aim to construct an off-policy evaluation oracle for tabular MDPs that is efficient in this sense. On a high level, our construction can be explained as follows. We first show that there is a one-to-one correspondence between the stationary state-action-next-state distribution of a controlled MDP and the combination of the MDP's transition kernel and the applied control policy. Thus, the state-action-next-state distribution corresponding to the behavioral policy encapsulates all the information needed to compute the long-run average reward of any given evaluation policy. Next, we demonstrate that the *empirical* state-action-next-state distribution, derived from the observed state-action trajectory of the behavioral policy, serves as a consistent estimator for the true state-action-next-state distribution. Moreover, this estimator satisfies a large deviations principle with a rate function reminiscent of the conditional relative entropy. Finally, we estimate the average reward of the evaluation policy under the unknown true transition kernel by considering the worst-case (infimal) average reward across all transition kernels whose state-action-next-state distributions under the behavioral policy deviate from the empirical state-action-next-state distribution by at most ρ, as measured by the rate function of the large deviations principle. We prove that the out-of-sample disappointment of the resulting distributionally robust off-policy evaluation oracle decays exponentially at a rate of ρ with the length of the observation history. Furthermore, we show that this oracle is the least conservative among all oracles whose out-of-sample disappointment decays at least at rate ρ, making the proposed off-policy oracle asymptotically efficient.

We emphasize that our efficiency guarantees remain valid even when only a single trajectory of serially correlated states and actions is available, and even if the behavioral policy generating this trajectory is unknown. Moreover, we prove that our efficient off-policy evaluation oracle naturally leads to an efficient estimate for the optimal value of the corresponding offline reinforcement learning problem. Computing this optimal value requires solving a robust MDP with a non-rectangular uncertainty set. Unfortunately, such problems are generically NP-hard (Wiesemann et al., 2013). To address this, we tailor the actor-critic algorithm proposed by Li et al. (2023), originally designed for robust discounted reward MDPs with arbitrary non-rectangular uncertainty sets, to the problem at hand. Given an oracle for approximating the (NP-hard) robust policy evaluation subproblem, we show that this algorithm finds an ϵ-optimal solution for the corresponding policy improvement problem in $O(1/\epsilon^4)$ iterations. Approximate solutions for the robust policy evaluation subproblem can be obtained using the randomized projected Langevin dynamics algorithm proposed by Li et al.

(2023). Although the runtime of this algorithm scales exponentially with the number of states and actions, numerical experiments suggest that it remains effective in practice.

In a nutshell, the main contributions of this paper can be summarized as follows.

- We propose a novel approach to offline reinforcement learning that applies even when only a *single trajectory* of correlated data is available, generated under an *unknown* behavioral policy.
- We develop a distributionally robust off-policy evaluation oracle that is statistically efficient, thus optimally balancing in-sample performance against out-of-sample disappointment. Furthermore, we show that this oracle yields an efficient estimator for the optimal value of the corresponding offline reinforcement learning problem.
- Computing the proposed estimators reduces to solving a robust MDP with a non-rectangular uncertainty set. To address this hard problem, we adapt an existing actor-critic algorithm to solve the robust MDP approximately. Numerical experiments show that the resulting estimators are competitive with state-of-the-art baselines on standard test problems.

Our paper contributes to a stream of research that exploits large deviations theory in order to construct distributionally robust estimators for the optimal solutions of data-driven decision problems that enjoy statistical efficiency guarantees. Van Parys et al. (2021) develop efficient estimators for *static* stochastic programs assuming only access to *independent* samples from the distribution of the uncertain problem parameters. They show that a distributionally robust optimization model with a relative entropy uncertainty set is statistically optimal. Sutter et al. (2024) extend this model to more general data-generating processes with serially dependent observations. They show that, when the data process is governed by a parametric distribution and the underlying parameters admit a sufficient statistic satisfying a large deviations principle, then solving a distributionally robust optimization problem with an uncertainty set constructed via the rate function of the large deviations principle is statistically optimal. Li et al. (2021) propose a customized Frank-Wolfe algorithm to compute efficient distributionally robust estimators under the assumption that the data is generated by a Markov chain. However, this algorithm only guarantees convergence to a stationary point, which may lack the efficiency properties of the global optimizer. Bennouna and Van Parys (2021) explore the same setting as Van Parys et al. (2021), analyzing the effects of imposing different decay rates on out-of-sample disappointment. They show that distributionally robust estimators with a relative entropy uncertainty sets are optimal in the exponential regime, variance-regularized empirical estimators are optimal in the sub-exponential regime, and worst-case robust estimators are optimal in the super-exponential regime. Similarly, Ganguly and Sutter (2024) examine the construction of confidence intervals in the moderate deviation regime and establish the efficiency of distributionally robust estimators. Compared to all these works, our model is the only one to simultaneously offer the following benefits. It realistically assumes that the available data is limited to a single finite trajectory of states and actions generated by an MDP controlled by a stationary behavioral policy. It does *not* rely on the restrictive assumption that all transitions between arbitrary state-action pairs have a positive probability. It recognizes that transitions between certain state-action pairs must have identical probabilities because the behavioral policy is stationary. It explicitly accounts for the distribution shift between the state-action trajectory distributions under the behavioral policy and the evaluation policy, respectively. In addition, our model delivers estimators that enjoy both asymptotic and finite-sample guarantees on the out-of-sample disappointment. Finally, we provide an algorithm that solves the relevant robust MDPs to global optimality, thus ensuring that

the proposed efficient estimators are accessible. Sutter et al. (2021) also study off-policy evaluation problems that explicitly account for distribution shifts. However, they assume to have access to independent samples from the stationary state-action-next-state distribution, and they assume that only the reward function is unknown, whereas the transition kernel and the behavioral policy are known. Finally, their methods do not extend easily to offline policy optimization.

Structure. The rest of the paper is structured as follows. Section 2 reviews and extends the large deviations theory for Markov chains and Markov decision processes. Leveraging this theory, Sections 3 and 4 propose statistically optimal solutions for the off-policy evaluation and the offline policy optimization problems, respectively. Section 5 then develops a projected Langevin dynamics method for solving the robust policy evaluation problem and an actor-critic method for solving the offline policy optimization problem. Finally, Section 6 validates the efficiency of the proposed estimators on standard test problems from reinforcement learning and operations research.

Notation. The probability simplex over a finite set $\mathcal{X}$ is defined as $\Delta(\mathcal{X}) = \{p \in \mathbb{R}_+^{|\mathcal{X}|} : \sum_{x\in\mathcal{X}} p(x) = 1\}$. The relative entropy of $p \in \Delta(\mathcal{X})$ with respect to $q \in \Delta(\mathcal{X})$ is defined as $\mathsf{D}(p\|q) = \sum_{x\in\mathcal{X}} p(x) \log\left(p(x)/q(x)\right)$, where we use the conventions $0\log(0/t) = 0$ for any $t \geq 0$ and $t\log(t/0) = \infty$ for any $t > 0$. The support of $p \in \Delta(\mathcal{X})$ is the set $\mathrm{supp}(p) = \{x \in \mathcal{X} : p(x) > 0\}$.

2 Statistics of Markov Chains and Markov Decision Processes

Off-policy evaluation and offline policy optimization constitute statistical learning problems based on Markovian data. In Section 2.1 we thus review and generalize a large deviations principle for the doublet distribution of an irreducible Markov chain, which provides a mathematical framework for studying the probabilities of rare events. Building on these insights, in Section 2.2 we then derive a large deviations principle for the state-action-next-state distribution of a Markov decision process.

2.1 Markov Chains

Consider a time-homogeneous irreducible Markov chain given by a triple $(\mathcal{X}, P, \gamma)$ consisting of a finite state space $\mathcal{X} = \{1, \ldots, d\}$, a transition probability matrix $P \in \Delta(\mathcal{X})^d$ and an initial distribution $\gamma \in \Delta(\mathcal{X})$. If the system underlying the Markov chain is in state $x_t \in \mathcal{X}$ at time t, then it moves to state $x_{t+1} \in \mathcal{X}$ at time $t+1$ with probability $P(x_t, x_{t+1})$. Thus, $P(x_t, \cdot)$ represents the distribution of X_{t+1} conditional on $X_t = x_t$. There exists a unique probability measure $\mathbb{P}_P$ defined on the canonical sample space $\Omega = \mathcal{X}^\infty$ equipped with its power set σ-algebra $\mathcal{F} = 2^\Omega$ such that

$$\mathbb{P}_P(X_1 = x_1) = \gamma(x_1) \quad \forall x_1 \in \mathcal{X}$$

and

$$\mathbb{P}_P(X_{t+1} = x_{t+1} | X_t = x_t, \ldots, X_1 = x_1) = P(x_t, x_{t+1}) \quad \forall x_1, \ldots x_{t+1} \in \mathcal{X}.$$

Note that $\mathbb{P}_P$ also depends on γ, but we suppress this dependence to avoid clutter. As the Markov chain at hand is irreducible, the Perron-Frobenius theorem implies that there exists a unique stationary state distribution $\mu \in \Delta(\mathcal{X})$ that satisfies $\mu(y) = \sum_{x\in\mathcal{X}} \mu(x)P(x,y) > 0$ for all $y \in \mathcal{X}$. Using P and μ, we can further define the stationary doublet distribution $\theta \in \Delta(\mathcal{X} \times \mathcal{X})$ through $\theta(x,y) = \mu(x)P(x,y)$ for all $x, y \in \mathcal{X}$. From the last formula it is evident that the transition probability matrix P can be recovered from θ, that is, we have $P(x,y) = \theta(x,y)/\mu(x)$ for all $x, y \in \mathcal{X}$. Hence, there is a one-to-one correspondence between P and θ. Without loss of generality, we can

thus denote the probability measure governing the Markov chain $\{X_t\}_{t=1}^{\infty}$ by $\mathbb{P}_\theta$ instead of $\mathbb{P}_P$. We prefer θ to P because it admits a simple estimator that satisfies a large deviations principle.

Note that θ has balanced marginals in the sense that

$$\sum_{x\in\mathcal{X}} \theta(x,y) = \mu(y) = \sum_{y\in\mathcal{X}} \theta(y,x) \quad \forall y \in \mathcal{X}, \tag{1}$$

where the first equality follows from the definition of μ, and the second equality follows from the definition of θ. In the following we use Θ to denote the set of all doublet distributions $\theta \in \Delta(\mathcal{X}\times\mathcal{X})$ that satisfy (1). We emphasize that not every $\theta \in \Theta$ represents the doublet distribution of an irreducible Markov chain, that is, it is not sufficient for θ to have balanced marginals. In addition, the pair $(\mathcal{X}, \mathrm{supp}(\theta))$ must represent a strongly connected directed graph with vertex set $\mathcal{X}$ and edge set $\mathrm{supp}(\theta)$. The strong connectedness ensures that the underlying Markov chain has only one single communicating class of states. In the following, we use $\Theta_0 \subseteq \Theta$ to denote the set of all doublet distributions that induce an irreducible Markov chain.

The Markov law of large numbers (Norris, 1998, Theorem 1.7.6) implies that

$$\theta(x,y) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_\theta(X_t = x, X_{t+1} = y) \quad \forall x,y \in \mathcal{X}, \quad \forall \theta \in \Theta_0. \tag{2}$$

Given a finite trajectory of state observations $\{X_t\}_{t=1}^{T}$, a natural estimator for θ is thus given by the empirical doublet distribution $\widehat{\theta}_T \in \Delta(\mathcal{X}\times\mathcal{X})$, which is defined through

$$\widehat{\theta}_T(x,y) = \frac{1}{T}\left(\sum_{t=1}^{T-1} 1_{X_t=x, X_{t+1}=y} + 1_{X_T=x, X_1=y}\right) \quad \forall x,y \in \mathcal{X}. \tag{3}$$

The ghost transition from X_T to X_1 (Vidyasagar, 2014) in this definition ensures that $\widehat{\theta}_T$ has balanced marginals and thus lies in Θ. In the following, we refer to elements of Θ_0 as models and elements of Θ as estimator realizations. It is natural to measure the discrepancy between a model $\theta \in \Theta_0$ and an estimator realization $\theta' \in \Theta$ by their conditional relative entropy.

Definition 2.1 (Conditional relative entropy for Markov chains) *The conditional relative entropy of $\theta' \in \Theta$ with respect to $\theta \in \Theta_0$ is defined as*

$$\mathsf{D}_{\mathsf{mc}}(\theta'\|\theta) = \sum_{x,y\in\mathcal{X}} \theta'(x,y)\left(\log\frac{\theta'(x,y)}{\sum_{z\in\mathcal{X}}\theta'(x,z)} - \log\frac{\theta(x,y)}{\sum_{z\in\mathcal{X}}\theta(x,z)}\right). \tag{4}$$

The conditional relative entropy is well-defined for all $\theta' \in \Theta$ and $\theta \in \Theta_0$. Indeed, due to our conventions for the logarithm, $\mathsf{D}_{\mathsf{mc}}(\theta'\|\theta)$ is finite whenever $\mathrm{supp}(\theta') \subseteq \mathrm{supp}(\theta)$ and evaluates to $+\infty$ otherwise. If the doublet distributions θ' and θ belong to Θ_0, then they induce irreducible Markov chains with unique transition probability matrices P' and P and stationary distributions μ' and μ, respectively. In this case, the conditional relative entropy can be equivalently expressed as

$$\mathsf{D}_{\mathsf{mc}}(\theta'\|\theta) = \sum_{x,y\in\mathcal{X}} \theta'(x,y)\log\frac{P'(x,y)}{P(x,y)} = \sum_{x\in\mathcal{X}} \mu'(x)\,\mathsf{D}(P'(x,\cdot)\|P(x,\cdot)). \tag{5}$$

The last formula in (5) motivates the name "conditional relative entropy." In addition, $\mathsf{D}_{\mathsf{mc}}(\theta'\|\theta)$ admits a unique lower semi-continuous extension to $\Theta \times \Theta$, which is obtained by setting

$$\mathsf{D}_{\mathsf{mc}}(\theta'\|\theta) = \lim_{\delta \downarrow 0} \inf_{(\vartheta',\vartheta)\in\Theta\times\Theta_0} \left\{ \mathsf{D}_{\mathsf{mc}}(\vartheta'\|\vartheta) : \|(\vartheta',\vartheta) - (\theta',\theta)\| \leq \delta \right\} \quad \forall (\theta',\theta) \in \Theta \times (\Theta \backslash \Theta_0);$$

see (Sutter et al., 2024, p. 1996) and (Rockafellar and Wets, 2009, Definition 1.5). In the following, we will always mean this lower semi-continuous extension to $\Theta \times \Theta$ when referring to $\mathsf{D}_{\mathsf{mc}}(\theta'\|\theta)$.

The conditional relative entropy is significant because it represents the rate function of a large deviations principle for the empirical doublet distribution $\widehat{\theta}_T$. Large deviations theory provides bounds on the exponential rate at which the probability of a rare event $\widehat{\theta}_T \in \mathcal{D}$ decays as the length T of the observation history grows. These bounds are expressed in terms of the infimum of a rate function over the set $\mathcal{D}$ or its interior. The classical large deviations principle for Markov chains (Natarajan, 1985, Theorem 1) assumes that $\mathrm{supp}(\theta) = \mathcal{X} \times \mathcal{X}$. We generalize this classical result to arbitrary irreducible Markov chains. This generalization necessitates only cosmetic changes in the proof. We provide a proof sketch to keep this paper self-contained.

