

# ROBUSTNESS TO MODEL APPROXIMATION, MODEL LEARNING FROM DATA, AND SAMPLE COMPLEXITY IN WASSERSTEIN REGULAR MDPS

YICHEN ZHOU<sup>†</sup>, YANGLEI SONG<sup>†</sup>, AND SERDAR YÜKSEL<sup>†</sup>

**ABSTRACT.** The paper studies the robustness properties of discrete-time stochastic optimal control under Wasserstein model approximation for both discounted cost and average cost criteria. Specifically, we study the performance loss when applying an optimal policy designed for an approximate model to the true dynamics compared with the optimal cost for the true model under the sup-norm-induced metric, and relate it to the Wasserstein-1 distance between the approximate and true transition kernels. A primary motivation of this analysis is empirical model learning, as well as empirical noise distribution learning, where Wasserstein convergence holds under mild conditions but stronger convergence criteria, such as total variation, may not. We discuss applications of the results to the disturbance estimation problem, where sample complexity bounds are given, and also to a general empirical model learning approach, obtained under either Markov or i.i.d. learning settings.

## 1. INTRODUCTION

In this paper, we study the robustness properties of discrete-time stochastic optimal control under model approximation across various performance criteria. Specifically, we characterize how the accuracy of a model approximation affects the performance loss incurred when applying an optimal policy designed for the approximate model to the true system dynamics. This notion of robustness is of practical importance, as learning-based algorithms are seldom implemented with exact model knowledge.

As discussed in the literature review, the term *robustness* has been interpreted in various ways across diverse contexts and methodological frameworks. In this paper, we define robustness in terms of performance degradation—specifically, the loss incurred when a control policy designed for an approximate or incorrect model is applied to the true system, measured relative to the optimal cost achievable with full knowledge of the true model.

**1.1. Preliminaries on Markov Decision Processes.** Consider a discrete-time controlled Markov process  $\{(X_t, U_t) : t \geq 0\}$  defined on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , with Polish state and action spaces  $(\mathbb{X}, d_{\mathbb{X}})$  and  $(\mathbb{U}, d_{\mathbb{U}})$ , respectively. These spaces are endowed with their corresponding Borel  $\sigma$ -algebras, denoted by  $\mathcal{F}_{\mathbb{X}}$  and  $\mathcal{F}_{\mathbb{U}}$ . Let  $(\mathbb{X} \times \mathbb{U}, \sigma(\mathcal{F}_{\mathbb{X}} \times \mathcal{F}_{\mathbb{U}}))$  be the product measurable space.

**Definition 1.1.** A map  $\mathcal{T} : \mathcal{F}_{\mathbb{X}} \times \mathbb{X} \times \mathbb{U} \rightarrow [0, 1]$  is a *controlled transition kernel* if a) for every  $A \in \mathcal{F}_{\mathbb{X}}$ , the map  $(x, u) \mapsto \mathcal{T}(A|x, u)$  is  $\sigma(\mathcal{F}_{\mathbb{X}} \times \mathcal{F}_{\mathbb{U}})$ -measurable, and b) for every  $(x, u) \in \mathbb{X} \times \mathbb{U}$ , the map  $A \mapsto \mathcal{T}(A|x, u)$  is a probability measure on  $(\mathbb{X}, \mathcal{F}_{\mathbb{X}})$ .

We assume that the true dynamics of  $\{X_t, U_t : t \geq 0\}$  is given by a controlled transition kernel  $\mathcal{T}$ , that is, for  $(x, u) \in \mathbb{X} \times \mathbb{U}$  and  $A \in \mathcal{F}_{\mathbb{X}}$ ,  $\mathbb{P}(X_{t+1} \in A | X_t = x, U_t = u) = \mathcal{T}(A|x, u)$  for  $t \geq 0$ . Further, we define the history (or path) space at time  $t$  as  $(\mathbb{H}_t, \mathcal{F}_{\mathbb{H}_t}) := (\mathbb{X}^{t+1} \times \mathbb{U}^t, \sigma(\mathcal{F}_{\mathbb{X}}^{t+1} \times \mathcal{F}_{\mathbb{U}}^t))$ . An admissible policy at time  $t$  is a measurable function  