Theorem 2.2 (Large deviations principle for Markov chains) *For all $\theta \in \Theta_0$ and Borel sets $\mathcal{D} \subseteq \Theta$, the empirical doublet distribution $\widehat{\theta}_T$ defined in* (3) *satisfies*

$$\begin{aligned} -\inf_{\theta' \in \operatorname{int} \mathcal{D}} \mathsf{D}_{\mathsf{mc}}(\theta'\|\theta) &\leq \liminf_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta \left(\widehat{\theta}_T \in \mathcal{D} \right) \\ &\leq \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta \left(\widehat{\theta}_T \in \mathcal{D} \right) \leq -\inf_{\theta' \in \mathcal{D}} \mathsf{D}_{\mathsf{mc}}(\theta'\|\theta). \end{aligned}$$

Proof Fix an arbitrary $\theta \in \Theta_0$, and define $\Theta_T = \Theta \cap \frac{1}{T}\{0,\ldots,T\}^{d\times d}$ as the set of all possible realizations of $\widehat{\theta}_T$. One readily verifies that

$$\begin{aligned} \frac{1}{T} \log \mathbb{P}_\theta \left(\widehat{\theta}_T \in \mathcal{D} \right) &= \frac{1}{T} \log \mathbb{P}_\theta \left(\widehat{\theta}_T \in \mathcal{D} \cap \Theta_T \right) \leq \frac{1}{T} \log \left(|\mathcal{D} \cap \Theta_T| \sup_{\theta' \in \mathcal{D} \cap \Theta_T} \mathbb{P}_\theta \left(\widehat{\theta}_T = \theta' \right) \right) \\ &\leq \frac{1}{T} \log \left(|\Theta_T| \sup_{\theta' \in \mathcal{D}} \mathbb{P}_\theta \left(\widehat{\theta}_T = \theta' \right) \right). \end{aligned}$$

Next, define the type class

$$J(\theta') = \left\{ \{x_t\}_{t=1}^T \in \mathcal{X}^T : \frac{1}{T} \left(\sum_{t=1}^{T-1} \mathsf{1}_{x_t=x, x_{t+1}=y} + \mathsf{1}_{x_T=x, x_1=y} \right) = \theta'(x,y) \ \forall x,y \in \mathcal{X} \right\} \tag{6}$$

as the set of all sample paths consistent with the estimator realization $\theta' \in \Theta$. Thus, we have that $\widehat{\theta}_T = \theta'$ if and only if $\{X_t\}_{t=1}^T \in J(\theta')$. The above reasoning therefore implies that

$$\begin{aligned} &\frac{1}{T} \log \mathbb{P}_\theta \left(\widehat{\theta}_T \in \mathcal{D} \right) \\ &\qquad \leq \frac{d^2}{T} \log(T+1) + \frac{1}{T} \log \sup_{\theta' \in \mathcal{D}} \mathbb{P}_\theta \left(\widehat{\theta}_T = \theta' \right) \\ &\qquad \leq \frac{d^2}{T} \log(T+1) + \frac{1}{T} \sup_{\theta' \in \mathcal{D}} \log \left(|J(\theta')| \sup_{\{x_t\}_{t=1}^T \in J(\theta')} \mathbb{P}_\theta(X_t = x_t \ \forall t = 1,\ldots,T) \right). \end{aligned} \tag{7}$$

Next, use μ and μ' to denote the stationary state distributions corresponding to the stationary doublet distributions θ and θ', respectively. It follows from (Vidyasagar, 2014, Theorem 11) that

$$-d^2 \log(2T) + T\,\mathsf{H_c}(\theta') \leq \log|J(\theta')| \leq \log T + T\,\mathsf{H_c}(\theta'), \tag{8}$$

where

$$\mathsf{H_c}(\theta') = \sum_{x\in\mathcal{X}} \mu'(x)\log\mu'(x) - \sum_{x,y\in\mathcal{X}} \theta'(x,y)\log\theta'(x,y)$$

is the conditional entropy of the estimator realization $\theta' \in \Theta$. Similarly, it follows from inequalities (35) and (36) in (Vidyasagar, 2014) that

$$-T\left(\mathsf{H_c}(\theta') + \mathsf{D_{mc}}(\theta'\|\theta)\right) + \underline{c} \leq \log \mathbb{P}_\theta(X_t = x_t\ \forall t = 1,\ldots,T) \leq -T\left(\mathsf{H_c}(\theta') + \mathsf{D_{mc}}(\theta'\|\theta)\right) + \bar{c}, \tag{9}$$

where $\underline{c} = \min_{x,y\in\mathcal{X}} \log(\mu(x)\mu(y))/\theta(x,y))$ and $\bar{c} = \max_{x,y\in\mathcal{X}} \log(\mu(x)\mu(y))/\theta(x,y))$. Substituting these estimates into (7) yields

$$\begin{aligned}\frac{1}{T}\log\mathbb{P}_\theta\left(\widehat{\theta}_T \in \mathcal{D}\right) &\leq \frac{d^2}{T}\log(T+1) + \frac{1}{T}\log(T) + \sup_{\theta'\in\mathcal{D}}(-\mathsf{D_{mc}}(\theta'\|\theta) + \frac{1}{T}\bar{c}) \\ &= \frac{d^2}{T}\log(T+1) + \frac{1}{T}(\log(T) + \bar{c}) - \inf_{\theta'\in\mathcal{D}} \mathsf{D_{mc}}(\theta'\|\theta),\end{aligned}$$

from which the upper bound on $\limsup_{T\to\infty} \frac{1}{T}\log\mathbb{P}_\theta(\widehat{\theta}_T \in \mathcal{D})$ follows immediately.

As for the lower bound, note that $\cup_{T\in\mathbb{N}}\Theta_T$ is dense in $\mathrm{int}\mathcal{D}$. Hence, there is $T_0 \in \mathbb{N}$ that only depends on $\mathcal{D}$ and a deterministic sequence $\{\theta'_T\}_{T\in\mathbb{N}}$ such that $\theta'_T \in \Theta_T \cap \mathrm{int}\mathcal{D}$ for all $T \geq T_0$ and

$$\inf_{\theta'\in\mathrm{int}\mathcal{D}} \mathsf{D_{mc}}(\theta'\|\theta) = \liminf_{T\to\infty} \mathsf{D_{mc}}(\theta'_T\|\theta).$$

For all $T \geq T_0$, we then have

$$\begin{aligned}\frac{1}{T}\log\mathbb{P}_\theta\left(\widehat{\theta}_T \in \mathcal{D}\right) &\geq \frac{1}{T}\log\mathbb{P}_\theta\left(\widehat{\theta}_T = \theta'_T\right) \\ &\geq \frac{1}{T}\log\left(|J(\theta'_T)| \inf_{\{x_t\}_{t=1}^T \in J(\theta'_T)} \mathbb{P}_\theta(X_t = x_t\ \forall t = 1,\ldots,T)\right) \\ &\geq -\frac{d^2}{T}\log(2T) + \frac{1}{T}(\log T + \underline{c}) - \mathsf{D_{mc}}(\theta'_T\|\theta),\end{aligned}$$

where the first equality holds because $T \geq T_0$, and the third inequality follows again from (8) and (9). Taking the limit inferior as T tends to infinity finally yields the desired lower bound. ■

Corollary 2.3 below establishes a finite-sample version of the large deviations upper bound in Theorem 2.2. Its proof follows immediately from that of Theorem 2.2 and is therefore omitted.

Corollary 2.3 (Finite-sample version of Theorem 2.2) *For all $\theta \in \Theta_0$ and Borel sets $\mathcal{D} \subseteq \Theta$, the empirical doublet distribution $\widehat{\theta}_T$ defined in* (3) *satisfies*

$$\frac{1}{T}\log\mathbb{P}_\theta\left(\widehat{\theta}_T \in \mathcal{D}\right) \leq \frac{1}{T}(\log T + \bar{c} + d^2\log(T+1)) - \inf_{\theta'\in\mathcal{D}} \mathsf{D_{mc}}(\theta'\|\theta) \qquad \forall T \in \mathbb{N},$$

where $\bar{c} > 0$ is a universal constant that depends only on θ.

Besides representing the rate function of a large deviations principle, the conditional relative entropy $\mathsf{D_{mc}}$ has many useful properties including level compactness, radial monotonicity and coercivity. We refer to (Sutter et al., 2024, Proposition 5.1) for a discussion of these properties.

2.2 Markov Decision Processes

Consider a *Markov decision process* (MDP) given by a five-tuple $(\mathcal{S}, \mathcal{A}, Q, r, \eta)$ consisting of a finite state space $\mathcal{S} = \{1, \ldots, S\}$, a finite action space $\mathcal{A} = \{1, \ldots, A\}$, a transition kernel $Q \in \mathcal{Q} = \Delta(\mathcal{S})^{SA}$, a reward-per-stage function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and an initial distribution $\eta \in \Delta(\mathcal{S})$. If the system underlying the MDP is in state $s_t \in \mathcal{S}$ at time t and action $a_t \in \mathcal{A}$ is executed, then an immediate reward $r(s_t, a_t)$ is earned, and the system moves to state s_{t+1} at time $t + 1$ with probability $Q(s_{t+1}|s_t, a_t)$. Thus, $Q(\cdot|s_t, a_t)$ represents the distribution of S_{t+1} conditional on $S_t = s_t$ and $A_t = a_t$. Actions are chosen according to a stationary policy, which is described by a stochastic kernel $\pi \in \Pi = \Delta(\mathcal{A})^S$. That is, the probability of choosing action a_t if the current state is s_t is characterized by $\pi(a_t|s_t)$. Thus, $\pi(\cdot|s_t) \in \Delta(\mathcal{A})$ represents the distribution of A_t conditional on $S_t = s_t$. Under a stationary policy π, there exists a unique probability measure $\mathbb{P}_{\pi,Q}$ defined on the canonical sample space $\Omega = (\mathcal{S} \times \mathcal{A})^\infty$ equipped with its power set σ-algebra $\mathcal{F} = 2^\Omega$ such that

$$\mathbb{P}_{\pi,Q}(S_1 = s_1) = \eta(s_1) \quad \forall s_1 \in \mathcal{S}, \tag{10a}$$

and for all $t \in \mathbb{N}$ we have

$$\begin{aligned} &\mathbb{P}_{\pi,Q}(S_{t+1} = s_{t+1}|S_t = s_t, A_t = a_t, \ldots, S_1 = s_1, A_1 = a_1) \\ &\qquad = Q(s_{t+1}|s_t, a_t) \quad \forall s_1, \ldots, s_{t+1} \in \mathcal{S},\ a_1, \ldots, a_t \in \mathcal{A}, \end{aligned} \tag{10b}$$

and

$$\begin{aligned} &\mathbb{P}_{\pi,Q}(A_{t+1} = a_{t+1}|S_{t+1} = s_{t+1}, S_t = s_t, A_t = a_t, \ldots, S_1 = s_1, A_1 = a_1) \\ &\qquad = \pi(a_{t+1}|s_{t+1}) \quad \forall s_1, \ldots, s_{t+1} \in \mathcal{S},\ a_1, \ldots, a_{t+1} \in \mathcal{A}. \end{aligned} \tag{10c}$$

Further details about the construction of $\mathbb{P}_{\pi,Q}$ are provided in (Hernández-Lerma and Lasserre, 1996, § 2.2). Note that $\mathbb{P}_{\pi,Q}$ also depends on η, but we suppress this dependence notationally to avoid clutter. To simplify notation, we will use X_t as a shorthand for the state-action pair (S_t, A_t). Note that X_t ranges over $\mathcal{X} = \mathcal{S} \times \mathcal{A}$. As in Section 2.1, $d = SA$ denotes the cardinality of $\mathcal{X}$.

Proposition 2.4 (State-action process $\{X_t\}_{t=1}^\infty$) *The stochastic process $\{X_t\}_{t=1}^\infty$ represents a time-homogeneous Markov chain under $\mathbb{P}_{\pi,Q}$, and its transition probability matrix P satisfies*

$$P((s, a), (s', a')) = \pi(a'|s')Q(s'|s, a) \quad \forall s, s' \in \mathcal{S},\ a, a' \in \mathcal{A}. \tag{11}$$

Proof By the construction of $\mathbb{P}_{\pi,Q}$, we have for all $t \in \mathbb{N}$ that

$$\begin{aligned} &\mathbb{P}_{\pi,Q}(X_{t+1} = x_{t+1}|X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_1 = x_1) \\ &\quad = \mathbb{P}_{\pi,Q}(S_{t+1} = s_{t+1}, A_{t+1} = a_{t+1}|S_t = s_t, A_t = a_t, \ldots, S_1 = s_1, A_1 = a_1) \\ &\quad = \mathbb{P}_{\pi,Q}(A_{t+1} = a_{t+1}|S_{t+1} = s_{t+1}, S_t = s_t, A_t = a_t, \ldots, S_1 = s_1, A_1 = a_1) \\ &\qquad \times \mathbb{P}_{\pi,Q}(S_{t+1} = s_{t+1}|S_t = s_t, A_t = a_t, \ldots, S_1 = s_1, A_1 = a_1) \\ &\quad = \pi(a_{t+1}|s_{t+1})Q(s_{t+1}|s_t, a_t) \quad \forall s_1, \ldots, s_{t+1} \in \mathcal{S},\ a_1, \ldots, a_{t+1} \in \mathcal{A}, \end{aligned}$$

where the second equality follows from Bayes' law, while the third equality exploits (10b) and (10c). Thus, the stochastic process $\{X_t\}_{t=1}^\infty$ satisfies the Markov property, and its transition probability matrix satisfies (11). Since π and Q are independent of t, the Markov chain is time-homogeneous. ■

Proposition 2.4 implies that the triple $(\mathcal{X}, P, \gamma)$ induces a Markov chain with initial distribution $\gamma \in \Delta(\mathcal{X})$ defined through $\gamma(s,a) = \eta(s)\pi(a,s)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. For this Markov chain to be irreducible, we require that $\pi > 0$. The family of all stationary policies with this property is denoted by Π_0. In addition, we require that the directed graph $(\mathcal{X}, \mathcal{E})$ with edge set

$$\mathcal{E} = \big\{((s,a),(s',a')) \in (\mathcal{S}\times\mathcal{A})^2 \,:\, Q(s'|s,a) > 0\big\}$$

is strongly connected. The family of all transition kernels with this property is denoted by $\mathcal{Q}_0$. The following lemma is elementary, and therefore its proof is omitted.

Lemma 2.5 *If $\pi \in \Pi_0$ and $Q \in \mathcal{Q}_0$, then the Markov chain $\{X_t\}_{t=1}^\infty$ is irreducible under $\mathbb{P}_{\pi,Q}$. Conversely, if $\pi \in \Pi\backslash\Pi_0$ or $Q \in \mathcal{Q}\backslash\mathcal{Q}_0$, then the Markov chain $\{X_t\}_{t=1}^\infty$ is not irreducible under $\mathbb{P}_{\pi,Q}$.*

As the Markov chain induced by $\pi \in \Pi_0$ and $Q \in \mathcal{Q}_0$ is irreducible, we know from Section 2.1 that it admits a unique stationary distribution $\mu \in \Delta(\mathcal{S}\times\mathcal{A})$ as well as a unique stationary doublet distribution $\theta \in \Delta((\mathcal{S}\times\mathcal{A})^2)$. In addition, π and Q induce a unique stationary state-action-next-state distribution $\xi \in \Delta(\mathcal{S}\times\mathcal{A}\times\mathcal{S})$ defined through $\xi(s,a,s') = \sum_{a'\in\mathcal{A}} \theta((s,a),(s',a'))$ for all $s,s' \in \mathcal{S}$ and $a \in \mathcal{A}$. Recall from Section 2.1 that Θ_0 denotes the family of doublet distributions corresponding to irreducible Markov chains on $\mathcal{X} = \mathcal{S}\times\mathcal{A}$. We emphasize, however, that not every Markov chain on $\mathcal{X}$ with doublet distribution $\theta \in \Theta_0$ has a transition probability matrix of the form (11). Instead, the state-action-next-state distribution ξ, which has fewer degrees of freedom than θ, provides sufficient information to reconstruct π and Q. To see this, note first that

$$\xi(s,a,s') = Q(s'|s,a)\mu(s,a) \quad \forall s,s' \in \mathcal{S},\ a,a' \in \mathcal{A} \tag{12}$$

and that $\mu(s,a) = \sum_{s'\in\mathcal{S}} \xi(s,a,s')$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Hence, we have

$$Q(s'|s,a) = \frac{\xi(s,a,s')}{\mu(s,a)} = \frac{\xi(s,a,s')}{\sum_{\tilde{s}\in\mathcal{S}} \xi(s,a,\tilde{s})} \quad \forall s,s' \in \mathcal{S},\ a \in \mathcal{A}. \tag{13a}$$

In addition, the defining equation of the stationary distribution μ implies that

$$\mu(s',a') = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mu(s,a)P((s,a),(s',a')) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mu(s,a)\pi(a'|s')Q(s'|s,a) \quad \forall s' \in \mathcal{S},\ a' \in \mathcal{A},$$

where the second equality follows from Proposition 2.4. This readily implies that

$$\pi(a'|s') = \frac{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mu(s,a)Q(s'|s,a)}{\mu(s',a')} = \frac{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \xi(s,a,s')}{\sum_{s\in\mathcal{S}} \xi(s,a',s')} \quad \forall s' \in \mathcal{S},\ a' \in \mathcal{A}, \tag{13b}$$

where the second equality exploits (12). In summary, (13) shows that both π and Q can indeed be reconstructed from ξ. Hence, there is a one-to-one correspondence between (π, Q) and ξ. Without loss of generality, we can thus denote the probability measure governing the Markov chain $\{X_t\}_{t=1}^\infty$ by $\mathbb{P}_\xi$ instead of $\mathbb{P}_{\pi,Q}$. We prefer ξ to (π, Q) because it admits a simple estimator akin to $\widehat{\theta}_T$.

Note that ξ has balanced marginals in the sense that

$$\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} \xi(s,a,s') = \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} \xi(s',a,s) \quad \forall s \in \mathcal{S}. \tag{14}$$

In the following we use Ξ to denote the set of all $\xi \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$ that satisfy (14). We emphasize that not every $\xi \in \Xi$ is induced by an irreducible Markov chain. To see this, recall that if $\pi \in \Pi_0$ and $Q \in \mathcal{Q}_0$, then the Markov chain $\{X_t\}_{t=1}^{\infty}$ is irreducible, and thus $\mu > 0$. By (12), this ensures that the edge set $\mathcal{E}$ admits the equivalent representation

$$\mathcal{E} = \left\{((s,a),(s',a')) \in (\mathcal{S} \times \mathcal{A})^2 \,:\, \xi(s,a,s') > 0\right\}.$$

Hence, if $\pi \in \Pi_0$ and $Q \in \mathcal{Q}_0$, then $(\mathcal{X}, \mathcal{E})$ must represent a strongly connected graph. We use $\Xi_0 \subseteq \Xi$ to denote the set of all $\xi \in \Xi$ with this property. By what has been said above, it is now easy to prove that every $\pi \in \Pi_0$ and $Q \in \mathcal{Q}_0$ induces a unique $\xi \in \Xi_0$ and vice versa.

Given a finite state-action trajectory $\{(S_t, A_t)\}_{t=1}^{T}$, a natural estimator for ξ is thus the empirical state-action-next-state distribution $\widehat{\xi}_T \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$ defined through

$$\widehat{\xi}_T(s,a,s') = \frac{1}{T}\left(\sum_{t=1}^{T-1} 1_{S_t=s, A_t=a, S_{t+1}=s'} + 1_{S_T=s, A_T=a, S_1=s'}\right) \quad \forall s, s' \in \mathcal{S},\ a \in \mathcal{A}. \tag{15}$$

The ghost transition from (S_T, A_T) to S_1 in the above definition ensures that the estimator $\widehat{\xi}_T$ has balanced marginals and thus lies in Ξ. One readily realizes that $\widehat{\xi}_T = F(\widehat{\theta}_T)$, where $\widehat{\theta}_T$ denotes the empirical doublet distribution defined in (3), and $F : \Theta \to \Xi$ satisfies

$$F(\theta)(s,a,s') = \sum_{a' \in \mathcal{A}} \theta((s,a),(s',a')) \quad \forall s, s' \in \mathcal{S},\ a \in \mathcal{A}. \tag{16}$$

We henceforth refer to elements of Ξ_0 as models and elements of Ξ as estimator realizations. The discrepancy between a model $\xi \in \Xi_0$ and an estimator realization $\xi' \in \Xi$ is naturally measured by a conditional relative entropy tailored to MDPs.

Definition 2.6 (Conditional relative entropy for MDPs) *The conditional relative entropy of $\xi' \in \Xi$ with respect to $\xi \in \Xi_0$ is defined as*

$$\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi) = \sum_{s,s' \in \mathcal{S}, a \in \mathcal{A}} \xi'(s,a,s') \left(\log \frac{\xi'(s,a,s')}{\sum_{\tilde{s} \in \mathcal{S}, \tilde{a} \in \mathcal{A}} \xi'(s,\tilde{a},\tilde{s})} - \log \frac{\xi(s,a,s')}{\sum_{\tilde{s} \in \mathcal{S}, \tilde{a} \in \mathcal{A}} \xi(s,\tilde{a},\tilde{s})}\right). \tag{17}$$

Due to our conventions for the logarithm, $\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi)$ is finite whenever $\mathrm{supp}(\xi') \subseteq \mathrm{supp}(\xi)$ and evaluates to $+\infty$ otherwise. In addition, $\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi)$ admits a unique lower semi-continuous extension to $\Xi \times \Xi$, which is obtained by setting

$$\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi) = \lim_{\delta \downarrow 0} \inf_{(\zeta',\zeta) \in \Xi \times \Xi_0} \left\{\mathsf{D}_{\mathsf{mdp}}(\zeta'\|\zeta) : \|(\zeta',\zeta) - (\xi',\xi)\| \leq \delta\right\} \quad \forall (\xi',\xi) \in \Xi \times (\Xi \backslash \Xi_0);$$

see (Sutter et al., 2024, p. 1996) and (Rockafellar and Wets, 2009, Definition 1.5). In the following, we will always mean this lower semi-continuous extension to $\Xi \times \Xi$ when referring to $\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi)$. The next proposition shows that the conditional relative entropy $\mathsf{D}_{\mathsf{mdp}}$ for MDPs is closely related to the conditional relative entropy D_{mc} for Markov chains introduced in Definition 2.1.

Proposition 2.7 (Relation between D_{mc} and $\mathsf{D}_{\mathsf{mdp}}$) *If $G : \Xi \to \Theta$ is defined through*

$$G(\xi)((s,a),(s',a')) = \begin{cases} \frac{\sum_{\tilde{s} \in \mathcal{S}} \xi(s',a',\tilde{s})}{\sum_{\tilde{s} \in \mathcal{S}, \tilde{a} \in \mathcal{A}} \xi(s',\tilde{a},\tilde{s})}\, \xi(s,a,s') & \text{if } \sum_{\tilde{s} \in \mathcal{S}} \xi(s',a',\tilde{s}) > 0 \\ 0 & \text{if } \sum_{\tilde{s} \in \mathcal{S}} \xi(s',a',\tilde{s}) = 0 \end{cases} \tag{18}$$

for all $s, s' \in \mathcal{S}$ and $a, a' \in \mathcal{A}$, then $\mathsf{D}_{\mathsf{mc}}(G(\xi')\|G(\xi)) = \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi)$ for all $\xi' \in \Xi$ and $\xi \in \Xi_0$.

An elementary calculation reveals that $\theta = G(\xi)$ belongs to $\Delta((\mathcal{S}\times\mathcal{A})^2)$ and satisfies (1) for any $\xi \in \Xi$. Therefore, Θ represents indeed the codomain of G. Note also that if $\xi \in \Xi_0$ and if we define the policy π via (13b), then $G(\xi)((s,a),(s',a')) = \pi(a'|s')\xi(s,a,s')$ for all $s,s' \in \mathcal{S}$ and $a,a' \in \mathcal{A}$.

Proof of Proposition 2.7 Select any $\xi \in \Xi_0$ and $\xi' \in \Xi$, and denote the stationary state-action and state distributions corresponding to ξ as μ and $\mu_{\mathcal{S}}$, respectively. Thus, we have

$$\mu(s,a) = \sum_{\tilde{s}\in\mathcal{S}} \xi(s,a,\tilde{s}) \quad \text{and} \quad \mu_{\mathcal{S}}(s) = \sum_{\tilde{s}\in\mathcal{S}}\sum_{\tilde{a}\in\mathcal{A}} \xi(s,\tilde{a},\tilde{s}) \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}.$$

The stationary state-action distribution μ' and the stationary state distribution $\mu'_{\mathcal{S}}$ corresponding to ξ' are defined analogously. We prove the claim first under the assumption that $\operatorname{supp}(\xi') \subseteq \operatorname{supp}(\xi)$. This implies that $\operatorname{supp}(G(\xi')) \subseteq \operatorname{supp}(G(\xi))$. By the definitions of D_{mc} and G, we thus have

$$\begin{aligned}
&\mathsf{D}_{\mathsf{mc}}(G(\xi')\|G(\xi)) \\
&= \sum_{s,s'\in\mathcal{S},a,a'\in\mathcal{A}} \frac{\mu'(s',a')}{\mu'_{\mathcal{S}}(s')}\xi'(s,a,s')\left(\log\left(\frac{\mu'(s',a')}{\mu'_{\mathcal{S}}(s')}\frac{\xi'(s,a,s')}{\mu'(s,a)}\right) - \log\left(\frac{\mu(s',a')}{\mu_{\mathcal{S}}(s')}\frac{\xi(s,a,s')}{\mu(s,a)}\right)\right) \\
&= \sum_{s,s'\in\mathcal{S},a\in\mathcal{A}} \xi'(s,a,s')\log\frac{\xi'(s,a,s')}{\xi(s,a,s')} + \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} \mu'(s',a')\log\frac{\mu'(s',a')}{\mu(s',a')} - \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu'(s,a)\log\frac{\mu'(s,a)}{\mu(s,a)} \\
&\quad - \sum_{s'\in\mathcal{S}} \mu'_{\mathcal{S}}(s')\log\frac{\mu'_{\mathcal{S}}(s')}{\mu_{\mathcal{S}}(s')} \\
&= \sum_{s,s'\in\mathcal{S},a\in\mathcal{A}} \xi'(s,a,s')\log\frac{\xi'(s,a,s')}{\xi(s,a,s')} - \sum_{s\in\mathcal{S}} \mu'_{\mathcal{S}}(s)\log\frac{\mu'_{\mathcal{S}}(s)}{\mu_{\mathcal{S}}(s)} \\
&= \sum_{s,s'\in\mathcal{S},a\in\mathcal{A}} \xi'(s,a,s')\left(\log\frac{\xi'(s,a,s')}{\sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}}\xi'(s,\tilde{a},\tilde{s})} - \log\frac{\xi(s,a,s')}{\sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}}\xi(s,\tilde{a},\tilde{s})}\right),
\end{aligned}$$

where we have repeatedly used the convention that $0\log(0/q) = 0$ for all $q \geq 0$. Assume next that $\operatorname{supp}(\xi') \not\subseteq \operatorname{supp}(\xi)$ such that $\operatorname{supp}(G(\xi')) \not\subseteq \operatorname{supp}(G(\xi))$. In this case, we find

$$\mathsf{D}_{\mathsf{mc}}(G(\xi')\|G(\xi)) = \infty = \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi)$$

thanks to our convention that $p\log(p/0) = \infty$ for all $p > 0$. Hence, the claim follows. ■

Equipped with Proposition 2.7, we are now ready to prove that the empirical state-action-next-state distribution $\widehat{\xi}_T$ satisfies a large deviations principle with rate function $\mathsf{D}_{\mathsf{mdp}}$.

Theorem 2.8 (Large deviations principle for MDPs) *For all $\xi \in \Xi_0$ and Borel sets $\mathcal{D} \subseteq \Xi$, the empirical state-action-next-state distribution $\widehat{\xi}_T$ defined in* (15) *satisfies*

$$\begin{aligned}
-\inf_{\xi'\in\operatorname{int}\mathcal{D}} \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi) &\leq \liminf_{T\to\infty}\frac{1}{T}\log\mathbb{P}_\xi\left(\widehat{\xi}_T \in \mathcal{D}\right) \\
&\leq \limsup_{T\to\infty}\frac{1}{T}\log\mathbb{P}_\xi\left(\widehat{\xi}_T \in \mathcal{D}\right) \leq -\inf_{\xi'\in\mathcal{D}} \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi).
\end{aligned}$$

Proof Fix any $\xi \in \Xi_0$, and set $\theta = G(\xi)$. One readily verifies that $\xi = F(\theta)$, where F is the transformation defined in (16). Recall also that $\widehat{\xi}_T = F(\widehat{\theta}_T)$, where $\widehat{\xi}_T$ and $\widehat{\theta}_T$ are the empirical estimators for ξ and θ defined in (15) and (3), respectively. We already know from Theorem 2.2 that $\widehat{\theta}_T$ satisfies a large deviations principle with rate function D_{mc}. By the contraction principle (Dembo and Zeitouni, 2009, Theorem 4.2), which applies because F is continuous, $\widehat{\xi}_T = F(\widehat{\theta}_T)$ thus satisfies a large deviations principle with rate function $I(\xi', \xi) = \inf_{\theta' \in \Theta, F(\theta') = \xi'} \mathsf{D}_{\mathsf{mc}}(\theta' \| \theta)$.

It remains to be shown that $I(\xi', \xi) = \mathsf{D}_{\mathsf{mdp}}(\xi' \| \xi)$ for all $\xi \in \Xi_0$ and $\xi' \in \Xi$. To this end, set again $\theta = G(\xi)$, and use μ and μ' to denote the stationary state-action distributions corresponding to ξ and ξ', respectively. For any $\theta' \in \Theta$ with $F(\theta') = \xi'$ we then have

$$\begin{aligned}
\mathsf{D}_{\mathsf{mc}}(\theta' \| \theta) &= \mathsf{D}(\theta' \| \theta) - \mathsf{D}(\mu' \| \mu) \\
&\geq \mathsf{D}(G(F(\theta')) \| G(F(\theta))) - \mathsf{D}(\mu' \| \mu) \\
&= \mathsf{D}(G(\xi') \| G(\xi)) - \mathsf{D}(\mu' \| \mu) \\
&= \mathsf{D}_{\mathsf{mc}}(G(\xi') \| G(\xi)) = \mathsf{D}_{\mathsf{mdp}}(\xi' \| \xi)
\end{aligned}$$

where the first equality exploits the definition of D_{mc} (see Definition 2.1), and the inequality follows from the data processing inequality (Csiszár and Körner, 2011, Lemma 3.11). The second equality holds because $F(\theta') = \xi'$ by assumption and because $F(\theta) = F(G(\xi)) = \xi$ for any $\xi \in \Xi_0$. The third equality exploits again the definition of D_{mc}, and the fourth equality follows from Proposition 2.7. Hence, the infimum in the definition of $I(\xi', \xi)$ is attained by $\theta' = G(\xi') \in \Theta$. This in turn implies that $I(\xi', \xi) = \mathsf{D}_{\mathsf{mdp}}(\xi' \| \xi)$. This observation completes the proof. ■

The next corollary akin to Corollary 2.3 establishes a finite-sample version of Theorem 2.8.

Corollary 2.9 (Finite-sample version of Theorem 2.8) *For all $\xi \in \Xi_0$ and Borel sets $\mathcal{D} \subseteq \Xi$, the empirical doublet distribution $\widehat{\xi}_T$ defined in* (15) *satisfies*

$$\frac{1}{T} \log \mathbb{P}_\xi \left(\widehat{\xi}_T \in \mathcal{D} \right) \leq \frac{1}{T} (\log T + \bar{c} + d^2 \log(T+1)) - \inf_{\xi' \in \mathcal{D}} \mathsf{D}_{\mathsf{mdp}}(\xi' \| \xi) \quad \forall T \in \mathbb{N},$$

where $\bar{c} > 0$ is a universal constant that depends only on ξ.

Proof Fix any $\xi \in \Xi_0$ and Borel set $\mathcal{D} \subseteq \Xi$. Set $\theta = G(\xi)$, and let F be the transformation defined in (16). As F is linear, $F^{-1}(\mathcal{D})$ is a Borel subset of Θ. As $\widehat{\xi}_T = F(\widehat{\theta}_T)$, we may conclude that

$$\begin{aligned}
\frac{1}{T} \log \mathbb{P}_\xi \left(\widehat{\xi}_T \in \mathcal{D} \right) &= \frac{1}{T} \log \mathbb{P}_\xi \left(\widehat{\theta}_T \in F^{-1}(\mathcal{D}) \right) \\
&\leq \frac{1}{T} (\log T + \bar{c} + d^2 \log(T+1)) - \inf_{\theta' \in F^{-1}(\mathcal{D})} \mathsf{D}_{\mathsf{mc}}(\theta' \| \theta) \\
&= \frac{1}{T} (\log T + \bar{c} + d^2 \log(T+1)) - \inf_{\xi' \in \mathcal{D}} \inf_{\theta' \in \Theta, F(\theta') = \xi'} \mathsf{D}_{\mathsf{mc}}(\theta' \| \theta) \\
&= \frac{1}{T} (\log T + \bar{c} + d^2 \log(T+1)) - \inf_{\xi' \in \mathcal{D}} \mathsf{D}_{\mathsf{mdp}}(\xi' \| \xi) \quad \forall T \in \mathbb{N},
\end{aligned}$$

where the inequality follows from Corollary 2.3, and the last equality uses Proposition 2.7. ■

The conditional relative entropy $\mathsf{D}_{\mathsf{mdp}}$ will be instrumental for modeling distribution shifts. Besides representing the rate function of a large deviations principle, it inherits many useful properties from D_{mc} including level compactness, radial monotonicity and coercivity. The properties of D_{mc} are established in (Sutter et al., 2024, Proposition 5.1). The corresponding properties of $\mathsf{D}_{\mathsf{mdp}}$ can be established similarly by adapting the proofs for D_{mc} in the obvious way.

Lemma 2.10 (Properties of $\mathsf{D}_{\mathsf{mdp}}$) *The following hold.*

(i) ***Lower semicontinuity.*** $\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi)$ *is lower semicontinuous on* $\Xi \times \Xi$*;*

(ii) ***Level-compactness.*** $\{(\xi', \xi) \in \Xi \times \Xi : \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi) \leq \rho\}$ *is compact for every* $\rho \geq 0$*.*

(iii) ***Coerciveness.*** $\lim_{\zeta \in \Xi, \zeta \to \xi} \mathsf{D}_{\mathsf{mdp}}(\xi'\|\zeta) = \infty$ *for all* $\xi' \in \Xi_0$ *and* $\xi \in \Xi$ *with* $\operatorname{supp}(\xi') \not\subseteq \operatorname{supp}(\xi)$*.*

(iv) ***Radial monotonicity in*** ξ*.* $\operatorname{cl}\{\xi \in \Xi_0 : \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi) < \rho\} = \{\xi \in \Xi : \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi) \leq \rho\}$ *for all* $\xi' \in \Xi$ *and* $\rho > 0$*;*

(v) ***Continuity of the sublevel set mapping.*** *The set-valued mapping* $\Gamma : \Xi \rightrightarrows \Xi$ *defined through* $\Gamma(\xi') = \{\xi \in \Xi : \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi) \leq \rho\}$ *is continuous in* $\xi' \in \Xi$ *for every* $\rho > 0$*.*

(vi) ***Convexity in*** ξ'***.*** $\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi)$ *is convex in* $\xi' \in \Xi$ *for every fixed* $\xi \in \Xi$*.*

Proof Assertion (i) follows from the definition of $\mathsf{D}_{\mathsf{mdp}}$. Assertions (ii), (iv), and (vi) can be shown by adapting the proof of (Sutter et al., 2024, Proposition 5.1) from D_{mc} to $\mathsf{D}_{\mathsf{mdp}}$. Assertion (iii) holds because $\lim_{p \to 0} p \log(p/q) = 0$ for any $q \geq 0$ and $\lim_{p \to 0} q \log(q/p) = \infty$ for any $p > 0$. Assertion (v) can be shown by adapting the proof of (Sutter et al., 2024, Proposition 3.1) from D_{mc} to $\mathsf{D}_{\mathsf{mdp}}$. ■

3 Distributionally Robust Off-Policy Evaluation

Fix any MDP $(\mathcal{S}, \mathcal{A}, Q, r, \eta)$ of the type studied in Section 2.2, and assume that $Q \in \mathcal{Q}_0$. In addition, fix any stationary policy $\pi \in \Pi_0$. Such policies are called *exploratory* because $\pi > 0$. We know from Section 2.2 that the state-action pairs of the MDP follow an irreducible Markov chain with a transition probability matrix of the form (11). By the Markov law of large numbers (Norris, 1998, Theorem 1.7.6), the state-action-next-state distribution $\xi \in \Xi_0$ of this Markov chain satisfies

$$\xi(s, a, s') = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\pi,Q} \left(S_t = s,\ A_t = a,\ S_{t+1} = s' \right) \quad \forall s, s' \in \mathcal{S},\ a \in \mathcal{A}, \tag{19}$$

where $\mathbb{P}_{\pi,Q}$ is defined as in Section 2.2. In addition, ξ is related to the stationary state-action distribution μ through (12). We define the long-run average reward generated by π under Q as

$$V(\xi) = \lim_{T \to \infty} \mathbb{E}_{\pi,Q} \left[\frac{1}{T} \sum_{t=1}^{T} r(S_t, A_t) \right].$$

Note that $V(\xi)$ is independent of η and depends on π and Q only indirectly through ξ because

$$V(\xi) = \lim_{T \to \infty} \sum_{s,s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\pi,Q} \left(S_t = s,\ A_t = a,\ S_{t+1} = s' \right) = \sum_{s,s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \xi(s, a, s').$$

This shows in particular that $V(\xi)$ is linear and thus continuous in ξ. Note also that since $\xi \in \Xi_0$, we have that $V(\xi) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r(S_t, A_t)$ $\mathbb{P}_\xi$-almost surely by virtue of the Markov law of large numbers (Norris, 1998, Theorem 1.7.6). We express the long-run average reward $V(\xi)$ as a function of ξ instead of μ because ξ contains full information about π and Q, whereas μ does not.

Assume now that Q is *unknown*, which is a standard assumption in reinforcement learning. In contrast, the reward function r and the initial state distribution η are *known*. We also distinguish an *unknown* behavioral policy $\pi_0 \in \Pi_0$ and a *known* evaluation policy $\pi \in \Pi_0$. In addition, we refer to the state-action-next-state distributions associated with π_0 and π as $\xi_0 \in \Xi_0$ and $\xi \in \Xi$, respectively. Note that both ξ_0 and ξ are unobservable because they depend on the unknown transition kernel Q.

In the following we assume to have access to a single state-action trajectory $X_1, \ldots, X_T$ generated under the behavioral policy π_0. This trajectory provides information about the unknown transition kernel Q and can thus be used to estimate the reward function at ξ_0 as well as at ξ. We thus distinguish two fundamental estimation problems. The *on-policy evaluation problem* asks for an estimate of $V(\xi_0)$, that is, the long-run average reward of the behavioral policy π_0 that generates the data. In contrast, the *off-policy evaluation problem* asks for an estimate of $V(\xi)$, that is, the long-run average reward of the evaluation policy π. Note that there is *no* data generated under π. The off-policy evaluation problem can be viewed as an estimation problem with a distribution shift because we aim to estimate the expected reward under $\mathbb{P}_\xi$ from data generated under $\mathbb{P}_{\xi_0}$.

We will see below that both ξ as well as ξ_0 admit asymptotically consistent estimators. A naïve solution of the on-policy evaluation problem would be to approximate $V(\xi_0)$ by the plug-in estimator $V(\widehat{\xi}_T)$, where $\widehat{\xi}_T$ is the empirical distribution defined in (15). By (Norris, 1998, Theorem 1.7.6), $\widehat{\xi}_T$ converges almost surely to ξ_0 as T grows because the Markov chain of state-action pairs is irreducible under π_0. Hence, $V(\widehat{\xi}_T)$ converges $\mathbb{P}_{\xi_0}$-almost surely to $V(\xi_0)$. However, we will later see that $V(\widehat{\xi}_T)$ is likely to overestimate the true average reward—especially at small sample sizes. Also, it does not admit a natural generalization to off-policy evaluation. An important ingredient for solving the off-policy evaluation problem is the following distribution shift function.

Definition 3.1 (Distribution shift function) *For every exploratory evaluation policy $\pi \in \Pi_0$, the distribution shift function $f_\pi : \Xi_0 \to \Xi$ satisfies $f_\pi(\xi_0) = \xi$, where ξ is the state-action-next-state distribution corresponding to (π, Q) and where Q is the transition kernel induced by ξ_0 through* (13a).

Note that f_π is well-defined because the evaluation policy $\pi \in \Pi_0$ is given and $Q \in \mathcal{Q}_0$ is determined by $\xi_0 \in \Xi_0$ via (13a) and because we know from Section 2.2 that (π, Q) induces a unique $\xi \in \Xi_0$. Using the distribution shift function, we can recast the long-run average reward of the evaluation policy as $V(\xi) = V(f_\pi(\xi_0))$. A naïve solution to the off-policy evaluation problem would therefore be to approximate $V(\xi)$ by the plug-in estimator $V(f_\pi(\widehat{\xi}_T))$. This simply amounts to replacing the unknown transition kernel Q with its maximum likelihood estimator, which is obtained by substituting $\widehat{\xi}_T$ into (13a). In the discounted reward setting this approach is sometimes referred to as the *direct method*. We emphasize that if T is small, then $\widehat{\xi}_T$ may adopt values in $\Xi \backslash \Xi_0$, and in these cases the direct estimator $V(f_\pi(\widehat{\xi}_T))$ is undefined. As $\widehat{\xi}_T$ converges almost surely to $\xi_0 \in \Xi_0$, however, $V(f_\pi(\widehat{\xi}_T))$ is eventually well-defined for all sufficiently large T. Moreover, $V(f_\pi(\widehat{\xi}_T))$ converges almost surely to $V(\xi)$ because V and f_π are continuous.

Lemma 3.2 (Continuity of f_π) *The distribution shift function f_π is continuous on Ξ_0.*

Proof By Definition 3.1, the distribution shift function satisfies $f_\pi(\xi_0) = \xi$, and (12) implies that

$$\xi(s, a, s') = Q(s'|s, a)\mu(s, a) \quad \forall s, s' \in \mathcal{S},\ a, a' \in \mathcal{A}.$$

By (13a), the transition kernel Q is a rational and therefore continuous function of $\xi_0 \in \Xi_0$. Hence, the transition probability matrix P defined via (11) is also continuous in ξ_0. By the Perron-Frobenius theorem, the stationary distribution μ is the unique positive solution of the linear equations

$$\sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu(s,a) = 1 \quad \text{and} \quad \mu(s',a') = \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu(s,a)P((s,a),(s',a')) \quad \forall s' \in \mathcal{S},\ a' \in \mathcal{A}.$$

Therefore it inherits continuity in ξ_0 from P. In summary, these insights imply that $f_\pi(\xi_0) = \xi$ is continuous in ξ_0 throughout Ξ_0. Hence, the claim follows. ■

The following extension of Lemma 3.2 will be useful in Section 4.

Corollary 3.3 (Continuity of f_π) *The function $f_\pi(\xi_0)$ is continuous in $(\pi, \xi_0) \in \Pi_0 \times \Xi_0$.*

Proof By (13a), Q is independent of π, and by (11), P is continuous in (π, ξ_0). The continuity of $f_\pi(\xi_0)$ can thus be shown by repeating the proof of Lemma 3.2 with obvious modifications. ■

Lemma 3.2 implies via the Markov law of large numbers (Norris, 1998, Theorem 1.7.6) that the direct estimator $V(f_\pi(\widehat{\xi}_T))$ converges $\mathbb{P}_{\xi_0}$-almost surely to $V(f_\pi(\xi_0)) = V(\xi)$ and is therefore asymptotically consistent. We now introduce an alternative—distributionally robust—estimator $V_\rho^\pi(\widehat{\xi}_T)$ for $V(\xi)$, which is defined through the worst-case value function $V_\rho^\pi : \Xi \to \mathbb{R}$ with

$$V_\rho^\pi(\xi') = \inf_{\xi_0\in\Xi_0} \big\{V(f_\pi(\xi_0)) : \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi_0) \le \rho\big\} \quad \forall \xi' \in \Xi. \tag{20}$$

By construction, the optimization problem in (20) seeks the worst-case reward of the evaluation policy π with respect to all state-action-next-state distributions ξ_0 close to a realization $\xi' \in \Xi$ of the empirical estimator $\widehat{\xi}_T$. Here, proximity between ξ' and ξ_0 is measured by the conditional relative entropy for MDPs introduced in Definition 2.6. Hence, the feasible set of problem (20) can be viewed as a conditional relative entropy ball of radius $\rho \ge 0$ in the space of state-action-next-state distributions. If $\rho = 0$, then the distributioinally robust estimator $V_\rho^\pi(\widehat{\xi}_T)$ collapses to the direct estimator $V(f_\pi(\widehat{\xi}_T))$, which is ill-defined unless $\widehat{\xi}_T \in \Xi_0$. If $\rho > 0$, on the other hand, then problem (20) is guaranteed to be feasible for any $\xi' \in \Xi$. Indeed, $\{\xi \in \Xi : \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi) \le \rho\}$ is non-empty because $\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi') = 0$. This implies via the radial monotonicity established in Lemma 2.10 (iv) that problem (20) is feasible. Hence, $V_\rho^\pi(\widehat{\xi}_T)$ is well-defined for all $\rho > 0$.

In addition, if $\rho > 0$, then the minimum in (20) is attained whenever $\xi' \in \Xi_0$. Indeed, the objective function $V(f_\pi(\xi_0))$ is continuous on Ξ_0 thanks to Lemma 3.2, and the feasible set satisfies

$$\big\{\xi_0 \in \Xi_0 : \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi_0) \le \rho\big\} = \big\{\xi_0 \in \Xi : \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi_0) \le \rho\big\}. \tag{21}$$

The equality holds because $\xi' \in \Xi_0$, which implies via Lemma 2.10 (iii) that $\lim_{\zeta\in\Xi,\zeta\to\xi_0} \mathsf{D}_{\mathsf{mdp}}(\xi'\|\zeta) = \infty$ for all $\xi_0 \in \Xi\backslash\Xi_0$. Hence the feasible set is compact by virtue of Lemma 2.10 (ii), and the infimum in (20) is attained by the Weierstrass extreme value theorem. From now on we assume that $\rho > 0$.

In the remainder of this section we will show that the proposed distributionally robust estimator $V_\rho^\pi(\widehat{\xi}_T)$ for $V(f_\pi(\xi_0))$ enjoys rigorous finite-sample and asymptotic consistency guarantees. In addition, we show that $V_\rho^\pi(\widehat{\xi}_T)$ is, in a precise sense, the least conservative estimator whose

out-of-sample disappointment decays exponentially at rate ρ. Throughout this section, we restrict attention to estimators of the form $\widehat{V}^\pi(\widehat{\xi}_T)$, where $\widehat{V}^\pi : \Xi \to \mathbb{R}$ is an arbitrary lower semicontinuous function. We refer to the probability $\mathbb{P}_{\xi_0}(V(f_\pi(\xi_0)) < \widehat{V}^\pi(\widehat{\xi}_T))$ as the *out-of-sample disappointment* of the estimator $\widehat{V}^\pi(\widehat{\xi}_T)$ under the model $\xi_0 \in \Xi_0$. It quantifies the probability that the actual expected reward of the evaluation policy is strictly smaller than the reward predicted by the estimator. If the out-of-sample disappointment is large, then $\widehat{V}^\pi(\widehat{\xi}_T)$ *overestimates* the expected reward $V(f_\pi(\xi_0))$ with high probability. Hence, the estimator is overly optimistic, which may lead to disappointment. In the following, we will restrict our attention to estimators $\widehat{V}^\pi(\widehat{\xi}_T)$ that satisfy

$$\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\xi_0}\left(V(f_\pi(\xi_0)) < \widehat{V}^\pi(\widehat{\xi}_T)\right) \leq -\rho \quad \forall \xi_0 \in \Xi_0. \tag{22}$$

This condition ensures that the out-of-sample disappointment decays asymptotically as $e^{-\rho T + o(T)}$. It implies that $\widehat{V}^\pi(\widehat{\xi}_T)$ becomes an increasingly reliable lower confidence bound on $V(f_\pi(\xi_0))$ as the sample size grows. Underestimating the true expected reward $V(f_\pi(\xi_0))$ means to err on the side of caution. We will now demonstrate that the distributionally robust estimator $V_\rho^\pi(\widehat{\xi}_T)$ satisfies (22) and that it is in fact the least conservative estimator satisfying (22). As a preparation, the next proposition shows that the function V_ρ^π is lower semicontinuous on Ξ and continuous on Ξ_0.

Lemma 3.4 (Continuity properties of V_ρ^π) *For any fixed $\rho > 0$ and $\pi \in \Pi_0$, the distributionally robust value function $V_\rho^\pi(\xi')$ is lower semicontinuous in $\xi' \in \Xi$ and continuous in $\xi' \in \Xi_0$.*

Proof Define $\overline{V}(\xi_0) = \lim_{\delta\downarrow 0} \inf_{\zeta\in\Xi_0}\{V(f_\pi(\zeta)) : \|\zeta - \xi_0\| \leq \delta\}$ as the unique lower semicontinuous extension of $V(f_\pi(\xi_0))$ to Ξ. Indeed, $\overline{V}$ is lower semicontinuous by construction and coincides with $V(f_\pi(\xi_0))$ on Ξ_0 by virtue of Lemma 3.2. Define now the set-valued mapping $\Gamma(\xi') = \{\xi_0 \in \Xi : \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi_0) \leq \rho\}$, which is continuous thanks to Lemma 2.10 (v). Noting that $\Gamma(\xi') \neq \emptyset$ for every $\xi' \in \Xi$ because $\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi') = 0$, (Aliprantis and Border, 2006, Lemma 17.29) implies that

$$\varphi(\xi') = \min_{\xi_0\in\Gamma(\xi')} \overline{V}(\xi_0)$$

is lower semicontinuous on Ξ. In addition, we know from (21) that $\Gamma(\xi') \subseteq \Xi_0$ for every $\xi' \in \Xi_0$. As $\overline{V}(\xi_0) = V(f_\pi(\xi_0))$ is continuous on Ξ_0 by virtue of Lemma 3.2, we may then use (Aliprantis and Border, 2006, Theorem 17.31) to conclude that $\varphi(\xi')$ is continuous on Ξ_0. The claim thus follows if we can show that $V_\rho^\pi(\xi') = \varphi(\xi')$ for every $\xi' \in \Xi$. By (21), this identity clearly holds for every $\xi' \in \Xi_0$. Assume now that $\xi' \in \Xi\backslash\Xi_0$, and select any $\xi_0^\star \in \arg\min_{\xi_0\in\Gamma(\xi')} \overline{V}(\xi_0)$, which exists because $\Gamma(\xi')$ is non-empty and compact and because $\overline{V}(\xi_0)$ is lower semicontinuous. Thus, we find

$$\varphi(\xi') \leq V_\rho^\pi(\xi') \leq \lim_{\delta\downarrow 0} \inf_{\zeta\in\Xi_0} \{V(f_\pi(\zeta)) : \|\zeta - \xi_0^\star\| \leq \delta\} = \overline{V}(\xi_0^\star) = \varphi(\xi'),$$

where the first inequality holds because the feasible set of (20) is obtained by restricting $\Gamma(\xi')$ to Ξ_0, and the two equalities follow from the definitions of $\overline{V}$ and $\xi_0^\star$, respectively. Hence, the identity $V_\rho^\pi(\xi') = \varphi(\xi')$ holds indeed for all $\xi' \in \Xi$. This observation completes the proof. ■

The following extension of Lemma 3.4 will be useful in Section 4.

Corollary 3.5 (Continuity properties of V_ρ^π) *For any fixed $\rho > 0$, the distributionally robust value function $V_\rho^\pi(\xi')$ is continuous in π and lower semicontinuous in ξ' throughout $\Pi_\epsilon \times \Xi$.*

Proof The proof parallels that of Lemma 3.4 with obvious minor modifications (*e.g.*, Corollary 3.3 must be used instead of Lemma 3.2). ■

We are now ready to prove that the out-of-sample disappointment of the distributionally robust estimator $V_\rho^\pi(\widehat{\xi}_T)$ decays exponentially at rate ρ with the sample size T.

Theorem 3.6 (Out-of-sample disappointment of $V_\rho^\pi(\widehat{\xi}_T)$) *For every $\rho > 0$, $\pi \in \Pi_0$ and $\xi_0 \in \Xi_0$, the distributionally robust estimator $V_\rho^\pi(\widehat{\xi}_T)$ satisfies*

$$\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\xi_0} \left(V(f_\pi(\xi_0)) < V_\rho^\pi(\widehat{\xi}_T) \right) \leq -\rho.$$

Theorem 3.6 asserts that the out-of-sample disappointment of $V_\rho^\pi(\widehat{\xi}_T)$ decays as $e^{-\rho T + o(T)}$. Its proof is omitted because it is a direct consequence of the following finite-sample guarantee.

Theorem 3.7 (Finite-sample guarantee for $V_\rho^\pi(\widehat{\xi}_T)$) *For every $\xi_0 \in \Xi_0$ there is $\bar{c} > 0$ such that the distributionally robust estimator $V_\rho^\pi(\widehat{\xi}_T)$ satisfies the following for all $\rho > 0$, $\pi \in \Pi_0$ and $T \in \mathbb{N}$.*

$$\frac{1}{T} \log \mathbb{P}_{\xi_0} \left(V(f_\pi(\xi_0)) < V_\rho^\pi(\widehat{\xi}_T) \right) \leq \frac{1}{T} (\log(T) + \bar{c} + (SA)^2 \log(T+1)) - \rho$$

Proof Define the disappointment set $\mathcal{D} = \{\xi' \in \Xi : V(f_\pi(\xi_0)) < V_\rho^\pi(\xi')\}$, which is open because $V_\rho^\pi(\xi')$ is lower semicontinuous by virtue of Lemma 3.4. Then, we have

$$\begin{aligned}
\frac{1}{T} \log \mathbb{P}_{\xi_0} \left(V(f_\pi(\xi_0)) < V_\rho^\pi(\widehat{\xi}_T) \right) &= \frac{1}{T} \log \mathbb{P}_{\xi_0} \left(\widehat{\xi}_T \in \mathcal{D} \right) \\
&\leq \frac{(SA)^2}{T} \log(T+1) + \frac{1}{T} (\log(T) + \bar{c}) - \inf_{\xi' \in \mathcal{D}} \mathsf{D}_{\mathsf{mdp}}(\xi' \| \xi_0) \\
&\leq \frac{1}{T} ((SA)^2 \log(T+1) + \log(T) + \bar{c}) - \rho \quad \forall T \in \mathbb{N},
\end{aligned}$$

where the first inequality follows from Corollary 2.9, and the second inequality holds because $\mathsf{D}_{\mathsf{mdp}}(\xi' \| \xi_0) > \rho$ for any $\xi' \in \mathcal{D}$. Indeed, the definition of $V_\rho^\pi(\xi')$ in (20) readily implies that $V(f_\pi(\xi_0)) \geq V_\rho^\pi(\xi')$ whenever $\mathsf{D}_{\mathsf{mdp}}(\xi' \| \xi_0) \leq \rho$. By contraposition, this means that $\mathsf{D}_{\mathsf{mdp}}(\xi' \| \xi_0) > \rho$ whenever $V(f_\pi(\xi_0)) < V_\rho^\pi(\xi')$. Recall also from Corollary 2.9 that $\bar{c}$ depends only on ξ_0. ■

We now show that the distributionally robust predictor $V_\rho^\pi(\widehat{\xi}_T)$ is efficient in the sense that it represents the least conservative estimator whose out-of-sample disappointment decays at rate ρ.

Theorem 3.8 (Statistical efficiency of $V_\rho^\pi(\widehat{\xi}_T)$) *For every $\rho > 0$, $\pi \in \Pi_0$ and lower semicontinuous function $\widehat{V}^\pi : \Xi \to \mathbb{R}$ such that the corresponding reward estimator $\widehat{V}^\pi(\widehat{\xi}_T)$ satisfies the out-of-sample guarantee* (22)*, we have $V_\rho^\pi(\xi') \geq \widehat{V}^\pi(\xi')$ for all $\xi' \in \Xi$.*

Proof We first show that the claim holds for all $\xi' \in \Xi_0$. Assume for the sake of contradiction that there exists a lower semicontinuous function $\widehat{V}^\pi : \Xi \to \mathbb{R}$ with $\widehat{V}^\pi(\widehat{\xi}_T)$ satisfying (22) and an estimator realization $\xi_0' \in \Xi_0$ with $\widehat{V}^\pi(\xi_0') > V_\rho^\pi(\xi_0')$, and define $\epsilon_0 = \widehat{V}^\pi(\xi_0') - V_\rho^\pi(\xi_0') > 0$. Next, let $\xi_1^\star \in \Xi_0$ be a minimizer of problem (20) for $\xi' = \xi_0'$, which exists because $\xi_0' \in \Xi_0$. We thus

have $\mathsf{D}_{\mathsf{mdp}}(\xi_0'\|\xi_1^\star) \leq \rho$ by feasibility and $V_\rho^\pi(\xi_0') = V(f_\pi(\xi_1^\star))$ by optimality. The radial monotonicity of $\mathsf{D}_{\mathsf{mdp}}$ established in Lemma 2.10(iv) further guarantees that there exists a sequence $\{\xi_k^\star\}_{k\in\mathbb{N}}$ in Ξ_0 such that $\mathsf{D}_{\mathsf{mdp}}(\xi_0'\|\xi_k^\star) < \rho$ for all $k \in \mathbb{N}$ and $\lim_{k\to\infty} \xi_k^\star = \xi_1^\star$. In addition, the continuity of $V(f_\pi(\xi_1^\star))$ on Ξ_0 implies that there exists a model $\xi_0^\star \in \Xi_0$ with $V(f_\pi(\xi_0^\star)) < V(f_\pi(\xi_1^\star)) + \epsilon_0/2$ and $\rho_0 = \mathsf{D}_{\mathsf{mdp}}(\xi_0'\|\xi_0^\star) < \rho$. In summary, we have thus shown that

$$V_\rho^\pi(\xi_0') > V(f_\pi(\xi_1^\star)) - \frac{\epsilon_0}{2} > V(f_\pi(\xi_0^\star)) - \epsilon_0,$$

which allows us to conclude that

$$\widehat{V}^\pi(\xi_0') = V_\rho^\pi(\xi_0') + \epsilon_0 > V(f_\pi(\xi_0^\star)). \tag{23}$$

Next, we introduce the disappointment set $\mathcal{D} = \{\xi' \in \Xi : V(f_\pi(\xi_0^\star)) < \widehat{V}^\pi(\xi')\}$ and observe that $\xi_0' \in \mathcal{D}$. As $\widehat{V}^\pi$ is lower semicontinuous on Ξ by assumption, the set $\Xi\backslash\mathcal{D} = \{\xi' \in \Xi : V(f_\pi(\xi_0^\star)) \geq \widehat{V}^\pi(\xi')\}$ is closed, which in turn implies that $\mathcal{D}$ is open. We thus find

$$\begin{aligned}\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\xi_0^\star}\left(V(f_\pi(\xi_0^\star)) < \widehat{V}^\pi(\widehat{\xi}_T)\right) &= \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\xi_0^\star}\big(\widehat{\xi}_T \in \mathcal{D}\big) \\ &\geq -\inf_{\xi'\in\mathrm{int}\mathcal{D}} \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi_0^\star) \geq -\rho_0 > -\rho,\end{aligned}$$

where the first inequality follows from Theorem 2.8, and the second inequality holds because $\xi_0' \in \mathcal{D} = \mathrm{int}\mathcal{D}$ and $\mathsf{D}_{\mathsf{mdp}}(\xi_0'\|\xi_0^\star) = \rho_0$. Hence, the out-of-sample disappointment of $\widehat{V}^\pi(\widehat{\xi}_T)$ fails to decay at rate ρ, which contradicts our initial assumption. Thus, the claim follows.

Next, assume that $\xi' \in \Xi\backslash\Xi_0$. For any $n \in \mathbb{N}$, let $\xi_n' \in \Xi_0$ be a $1/n$-optimal solution of

$$\inf_{\zeta'\in\Xi_0} \{\widehat{V}^\pi(\zeta') : \|\zeta' - \xi'\|_2 \leq 1/n\}. \tag{24}$$

The feasibility of ξ_n' in (24) implies that $\|\xi_n' - \xi'\|_2 \leq 1/n$ such that $\lim_{n\to\infty} \xi_n' = \xi'$. Thus, we have

$$\begin{aligned}\widehat{V}^\pi(\xi') \leq \liminf_{n\to\infty} \widehat{V}^\pi(\xi_n') &= \lim_{n\to\infty} \inf_{\zeta'\in\Xi_0} \{\widehat{V}^\pi(\zeta') : \|\zeta' - \xi'\|_2 \leq 1/n\} \\ &\leq \lim_{n\to\infty} \inf_{\zeta'\in\Xi_0} \{V_\rho^\pi(\zeta') : \|\zeta' - \xi'\|_2 \leq 1/n\} \\ &= \lim_{n\to\infty} \inf_{\zeta',\zeta\in\Xi_0} \{V(f_\pi(\zeta)) : \mathsf{D}_{\mathsf{mdp}}(\zeta'\|\zeta) \leq \rho,\ \|\zeta' - \xi'\|_2 \leq 1/n\},\end{aligned} \tag{25}$$

where the first inequality follows from the assumed lower semicontinuity of $\widehat{V}^\pi$. The equality holds because ξ_n' is a $1/n$-optimal solution of (24), and the second inequality follows from the first part of the proof, where we have shown that $\widehat{V}^\pi(\zeta') \leq V_\rho^\pi(\zeta')$ for all $\zeta' \in \Xi_0$. The second equality, finally, exploits the definition of $V_\rho^\pi(\zeta')$. Fix now any $\epsilon > 0$, and let $\zeta_\epsilon \in \Xi_0$ be an ϵ-optimal solution of (20). Also, set $\zeta_{\epsilon,n}' = (1 - 1/(2n))\xi' + 1/(2n)\zeta_\epsilon$, and note that $\zeta_{\epsilon,n}' \in \Xi_0$ because $\zeta_{\epsilon,n}'$ represents a strict convex combination of $\xi' \in \Xi$ and $\zeta_\epsilon \in \Xi_0$. By construction, we thus have

$$\|\zeta_{\epsilon,n}' - \xi'\|_2 = \|\zeta_\epsilon - \xi'\|_2/(2n) \leq (\|\xi'\|_2 + \|\zeta_\epsilon\|_2)/(2n) \leq 1/n,$$

where the last inequality holds because $\zeta_\epsilon, \xi' \in \Delta(\mathcal{S}\times\mathcal{A}\times\mathcal{S})$. In addition, we find

$$\mathsf{D}_{\mathsf{mdp}}(\zeta_{\epsilon,n}'\|\zeta_\epsilon) \leq (1 - 1/(2n))\mathsf{D}_{\mathsf{mdp}}(\xi'\|\zeta_\epsilon) + \mathsf{D}_{\mathsf{mdp}}(\zeta_\epsilon\|\zeta_\epsilon)/(2n) \leq \rho,$$

where the first inequality follows from the convexity of the conditional relative entropy in its first argument (see Lemma 2.10 (vi)), and the second inequality holds because ζ_ϵ is feasible in (20). In summary, we have shown that $(\zeta'_{\epsilon,n}, \zeta_\epsilon)$ constitutes a feasible solution for the last minimization problem in (25). This observation readily implies that

$$\widehat{V}^\pi(\xi') \leq V(f_\pi(\zeta_\epsilon)) \leq V_\rho^\pi(\xi') + \epsilon.$$

where the second inequality holds because ζ_ϵ is an ϵ-optimal solution of (20). As this estimate holds for every $\epsilon > 0$, we may finally conclude that $\widehat{V}^\pi(\xi') \leq V_\rho^\pi(\xi')$. Hence, the claim follows. ■

Theorem 3.8 is inspired by (Sutter et al., 2024, Theorem 3.1), which establishes a similar result for general data-generating processes that are *not* affected by a distribution shift. On the technical level, Theorem 3.8 is also more general because it establishes the optimality of the distributionally robust estimator within the class of estimators of the form $\widehat{V}^\pi(\widehat{\xi}_T)$ induced by lower semicontinuous functions $\widehat{V}^\pi$, whereas (Sutter et al., 2024, Theorem 3.1) focuses exclusively on estimators induced by continuous functions. Theorem 3.8 also implies that $V_\rho^\pi(\widehat{\xi}_T) \geq \widehat{V}^\pi(\widehat{\xi}_T)$ for all $T \in \mathbb{N}$ because $V_\rho^\pi(\xi') \geq \widehat{V}^\pi(\xi')$ holds also for all $\xi' \in \Xi \backslash \Xi_0$. In contrast, the optimality result in (Sutter et al., 2024, Theorem 3.1) is asymptotic, that is, it holds only in the limit when T tends to infinity.

The following proposition inspired by (Li et al., 2021, Theorem 3) shows that the estimator V_ρ^π becomes asymptotically consistent if the radius ρ of the ambiguity set shrinks with T.

Proposition 3.9 (Asymptotic consistency of $V_\rho^\pi(\widehat{\xi}_T)$) *If $\{\rho_T\}_{T=1}^\infty$ is a sequence of non-negative radii with $\lim_{T\to\infty} \rho_T = 0$, then*

$$\lim_{T\to\infty} V_{\rho_T}^\pi(\widehat{\xi}_T) = V(f_\pi(\xi_0)) \quad \mathbb{P}_{\xi_0}\text{-}a.s. \quad \forall \xi_0 \in \Xi_0. \tag{26}$$

Proof Fix any $\xi_0 \in \Xi_0$ and any sample $\omega \in \Omega$ for which $\widehat{\xi}_T(\omega) \in \Xi$ converges to ξ_0 as T grows. For this particular sample ω, one can proceed as in the proof of (Li et al., 2021, Theorem 3) to show that problem (20) with $\xi' = \widehat{\xi}_T(\omega)$ is solvable for all sufficiently large T and that the corresponding minimizer $\xi_T^\star(\omega) \in \Xi_0$ converges to ξ_0 as T grows. This allows us to conclude that

$$\lim_{T\to\infty} V_{\rho_T}^\pi(\widehat{\xi}_T(\omega)) = \lim_{T\to\infty} V(f_\pi(\xi_T^\star(\omega))) = V(f_\pi(\xi_0)),$$

where the first equality follows from the construction of $\xi_T^\star(\omega)$, whereas the second equality holds because $V(f_\pi(\xi_0))$ is continuous on Ξ_0. The claim now follows because $\widehat{\xi}_T(\omega)$ converges to ξ_0 for $\mathbb{P}_{\xi_0}$-almost every sample ω. ■

4 Distributionally Robust Offline Policy Optimization

The distributionally robust estimator $V_\rho^\pi(\widehat{\xi}_T)$ analyzed in Section 3 solves the off-policy evaluation problem in a statistically efficient manner. Thus, it estimates the average reward of a fixed evaluation policy π based on data generated under a behavioral policy $\pi_0 \neq \pi$. We now use $V_\rho^\pi(\widehat{\xi}_T)$ as a building block for an offline policy optimization problem that learns an optimal policy π from data generated under π_0. To this end, we define $\Pi_\epsilon = \{\pi \in \Pi : \pi(a|s) \geq \epsilon \; \forall s \in \mathcal{S}, \, a \in \mathcal{A}\}$ as the family

of all exploratory policies that select any action with probability at least $\epsilon > 0$, where ϵ is sufficiently small to ensure that $\Pi_\epsilon \neq \emptyset$. We then introduce the distributionally robust policy estimator $\pi_\rho(\widehat{\xi}_T)$ corresponding to a given radius $\rho \geq 0$, where $\pi_\rho : \Xi \to \Pi_\epsilon$ is any Borel measurable function with

$$\pi_\rho(\xi') \in \operatorname*{argmax}_{\pi \in \Pi_\epsilon} V_\rho^\pi(\xi') \quad \forall \xi' \in \Xi. \tag{27}$$

Restricting Π to Π_ϵ in (27) simplifies the analysis of $\pi_\rho(\widehat{\xi}_T)$ at the expense of sacrificing flexibility. Note, however, that any policy $\pi \in \Pi$ gives rise to an ϵ-greedy policy $\pi_\epsilon \in \Pi_\epsilon$ defined through $\pi_\epsilon(a|s) = \epsilon/A + (1-\epsilon)\pi(a|s)$ for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, which approximates π arbitrarily closely.

In the remainder of this section, we investigate the statistical properties of the distributionally robust policy estimator $\pi_\rho(\widehat{\xi}_T)$. Specifically, we show that it is efficient in comparison to all estimators $\widehat{\pi}(\widehat{\xi}_T)$ induced by admissible policy functions $\widehat{\pi}$ in the sense of the following definition.

Definition 4.1 (Admissible policy function) *Given $\rho > 0$ and $\epsilon > 0$, a policy function $\widehat{\pi} : \Xi \to \Pi_\epsilon$ is called admissible if there exist value functions $\widehat{V}^\pi : \Xi \to \mathbb{R}$ parametrized by $\pi \in \Pi_\epsilon$ such that*

(i) $\widehat{V}^\pi(\xi')$ is continuous in π and lower semicontinuous in ξ' throughout $\Pi_\epsilon \times \Xi$;

(ii) $\widehat{\pi}(\xi')$ is Borel measurable in $\xi' \in \Xi$, and $\widehat{\pi}(\xi') \in \arg\max_{\pi \in \Pi_\epsilon} \widehat{V}^\pi(\xi')$ for every $\xi' \in \Xi_0$;

(iii) the out-of-sample disappointment of the policy $\widehat{\pi}(\widehat{\xi}_T)$ decays exponentially at rate ρ, that is,

$$\limsup_{T \to \infty} \frac{1}{T} \log \mathbb{P}_{\xi_0} \left(V(f_{\widehat{\pi}(\widehat{\xi}_T)}(\xi_0)) < \widehat{V}^{\widehat{\pi}(\widehat{\xi}_T)}(\widehat{\xi}_T) \right) \leq -\rho \quad \forall \xi_0 \in \Xi_0, \tag{28}$$

where f_π is the distribution shift function and $\widehat{\xi}_T$ the empirical estimator defined in (15).

We say that the admissibility of a policy function $\widehat{\pi}$ is certified by the corresponding family $\widehat{V}^\pi$, $\pi \in \Pi_\epsilon$, of value functions. We emphasize that the condition of admissibility is very weak. For example, every continuous policy function $\widehat{\pi} : \Xi \to \Pi_\varepsilon$ is admissible. Indeed, $\widehat{\pi}(\xi')$ is the unique maximizer of the continuous function $\widehat{V}^\pi(\xi') = -\|\pi - \widehat{\pi}(\xi')\|_2^2 - C$ across all $\pi \in \Pi_\epsilon$, where C is an arbitrary real constant. To ensure that (28) holds, it suffices to make C large. On a related note, the following lemma shows that for every family of value function $\widehat{V}^\pi$, $\pi \in \Pi_\epsilon$, satisfying condition (i) there exists a policy function $\widehat{\pi}$ satisfying condition (ii) of Definition 4.1.

Lemma 4.2 (Borel measurability of $\widehat{\pi}$) *If $\epsilon > 0$ and $\widehat{V}^\pi(\xi')$ is continuous in π and lower semicontinuous in ξ' throughout $\Pi_\epsilon \times \Xi$, then there is a Borel measurable function $\widehat{\pi} : \Xi \to \Pi_\epsilon$ with*

$$\widehat{\pi}(\xi') \in \arg\max_{\pi \in \Pi_\epsilon} \widehat{V}^\pi(\xi') \quad \forall \xi' \in \Xi.$$

Proof As Π_ϵ is compact and $\widehat{V}^\pi(\xi')$ is continuous in $\pi \in \Pi_\epsilon$ for every fixed $\xi' \in \Xi$, the optimization problem $\max_{\pi \in \Pi_\epsilon} \widehat{V}^\pi(\xi')$ is solvable. Hence, the solution mapping $\operatorname{argmax}_{\pi \in \Pi_\epsilon} \widehat{V}^\pi(\xi')$ is non-empty- and closed-valued throughout Ξ. Next, note that $\widehat{V}^\pi(\xi')$ is continuous in π and lower semicontinuous and thus Borel measurable in ξ' throughout $\Pi_\epsilon \times \Xi_0$. In addition, the feasible set Π_ϵ is constant and thus trivially Borel measurable in ξ throughout Ξ. By (Rockafellar and Wets, 2009, Examples 14.29 & 14.32), the function $\psi : \Pi_\epsilon \times \Xi \to [-\infty, \infty)$ defined through $\psi(\pi, \xi') = -\widehat{V}^\pi(\xi')$ if $\pi \in \Pi_\epsilon$; $= \infty$ otherwise constitutes a normal integrand in the sense of (Rockafellar and Wets, 2009,

Definition 14.27). Consequently, the solution mapping $\arg\max_{\pi\in\Pi_\epsilon} \widehat{V}^\pi(\xi')$ is Borel measurable in ξ' throughout Ξ by virtue of (Rockafellar and Wets, 2009, Theorem 14.37). By (Rockafellar and Wets, 2009, Corollary 14.6), the solution mapping thus admits a Borel measurable selector. ■

Together with Corollary 3.5, Lemma 4.2 implies that the distributionally robust policy function π_ρ is well-defined, that is, there is indeed a Borel measurable function π_ρ satisfying (27). In order to show that π_ρ is admissible in the sense of Definition 4.1, it thus suffices to verify condition (iii).

Theorem 4.3 (Out-of-sample disappointment of $\pi_\rho(\widehat{\xi}_T)$) *For every $\rho > 0$, $\epsilon > 0$ and $\xi_0 \in \Xi_0$, the distributionally robust policy estimator $\pi_\rho(\widehat{\xi}_T)$ satisfies*

$$\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\xi_0}\left(V(f_{\pi_\rho(\widehat{\xi}_T)}(\xi_0)) < V_\rho^{\pi_\rho(\widehat{\xi}_T)}(\widehat{\xi}_T)\right) \leq -\rho.$$

Theorem 4.3 is a direct consequence of Theorem 4.4 below. Thus, its proof is omitted.

Theorem 4.4 (Finite-sample guarantee for $\pi_\rho(\widehat{\xi}_T)$) *For every $\xi_0 \in \Xi_0$, there is $\bar{c} > 0$ such that the distributionally robust estimator $\pi_\rho(\widehat{\xi}_T)$ satisfies the following for all $\rho > 0$, $\epsilon > 0$ and $T \in \mathbb{N}$.*

$$\frac{1}{T} \log \mathbb{P}_{\xi_0}\left(V(f_{\pi_\rho(\widehat{\xi}_T)}(\xi_0)) < V_\rho^{\pi_\rho(\widehat{\xi}_T)}(\widehat{\xi}_T)\right) \leq \frac{1}{T}(\log(T) + \bar{c} + S^2 A \log(T+1)) - \rho$$

Proof Define the disappointment set $\mathcal{D}(\pi) = \{\xi' \in \Xi : V(f_\pi(\xi_0)) < V_\rho^\pi(\xi')\}$ corresponding to a fixed policy $\pi \in \Pi_\epsilon$, which is open because $V_\rho^\pi(\xi')$ is lower semicontinuous by virtue of Lemma 3.4. Next, observe that, for every fixed estimator realization $\xi' \in \Xi$, the following implications hold.

$$\begin{aligned} V(f_{\pi_\rho(\xi')}(\xi_0)) < V_\rho^{\pi_\rho(\xi')}(\xi') &\implies \exists \pi \in \Pi_\epsilon \text{ with } V(f_\pi(\xi_0)) < V_\rho^\pi(\xi') \\ &\implies \xi' \in \cup_{\pi\in\Pi_\epsilon} \mathcal{D}(\pi) \end{aligned}$$

This in turn implies that

$$\begin{aligned} \frac{1}{T} \log \mathbb{P}_{\xi_0}\left(V(f_{\pi_\rho(\widehat{\xi}_T)}(\xi_0)) < V_\rho^{\pi_\rho(\widehat{\xi}_T)}(\widehat{\xi}_T)\right) &\leq \frac{1}{T} \log \mathbb{P}_{\xi_0}\left(\widehat{\xi}_T \in \cup_{\pi\in\Pi_\epsilon} \mathcal{D}(\pi)\right) \\ &\leq \frac{1}{T}(\log T + \bar{c} + d^2 \log(T+1)) - \inf_{\xi'\in\cup_{\pi\in\Pi_\epsilon}\mathcal{D}(\pi)} \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi) \\ &\leq \frac{1}{T}((SA)^2 \log(T+1) + \log(T) + \bar{c}) - \rho \quad \forall T \in \mathbb{N}, \end{aligned}$$

where the second inequality follows from Corollary 2.9. The third equality can be justified as follows. The definition of V_ρ^π in (20) readily implies that $V(f_\pi(\xi_0)) \geq V_\rho^\pi(\xi')$ whenever $\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi_0) \leq \rho$. By contraposition, this means that $\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi_0) > \rho$ whenever $V(f_\pi(\xi_0)) < V_\rho^\pi(\xi')$ for some $\pi \in \Pi$. Hence, the infimum in the second line of the above expression is bounded below by ρ. ■

We are now ready to show that the distributionally robust policy estimator $\pi_\rho(\widehat{\xi}_T)$ represents the least conservative admissible policy estimator in a sense made precise in the following theorem.

Theorem 4.5 (Statistical efficiency of $\pi_\rho(\widehat{\xi}_T)$) *For every fixed $\rho > 0$ and $\epsilon > 0$, and for any admissible policy function $\widehat{\pi}$ and the corresponding family of value functions $\widehat{V}^\pi$, $\pi \in \Pi_\epsilon$, we have*

$$V_\rho^{\pi_\rho(\xi')}(\xi') \geq \widehat{V}^{\widehat{\pi}(\xi')}(\xi') \quad \forall \xi' \in \Xi. \tag{29}$$

Proof Select any $\xi' \in \Xi$. We have

$$\widehat{V}^{\widehat{\pi}(\xi')}(\xi') \leq V_\rho^{\widehat{\pi}(\xi')}(\xi') \leq V_\rho^{\pi_\rho(\xi')}(\xi'),$$

where the first inequality follows from Theorem 3.8 for $\pi = \widehat{\pi}(\xi')$, and the second inequality exploits the definition of π_ρ in (27). Thus, the claim follows. ■

By condition (ii) of Definition 4.1, at $\xi' = \widehat{\xi}_T$ the right hand side of (29) equals $\max_{\pi \in \Pi_\epsilon} \widehat{V}^\pi(\widehat{\xi}_T)$ and can thus be viewed as a generic estimator for the optimal value of the offline policy optimization problem. Condition (iii) of Definition 4.1 ensures that this estimator is conservative, that is, it overestimates the achievable expected reward of the policy $\widehat{\pi}(\widehat{\xi}_T)$ only with a small probability of at most $e^{-\rho T + o(T)}$. Thus, it provides a lower confidence bound on the optimal expected reward. Similarly, by (27), at $\xi' = \widehat{\xi}_T$ the left hand side of (29) equals $\max_{\pi \in \Pi_\epsilon} V_\rho^\pi(\widehat{\xi}_T)$, and Theorem 4.3 ensures that it provides a lower confidence bound with the same significance level. Theorem 4.5 thus asserts that the distributionally robust policy estimator $\pi_\rho(\widehat{\xi}_T)$ provides the least conservative lower confidence bound among all policy estimators with the same significance level $e^{-\rho T + o(T)}$.

5 Numerical Solution Schemes

The distributionally robust value and policy estimators introduced in Sections 3 and 4 are statistically efficient. It remains to be discussed how these estimators can be computed efficiently. To this end, we will adapt the actor-critic algorithm by Li et al. (2023, § 4), which was originally developed for robust MDPs with discounted cost criteria, to robust MDPs with average reward criteria.

5.1 Reparametrization

As a preparation, we first show that the robust policy evaluation problem in (20) can be reformulated as an optimization problem over transition kernels with a non-convex objective function and a convex feasible set. Thus, for any state-action-next-state distributions $\xi', \xi \in \Xi_0$ we define the corresponding transition kernels $Q', Q \in \mathcal{Q}_0$ and policies $\pi', \pi \in \Pi_0$ through (13a) and (13b), respectively. In addition, we let $\mu', \mu \in \Delta(\mathcal{S} \times \mathcal{A})$ be the corresponding stationary state-action distributions. The conditional relative entropy (17) between ξ and ξ' can then be equivalently expressed as

$$\begin{aligned}
\mathsf{D}_{\mathsf{mdp}}(\xi' \| \xi) &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mu'(s,a) \left(\log \frac{\mu'(s,a)}{\sum_{\tilde{a} \in \mathcal{A}} \mu'(s,\tilde{a})} - \log \frac{\mu(s,a)}{\sum_{\tilde{a} \in \mathcal{A}} \mu(s,\tilde{a})} \right) \\
&\quad + \sum_{s,s' \in \mathcal{S}} \sum_{a,a' \in \mathcal{A}} \xi'(s,a,s') \left(\log \frac{\xi'(s,a,s')}{\mu'(s,a)} - \log \frac{\xi(s,a,s')}{\mu(s,a)} \right) \\
&= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mu'(s,a) \mathsf{D}(\pi'(\cdot|s) \| \pi(\cdot|s)) + \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mu'(s,a) \mathsf{D}(Q'(\cdot|s,a) \| Q(\cdot|s,a)),
\end{aligned} \tag{30}$$

where the second equality follows from (13a) and (13b). We can use (30) to recast the distributionally robust value function (20) as the optimal value of a parametric minimization problem over Q.

Lemma 5.1 (Reformulation of $V_\rho^\pi(\xi')$) *For any fixed $\pi \in \Pi_0$, $\rho \geq 0$ and $\xi' \in \Xi_0$, we have*

$$V_\rho^\pi(\xi') = \min_{Q \in \mathcal{Q}_0} \left\{ \sum_{x \in \mathcal{X}} r(x)\mu_{\pi,Q}(x) : \sum_{x \in \mathcal{X}} \mu'(x)\mathsf{D}(Q'(\cdot|x)\|Q(\cdot|x)) \leq \rho \right\}, \tag{31}$$

where Q' is the transition kernel induced by ξ' via (13a) *and $\mu_{\pi,Q}$ is the unique positive solution of the equations $\sum_{x \in \mathcal{X}} \mu_{\pi,Q}(x) = 1$ and $\mu_{\pi,Q}(x) = \sum_{y \in \mathcal{X}} \mu_{\pi,Q}(y)P(y,x)$ with P defined as in* (11).

The stationary distribution $\mu_{\pi,Q}$ is well-defined thanks to the Perron-Frobenius theorem. Note that the minimization problem (31) is independent of the policy π' corresponding to ξ'. In the remainder of the paper we use $\mathcal{Q}_\rho(\xi')$ as a shorthand for the feasible set of (31).

Proof of Lemma 5.1 Recall that $V_\rho^\pi(\xi')$ is defined as the optimal value of problem (20). In the first part of the proof we show that the optimal value of (20) is at least as large as that of (31). To this end, suppose that $\xi_0 \in \Xi_0$ is a feasible solution of (20), and define $\xi = f_\pi(\xi_0)$, where f_π is as in Definition 3.1. Then, the transition kernel $Q \in \mathcal{Q}_0$ induced by ξ_0 through (13a) satisfies

$$\sum_{x \in \mathcal{X}} r(x)\mu_{\pi,Q}(x) = \sum_{s,s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s,a)\xi(s,a,s') = V(f_\pi(\xi_0)),$$

where the two equalities follow from (12) and from the definition of V at the beginning of Section 3, respectively. Thus, the objective function value of Q in (31) equals that of ξ_0 in (20). Next, let $\pi_0 \in \Pi_0$ be the policy induced by ξ_0 through (13b). By construction, we also have

$$\sum_{x \in \mathcal{X}} \mu'(x)\mathsf{D}(Q'(\cdot|x)\|Q(\cdot|x)) = \mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi_0) - \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mu'(s,a)\mathsf{D}(\pi'(\cdot|s)\|\pi_0(\cdot|s)) \leq \rho,$$

where the equality follows from (30), while the inequality holds because ξ_0 is feasible in (20) such that $\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi_0) \leq \rho$ and because the relative entropy is non-negative. Hence, Q is feasible in (31). In summary, these insights confirm that the optimal value of (20) is at least as large as that of (31).

In the second part of the proof we show that the optimal value of (20) is at most as large as that of (31). To this end, suppose that $Q \in \mathcal{Q}_0$ is feasible in (31). Defining $\xi_0 \in \Xi_0$ as the stationary state-action-next-state distribution corresponding to π' and Q and setting $\xi = f_\pi(\xi_0)$, we then find

$$V(f_\pi(\xi_0)) = \sum_{s,s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s,a)\xi(s,a,s') = \sum_{x \in \mathcal{X}} r(x)\mu_{\pi,Q}(x),$$

where the equalities follow again from the definition of V and from (12), respectively. Thus, the objective function value of ξ_0 in (20) equals that of Q in (31). By construction of ξ_0, we also have

$$\begin{aligned}\mathsf{D}_{\mathsf{mdp}}(\xi'\|\xi_0) &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mu'(s,a)\mathsf{D}(\pi'(\cdot|s)\|\pi'(\cdot|s)) + \sum_{x \in \mathcal{X}} \mu'(x)\mathsf{D}(Q'(\cdot|x)\|Q(\cdot|x)) \\ &= \sum_{x \in \mathcal{X}} \mu'(x)\mathsf{D}(Q'(\cdot|x)\|Q(\cdot|x)) \leq \rho,\end{aligned}$$

where the two equalities follow from (30) and from the trivial relation $\mathsf{D}(\pi'(\cdot|s)\|\pi'(\cdot|s)) = 0$, respectively, whereas the inequality holds because Q is feasible in (31). Hence, ξ is feasible in (20). In summary, these insights confirm that the optimal value of (20) is at most as large as that of (31).

The above reasoning implies that the optimal values of (20) and (31) match. Recalling that the minimum of (20) is attained because $\xi' \in \Xi_0$, the first part of the proof thus implies that the minimum of problem (31) is attained, too. This observation completes the proof. ■

Note that the feasible region of problem (31) is convex because $\mathcal{Q}_0$ is a convex set and the relative entropy is a convex function. However, the objective function of (31) is generically non-convex because the stationary distribution $\mu_{\pi,Q}$ fails to be convex in Q (Li et al., 2021, Remark 13). Note also that problem (31) evaluates the worst-case average reward of the policy π across all transition kernels Q in the non-rectangular uncertainty set $\mathcal{Q}_\rho(\xi')$ *without* a distribution shift.

Example 5.1 (Non-rectangularity of $\mathcal{Q}_\rho(\xi')$) *Assume that $\mathcal{S} = \{s_1, s_2\}$ and $\mathcal{A} = \{a\}$, and fix an arbitrary $\xi' \in \Xi_0$. In this case, the uncertainty set $\mathcal{Q}_\rho(\xi')$ simplifies to*

$$\mathcal{Q}_\rho(\xi') = \left\{Q \in \mathcal{Q}_0 : \mu'_{\mathcal{S}}(s_1)\mathsf{D}(Q'(\cdot|s_1,a)\|Q(\cdot|s_1,a)) + \mu'_{\mathcal{S}}(s_2)\mathsf{D}(Q'(\cdot|s_2,a)\|Q(\cdot|s_2,a)) \leq \rho\right\}.$$

Thus, any transition kernel $Q \in \mathcal{Q}_\rho(\xi')$ must respect the following inequality:

$$\mu'_{\mathcal{S}}(s_1)\mathsf{D}(Q'(\cdot|s_1,a)\|Q(\cdot|s_1,a)) \leq \rho - \mu'_{\mathcal{S}}(s_2)\mathsf{D}(Q'(\cdot|s_2,a)\|Q(\cdot|s_2,a)).$$

This inequality demonstrates that the permissible next-state distribution of $Q(\cdot|s_1,a)$ for state s_1 and action a is contingent upon $Q(\cdot|s_2,a)$ for state s_2 and action a. Such interdependence directly contradicts the definition of s-rectangularity (Wiesemann et al., 2013, § 2.1), which requires that the ambiguity set decompose into independent, state-specific components.

Robust policy evaluation problems (and thus robust MDPs) with non-rectangular uncertainty sets are generically intractable. Theorem 1 in (Wiesemann et al., 2013) shows that when the uncertainty set of the transition kernel Q is a convex polytope that does not satisfy any rectangularity conditions—specifically, it is neither (s,a)-rectangular, s-rectangular, nor r-rectangular—then the robust policy evaluation problem can be reduced to an integer feasibility problem and thus is NP-hard.

5.2 Actor-Critic Algorithm

Given an oracle that outputs approximate solutions for the robust policy evaluation problem in (31), the robust policy optimization problem (27) can be addressed with a variant of the actor-critic algorithm developed by Li et al. (2023) for robust MDPs with a *discounted* cost objective; see Algorithm 1. Throughout this section we use $V_Q^\pi = \sum_{x\in\mathcal{X}} r(x)\mu_{\pi,Q}(x)$ as shorthand for the long-run average reward of the policy $\pi \in \Pi_\epsilon$ under the transition kernel $Q \in \mathcal{Q}_0$.

In each iteration k, Algorithm 1 first computes a δ-optimal solution $Q^{(k)}$ of the robust policy evaluation problem (31) associated with the current policy $\pi^{(k)}$ (*critic*) and then applies a projected gradient step to find a new policy $\pi^{(k+1)}$ that locally improves the value function associated with the current transition kernel $Q^{(k)}$ (*actor*). The critic's subproblem can be solved with Algorithm 2 described in Section 5.2.2 below, for example, which outputs a δ-optimal solution of the robust policy evaluation problem with high probability. The actor's subproblem simply consists in computing a policy gradient and projecting a vector onto the simplex Π_ϵ as described in Section 5.2.1 below.

Algorithm 1 Actor-critic algorithm for solving the robust policy optimization problem (27)

Require: Iteration number K, step size $\eta > 0$, tolerance $\delta > 0$

1: Initialize $\pi^{(0)} \in \Pi_\epsilon$, $k = 0$
2: **while** $k \leq K$ **do**
3: *Critic*: Find $Q^{(k)} \in \mathcal{Q}_\rho(\xi')$ such that $V_{Q^{(k)}}^{\pi^{(k)}} \leq V_\rho^{\pi^{(k)}}(\xi') + \delta$
4: *Actor*: $\pi^{(k+1)} = \mathrm{Proj}_{\Pi_\epsilon}\left(\pi^{(k)} + \eta \nabla_\pi V_{Q^{(k)}}^{\pi^{(k)}}\right)$
5: $k \to k + 1$
6: **end while**

5.2.1 Actor

Projecting a vector onto Π_ϵ is a standard operation that admits highly efficient implementations; see, *e.g.*, (Wang and Carreira-Perpinán, 2013). Thus, the main computational burden of the actor's subproblem is associated with the computation of the policy gradient $\nabla_\pi V_Q^\pi$. In order to derive a concise formula for the policy gradient, it is useful to introduce a differential action-value function.

Definition 5.2 (Differential action-value function) *For any fixed policy $\pi \in \Pi_\epsilon$ and transition kernel $Q \in \mathcal{Q}_0$, the differential action-value function $H_Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined through*

$$H_Q^\pi(s,a) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}_{\pi,Q}} \left[\sum_{\tau=1}^{t} \left(r(s_\tau, a_\tau) - V_Q^\pi \right) \,\middle|\, s_1 = s,\, a_1 = a \right].$$

We emphasize that the limit in Definition 5.2 exists and is finite (Puterman, 2005, Section 8.2.1). Note also that, strictly speaking, the gradient $\nabla_\pi V_Q^\pi$ fails to exist because the function V_Q^π was only defined on $\Pi_0 \times \mathcal{Q}_0$ and because the interior of Π_0 is empty. Using (Li et al., 2021, Lemma 4), however, one can show that V_Q^π is given by a rational (and thus analytic) function of the matrix P defined as in (11). As P is linear in π, V_Q^π thus constitutes a rational function of π. One can also show that the denominator of this rational function is bounded away from zero on a neighborhood of Π_ϵ for any $\epsilon > 0$. In the following, we interpret $\nabla_\pi V_Q^\pi$ as the gradient of this analytic extension of V_Q^π, which is well-defined on Π_ϵ because the interior of any neighborhood of Π_ϵ covers Π_ϵ.

Lemma 5.3 (Policy gradient (Li et al., 2024b, Lemma 7)) *If $Q \in \mathcal{Q}_0$ and $\pi \in \Pi_\epsilon$, then*

$$\frac{\partial V_Q^\pi}{\partial \pi(a|s)} = \sum_{\tilde{a} \in \mathcal{A}} \mu_{\pi,Q}(s,\tilde{a}) H_Q^\pi(s,a) \quad \forall a \in \mathcal{A}, s \in \mathcal{S},$$

where $\mu_{\pi,Q}$ is defined as in Lemma 5.1.

Note that $\nabla_\pi V_Q^\pi$ is Lipschitz continuous in π throughout Π_ϵ thanks to (Li et al., 2021, Lemma 17). As Π_ϵ is compact, this readily implies that V_Q^π is also Lipschitz continuous in π throughout Π_ϵ. By using a variant of (Li et al., 2023, Theorem 4.5), one can thus show that Algorithm 1 converges to a global maximizer of the robust offline policy optimization problem in (27).

Theorem 5.4 (Convergence of Algorithm 1) *Assume that V_Q^π is L-Lipschitz and $\nabla_\pi V_Q^\pi$ is ℓ-Lipschitz in $\pi \in \Pi_\epsilon$ uniformly for all $Q \in \mathcal{Q}_\rho(\xi')$, and set the distribution mismatch coefficient to*

$$C = \max_{\pi,\pi' \in \Pi_\epsilon} \max_{Q \in \mathcal{Q}_\rho(\xi')} \max_{s \in \mathcal{S}} \frac{\mu_{\pi,Q}(s)}{\mu_{\pi',Q}(s)}.$$

If $\delta = \frac{L}{2}\sqrt{2S/K}$ *and* $\eta = \frac{1}{L}\sqrt{2S/K}$*, then the iterates* $\pi^{(k)}$ *of Algorithm 1 satisfy*

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_\rho^{\pi^{(k)}}(\xi') - \min_{\pi\in\Pi_\epsilon} V_\rho^\pi(\xi')\right) \leq \frac{(72S)^{1/4}(C\sqrt{2\ell LS} + \frac{L}{2}\sqrt{L/\ell})}{K^{1/4}}.$$

Theorem 5.4 shows that computing a δ-optimal solution for the robust offline policy optimization problem in (27) requires at most $K = \mathcal{O}(\delta^{-4})$ iterations. We point out that the convergence rate of Algorithm 1 is of the order $\mathcal{O}(K^{-1/4})$ whenever $\eta = \mathcal{O}(K^{-1/2})$, and thus it is not necessary to know the Lipschitz constant L in practice. Setting $\eta = \frac{1}{L}\sqrt{2S/K}$ leads to theoretically optimal constants.

Proof of Theorem 5.4 Note that $\mu_{\pi,Q} > 0$ for every $\pi \in \Pi_\epsilon$ and $Q \in \mathcal{Q}_\rho(\xi') \subseteq \mathcal{Q}_0$ thanks to Lemma 2.5. In addition, $\mu_{\pi,Q}$ is jointly continuous in π and Q by (Li et al., 2021, Lemma 17). Weierstrass' extreme value theorem thus implies that the maxima in the definition of C are attained such that C is finite and strictly positive. Next, define the Moreau envelope $\Phi_\gamma : \mathbb{R}^{S\times A} \to \mathbb{R}$ of the distributionally robust value function V_ρ^π corresponding to the smoothness parameter $\gamma > 0$ by

$$\Phi_\gamma(\pi) = \min_{\pi'\in\Pi_\epsilon} V_\rho^{\pi'}(\xi') + \frac{1}{2\gamma}\|\pi' - \pi\|_{\mathbf{F}}^2,$$

where $\|\cdot\|_{\mathbf{F}}$ stands for the Frobenius norm. We then have

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\left(V_\rho^{\pi^{(k)}}(\xi') - \min_{\pi\in\Pi_\epsilon} V_\rho^\pi(\xi')\right) &\leq \frac{(C\sqrt{2S} + L/(2\ell))}{K}\sum_{k=0}^{K-1}\|\nabla\Phi_{1/(2\ell)}(\pi^{(k)})\|_{\mathbf{F}} \\
&\leq \frac{(C\sqrt{2S} + L/(2\ell))}{\sqrt{K}}\sqrt{\sum_{k=0}^{K-1}\|\nabla\Phi_{1/(2\ell)}(\pi^{(k)})\|_{\mathbf{F}}^2} \\
&\leq \frac{(C\sqrt{2S} + L/(2\ell))(72S)^{1/4}(\ell L)^{1/2}}{K^{1/4}},
\end{aligned}$$

where the first inequality follows from (Li et al., 2023, Lemma 4.4), which applies because V_Q^π is L-Lipschitz and $\nabla_\pi V_Q^\pi$ is ℓ-Lipschitz in $\pi \in \Pi_\epsilon$ uniformly for all $Q \in \mathcal{Q}_\rho(\xi')$. The second and the third inequalities exploit Jensen's inequality and (Li et al., 2023, Lemma 4.3), respectively. ■

5.2.2 Critic

We now show that the critic's robust policy evaluation problem (31) can be solved approximately with a randomized policy gradient method that offers global convergence guarantees; see Algorithm 2.

Throughout this section we assume that the uncertainty set admits a reparametrization of the form $\mathcal{Q}_\rho(\xi') = \{Q^\lambda : \lambda \in \Lambda\}$, where $\Lambda \subseteq \mathbb{R}^q$ is a solid parameter set, and Q^λ is an affine function. As Λ is solid, its linear span coincides with the ambient space $\mathbb{R}^q$. A reparametrization of the uncertainty set with these properties and $q = A(S-1)$ exists; see (Wiesemann et al., 2013, § 5).

In each iteration Algorithm 2 applies a projected stochastic gradient step. The projection onto the convex set Λ is a standard operation that admits efficient implementations (Usmanova et al., 2021). We now address the computation of the adversary's policy gradient. To this end, note that the gradient $\nabla_Q V_Q^\pi$ fails to exist because V_Q^π was only defined on $\Pi_0 \times \mathcal{Q}_0$ and because the interior of $\mathcal{Q}_0$ is empty. Using a similar reasoning as in Section 5.2.1 and recalling that $\xi' \in \Xi_0$, however,

Algorithm 2 Projected Langevin dynamics for solving the robust policy evaluation problem (31)

Require: Iteration number $M \in \mathbb{N}$, step size $\eta > 0$, Gibbs parameter $\beta > 1$

1: Initialize $\lambda^{(0)} \in \Lambda$, $m = 0$
2: **while** $m \leq M - 1$ **do**
3: Sample $w_{m+1} \sim \mathcal{N}(0, I_q)$
4: Find $\lambda^{(m+1)} = \mathrm{Proj}_\Lambda \left(\lambda^{(m)} - \eta \, \nabla_\lambda V^\pi_{Q^\lambda} \Big|_{\lambda=\lambda^{(m)}} + \sqrt{2\eta/\beta} w_{m+1} \right)$
5: $m \to m + 1$
6: **end while**

V^π_Q can be extended to an analytic function of Q on a neighborhood of $\mathcal{Q}_\rho(\xi')$. In the following, we interpret $\nabla_Q V^\pi_Q$ as the gradient of this analytic extension of V^π_Q. In order to derive a concise formula for the adversary's policy gradient, we introduce a differential action-next-state value function.

Definition 5.5 (Differential action-next-state value function) *For any fixed policy $\pi \in \Pi_\epsilon$ and transition kernel $Q \in \mathcal{Q}_0$, the differential action-next-state value function $J^\pi_Q : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is defined through*

$$J^\pi_Q(s, a, s') = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}_{\pi,Q}} \left[\sum_{\tau=1}^{t} \left(r(s_\tau, a_\tau) - V^\pi_Q \right) | s_1 = s, a_1 = a, s_2 = s' \right].$$

One can show that the limit in Definition 5.5 exists and is finite (Puterman, 2005, Section 8.2.1). The next lemma provides a formula for the adversary's policy gradient. Its proof widely parallels that of (Li et al., 2023, Lemma 1) and is thus omitted.

Lemma 5.6 (Adversary's policy gradient) *For any $\pi \in \Pi_0$ and $\lambda \in \Lambda$, we have*

$$\nabla_\lambda V^\pi_{Q^\lambda} = \sum_{s,s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mu_{\pi,Q^\lambda}(s, a) J^\pi_{Q^\lambda}(s, a, s') \nabla_\lambda Q^\lambda(s'|s, a).$$

By adapting (Li et al., 2023, Theorem 3) to our setting, we can now show that Algorithm 2 converges in expectation to a global minimizer of the robust policy evaluation problem in (20).

Theorem 5.7 (Convergence of Algorithm 2) *If $\delta > 0$, $\eta < 1/2$, $\pi \in \Pi_0$, and $\gamma \in (0, 1)$, there exist universal constants $a > 4$, $b > 1$ and $c_1, c_2, c_3 > 0$ such that for all $\beta \geq c_1^{-1} \left(2q/(c_1(1-\gamma)\delta e)\right)^{1/\gamma}$ and $M \geq \max\{4, c_2 \exp(c_3 q^b)/\delta^a\}$ the distribution ν_M of the output $\lambda^{(M)}$ of Algorithm 2 satisfies*

$$\mathbb{E}_{\lambda\sim\nu_M}[V^\pi_{Q^\lambda}] \leq V^\pi_\rho(\xi') + \delta.$$

Proof Recall that Λ is a solid convex body. In addition, note that $\nabla_Q V^\pi_Q$ is Lipschitz continuous in $Q \in \mathcal{Q}_\rho(\xi')$ thanks to (Li et al., 2021, Lemma 17). As Q_λ is affine in λ and $\mathcal{Q}_\rho(\xi') = \{Q^\lambda : \lambda \in \Lambda\}$, we may then use the chain rule to deduce that $\nabla_\lambda V^\pi_{Q^\lambda}$ is Lipschitz continuous in $\lambda \in \Lambda$. Thus, the claim follows directly from (Li et al., 2023, Theorem 3). ■

Theorem 5.7 implies that the number of iterations M required by Algorithm 2 to compute a δ-optimal solution for the robust policy evaluation problem (20) grows exponentially with both

the dimension q of the parameter λ and the number of accuracy digits $\log(1/\delta)$. This curse of dimensionality is expected in view of the hardness result by Wiesemann et al. (2013, Theorem 1). However, Algorithm 2 is conceptually simple and guarantees convergence to a global minimum, even though the uncertainty set fails to be rectangular. Moreover, by leveraging Markov's inequality, we can transform the convergence-in-expectation result from Theorem 5.7 into a probabilistic bound.

Corollary 5.8 (Probabilistic suboptimality guarantee) *If all assumptions of Theorem 5.7 hold, then we have* $\mathbb{P}_{\lambda\sim\nu_M}[V^\pi_{Q^\lambda} > V^\pi_\rho(\xi') - \delta/\gamma] \geq 1-\gamma$ *for all* $\gamma \in (0,1)$.

6 Numerical Results

We now assess the out-of-sample properties of the proposed distributionally robust value and policy estimators in two numerical experiments. The first experiment revolves around a stochastic GridWorld system and tests the distributionally robust value estimator from Section 3. The second experiment tests the distributionally robust policy estimator from Section 4 in the context of a standard machine replacement problem. All experiments are implemented in Python, and the code for reproducing all numerical results is available from `https://github.com/mengmenglior/offline-rl`.

6.1 Off-Policy Evaluation: GridWorld System

The first experiment is built around a GridWorld problem ubiquitous in reinforcement learning (Sutton and Barto, 2018). The state space $\mathcal{S}$ consists of the 25 cells of a 5×5 grid, and the action space $\mathcal{A} = \{0, 1, 2, 3\}$ includes the 4 directions "up," "down," "left" and "right." An agent aims to reach the Goal State in cell 1 (top left) while avoiding the Bad State in cell 25 (bottom right). Selecting action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ moves the agent to $s' \in \mathcal{S}$ with probability $Q(s'|s,a)$. The agent receives a reward of 0 in the Goal State, -5 in the Bad State, and -1.5 elsewhere. The initial state s_0 follows the uniform distribution on $\mathcal{S}$, which we denote by ρ. The transition probabilities are defined as follows. Let $\mathcal{S}(s) \subseteq \mathcal{S}$ be the set of cells adjacent to s. If s' is the adjacent cell in direction a, then $Q(s'|s,a) = 0.7$; if s' is any other adjacent cell, $Q(s'|s,a) = 0.1$; if $s' = s$, $Q(s'|s,a) = 1 - \sum_{s''\in\mathcal{S}(s)} Q(s''|s,a)$; otherwise, $Q(s'|s,a) = 0$. These rules apply even if no adjacent cell exists in direction a. The agent observes a state-action trajectory $\{(S_t, A_t)\}_{t=1}^T$ generated under the behavioral policy $\pi_0 \in \Pi_0$, which selects each action $a \in \mathcal{A}$ with probability $0.9 \times 0.1^a/(1-0.1^4)$, and leverages this data to estimate the average reward $V(f_\pi(\xi_0))$ of the evaluation policy $\pi \in \Pi_0$, which selects each action $a \in \mathcal{A}$ with probability A^{-1}. Thus, π_0 resembles a geometric distribution on the action space, and the density ratio $\pi(a|s)/\pi_0(a|s)$ increases rapidly with a, which poses a well-known challenge in off-policy evaluation literature (Liu et al., 2018b).

We compare our distributionally robust value estimator $V^\pi_\rho(\widehat{\xi}_T)$ against the marginalized importance sampling estimator $\widehat{V}^\pi_{\mathrm{MIS}}(\widehat{\xi}_T)$ by Liu et al. (2018b), which is known to have low variance, and against the distributionally robust estimator $\widehat{V}^\pi_{\mathrm{OT}}(\widehat{\xi}_T)$ by Wang et al. (2024), which uses an optimal transport uncertainty set. As this estimator is tailored to MDPs with a discounted cost criterion, we set the corresponding discount factor to $\gamma = 85\%$ and normalize the estimator by $(1-\gamma)^{-1}$ to approximate the long-run average cost of π as proposed by Tsitsiklis and Van Roy (2002).

The goal of the first experiment is to empirically validate the statistical optimality of $V^\pi_\rho(\widehat{\xi}_T)$. To this end, we fix a sample size $T \in \{500, 1{,}000, 2{,}000\}$ and sample 20 independent state-action trajectories of length T from the Markov chain induced by the transition kernel Q and the behavioral

policy π_0. For each trajectory, we then compute the three distinct estimators. The empirical out-of-sample disappointment $\widehat{\beta}$ of any estimator is defined as the percentage of the trajectories for which the estimator falls below the true expected long-run average reward $V(f_\pi(\xi_0))$ of the evaluation policy π. The out-of-sample disappointment of the two distributionally robust estimators can be tuned by changing the radii of the underlying uncertainty sets, while that of the marginalized importance sampling estimator can be tuned by applying an additive offset. All hyperparameter values that were tested in the first experiment are reported in Table 1.

Table 1: Tested hyperparameter values (uncertainty radii of the two distributionally robust estimators and additive offsets of the marginalized importance sampling estimator), where $\mathcal{K} = \{0, \ldots, 9\}$

T	Uncertainty radii of $V_\rho^\pi(\widehat{\xi}_T)$	Uncertainty radii of $\widehat{V}_{\mathrm{OT}}^\pi(\widehat{\xi}_T)$	Offsets of $\widehat{V}_{\mathrm{MIS}}^\pi(\widehat{\xi}_T)$
500	$\{0.01 - 1.111 \times 10^{-3} k : k \in \mathcal{K}\}$	$\{5 - 0.44k : k \in \mathcal{K}\}$	$\{0.2 - 0.02k : k \in \mathcal{K}\}$
1,000	$\{0.01 - 1.111 \times 10^{-3} k : k \in \mathcal{K}\}$	$\{2.5 - 0.17k : k \in \mathcal{K}\}$	$\{0.15 - 0.011k : k \in \mathcal{K}\}$
2,000	$\{0.01 - 1.111 \times 10^{-3} k : k \in \mathcal{K}\}$	$\{2.5 - 0.17k : k \in \mathcal{K}\}$	$\{0.15 - 0.016k : k \in \mathcal{K}\}$

Theorem 3.8 ensures that the distributionally robust value estimator $V_\rho^\pi(\widehat{\xi}_T)$ is the least conservative (*i.e.*, largest) of all estimators with a prescribed decay rate of the out-of-sample disappointment. For large finite values of T and for any choices of the hyperparameters that induce the same empirical out-of-sample disappointment, we thus expect our distributionally robust value estimator to exceed the two baseline estimators. This conjecture is supported by the numerical results in Figure 1, which shows that our estimator Pareto dominates the two baselines in that it predicts the highest rewards for any given upper bound on the out-of-sample disappointment. The first experiment thus complements the theoretical analysis in Section 3 by showing that the statistical efficiency of the proposed estimator persists even for (practically relevant) finite sample sizes and fixed out-of-sample disappointment levels. We also emphasize that, while the two benchmark estimators require explicit knowledge of the behavioral policy π_0, the proposed distributionally robust estimator can be evaluated without this knowledge.

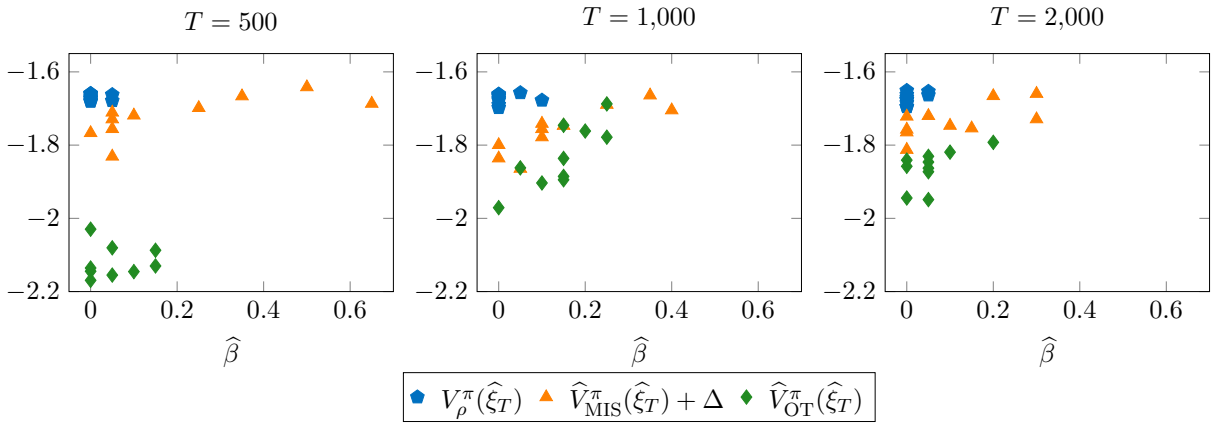

Figure 1: Scatter plot of the average reward predicted by different estimators against the empirical out-of-sample disappointment $\widehat{\beta}$. Points correspond to different hyperparameter values from Table 1.

6.2 Offline Policy Optimization: Machine Replacement

The second experiment is based on the machine replacement problem described in (Wiesemann et al., 2013). The objective is to determine an optimal repair strategy for a machine whose condition is represented by eight "operative" states, labeled $1, \ldots, 8$, and two "repair" states, labeled R1 and R2. The available actions are either "do nothing" or "repair." There is no reward in any of the operative states, while states 8, R1, and R2 yield rewards of -20, -2, and -10 per time period, respectively. The initial state s_0 is governed by the uniform distribution over $\mathcal{S}$, and the transition kernel Q is defined as in (Wiesemann et al., 2013, § 6). Details are omitted for brevity.

In this experiment we compare the out-of-sample performance of our distributionally robust policy estimator $\pi_\rho(\widehat{\xi}_T)$ against that of two benchmark estimators. The first benchmark, denoted by $\widehat{\pi}_{\mathrm{KL}}(\widehat{\xi}_T)$, is obtained from a robust offline reinforcement learning problem with an (s,a)-rectangular uncertainty set defined using the Kullback-Leibler divergence (Shi and Chi, 2024, § 4). This estimator enjoys near-optimal sample complexity. The second benchmark, denoted by $\widehat{\pi}_{\mathrm{PI}}(\widehat{\xi}_T)$, is based on a non-robust, model-based ("plug-in") approach, which achieves minimax-optimal sample complexity (Li et al., 2024a). Both baseline estimators are designed to maximize discounted reward. Since the average reward of a policy is the leading term in the expansion of the discounted reward as the discount factor approaches 1, a policy that is optimal for large discount factors must be nearly optimal for the average reward criterion (Puterman, 2005, § 10.1.2). We thus set the discount factor to 0.95 when computing the baseline estimators. We also emphasize that both baseline estimators require access to independent samples from the stationary state-action-next-state distribution. In contrast, the proposed distributionally robust policy estimator only requires access to one single state-action trajectory generated under an unknown behavioral policy. To ensure a fair comparison, we construct the baseline estimators from the $T-1$ (dependent) state-action-next-state triples in the state-action trajectory of length T that is made available to all estimators. Finally, we set the radii of the uncertainty sets of the distributionally robust estimators $\widehat{\pi}_{\mathrm{KL}}(\widehat{\xi}_T)$ and $\widehat{\pi}_{\mathrm{PI}}(\widehat{\xi}_T)$ to $4.5/T$. This scaling is recommended by Duchi et al. (2021) for $\widehat{\pi}_{\mathrm{KL}}(\widehat{\xi}_T)$ and ensures that the out-of-sample disappointment of $\pi_\rho(\widehat{\xi}_T)$ remains approximately constant at $\beta \approx 1\%$ (see Theorem 4.3).

We assume now that the behavioral policy $\pi_0 \in \Pi_0$ selects each action $a \in \mathcal{A}$ with probability $1/A$ irrespective of the current state. For any fixed sample size T, we first generate a state-action trajectory of length T from the Markov chain induced by π_0 and Q. For each of these trajectories, we then construct the three policy estimators and compute their true long-run average rewards. Finally, we record the frequency with which each estimator achieves the highest reward across the 100 simulation runs; see Figure 2. We observe that our distributionally robust policy estimator wins most often for all sample sizes $T \lesssim 400$. For larger sample sizes, the minimax-optimal plug-in estimator dominates. The robust policy estimator $\widehat{\pi}_{\mathrm{KL}}(\widehat{\xi}_T)$ displays a similar performance as $\pi_\rho(\widehat{\xi}_T)$ for small values of T. We believe that the performance of $\pi_\rho(\widehat{\xi}_T)$ deteriorates in this regime because of Algorithm 2, which outputs suboptimal reward estimates when the uncertainty set is large.

Figure 2 does not change substantially even if the baseline estimators are given unfair access to independent samples from the *true* stationary state-action-next-state distribution (not shown).

Acknowledgement

This work was supported as a part of the NCCR Automation, a National Center of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40_225155).

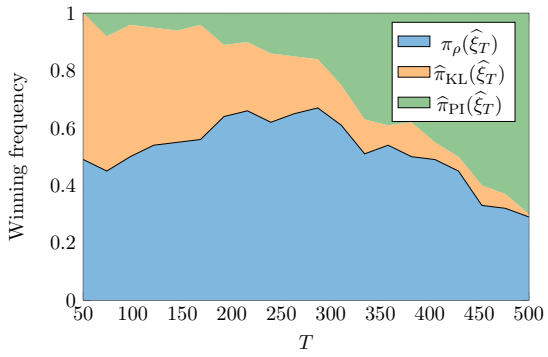

Figure 2: Frequencies at which each of the three policy estimators achieves the highest long-run average reward across 100 independent simulation runs, as a function of T.

References

Charalambos D Aliprantis and Kim C Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2006.

Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

Amine Bennouna and Bart Van Parys. Learning and decision-making with data: Optimal formulations and phase transitions. *arXiv preprint arXiv:2109.06911*, 2021.

Dimitri Bertsekas. *A Course in Reinforcement Learning*. Athena Scientific, 2023.

Mohak Bhardwaj, Tengyang Xie, Byron Boots, Nan Jiang, and Ching-An Cheng. Adversarial model for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 2024.

Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2009.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(18):503–556, 2005.

Rasool Fakoor, Jonas W Mueller, Kavosh Asadi, Pratik Chaudhari, and Alexander J Smola. Continuous doubly constrained batch reinforcement learning. *Advances in Neural Information Processing Systems*, 2021.

Arnab Ganguly and Tobias Sutter. Optimal learning via moderate deviations theory. *arXiv preprint arXiv:2305.14496*, 2024.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature*, 25(1):16–18, 2019.

Onésimo Hernández-Lerma and Jean B Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Yichun Hu, Nathan Kallus, and Masatoshi Uehara. Fast rates for the regret of offline reinforcement learning. *Mathematics of Operations Research*, 50(1):633–655, 2025.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. *International Conference on Machine Learning*, 2016.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6):3282–3302, 2022.

Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. *International Conference on Machine Learning*, 2022.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. *International Conference on Machine Learning*, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint, arXiv:2005.01643*, 2020.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024a.

Mengmeng Li, Tobias Sutter, and Daniel Kuhn. Distributionally robust optimization with Markovian data. *International Conference on Machine Learning*, 2021.

Mengmeng Li, Daniel Kuhn, and Tobias Sutter. Policy gradient algorithms for robust MDPs with non-rectangular uncertainty sets. *arXiv preprint arXiv:2305.19004*, 2023.

Tianjiao Li, Feiyang Wu, and Guanghui Lan. Stochastic first-order methods for average-reward Markov decision processes. *Mathematics of Operations Research (ahead of print)*, 2024b.

Feng Liu, Ruiming Tang, Xutao Li, Weinan Zhang, Yunming Ye, Haokun Chen, Huifeng Guo, and Yuzhou Zhang. Deep reinforcement learning based recommendation with explicit user-item interactions modeling. *arXiv preprint arXiv:1810.12027*, 2018a.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 2018b.

Xiaoteng Ma, Zhipeng Liang, Li Xia, Jiheng Zhang, Jose Blanchet, Mingwen Liu, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. *International Conference on Autonomous Agents and Multiagent Systems*, 2014.

Shie Mannor, Duncan Simester, Peng Sun, and John N. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

S Natarajan. Large deviations, hypotheses testing, and source coding for finite Markov chains. *IEEE Transactions on Information Theory*, 31(3):360–365, 1985.

James R Norris. *Markov Chains*. Cambridge University Press, 1998.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. *International Conference on Machine Learning*, 2019.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving Rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in Neural Information Processing Systems*, 2022.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. *International Conference on Machine Learning*, 2000.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 2005.

Shyam Sundhar Ramesh, Pier Giuseppe Sessa, Yifan Hu, Andreas Krause, and Ilija Bogunovic. Distributionally robust model-based reinforcement learning with large state spaces. *International Conference on Artificial Intelligence and Statistics*, 2024.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer, 2009.

Naman Saxena, Subhojyoti Khastagir, Shishir Kolathaya, and Shalabh Bhatnagar. Off-policy average reward actor-critic with deterministic policy search. *International Conference on Machine Learning*, 2023.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024.

Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust batch contextual bandits. *Management Science*, 69(10):5772–5793, 2023.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Jeffrey E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

Tobias Sutter, Andreas Krause, and Daniel Kuhn. Robust generalization despite distribution shift via minimum discriminating information. *Advances in Neural Information Processing Systems*, 2021.

Tobias Sutter, Bart Van Parys, and Daniel Kuhn. A Pareto dominance principle for data-driven optimization. *Operations Research*, 72(5):1976–1999, 2024.

Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *Advances in Neural Information Processing Systems*, 2015.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. *International Conference on Machine Learning*, 2016.

John N Tsitsiklis and Benjamin Van Roy. On average versus discounted reward temporal-difference learning. *Machine Learning*, 49:179–191, 2002.

Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Offline minimax soft-Q-learning under realizability and partial coverage. *Advances in Neural Information Processing Systems*, 2023.

Ilnura Usmanova, Maryam Kamgarpour, Andreas Krause, and Kfir Levy. Fast projection onto convex smooth constraints. *International Conference on Machine Learning*, 2021.

Bart Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402, 2021.

Mathukumalli Vidyasagar. An elementary derivation of the large deviation rate function for finite state Markov chains. *Asian Journal of Control*, 16(1):1–19, 2014.

Jie Wang, Rui Gao, and Hongyuan Zha. Reliable off-policy evaluation for reinforcement learning. *Operations Research*, 72(2):699–716, 2024.

Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 2021.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous Q-learning. *IEEE Transactions on Information Theory*, 69(11):7185–7219, 2023.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in Neural Information Processing Systems*, 2021.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. *Conference on Learning Theory*, 2022.

Shangtong Zhang, Yi Wan, Richard S Sutton, and Shimon Whiteson. Average-reward off-policy policy evaluation with function approximation. *International Conference on Machine Learning*, 2021.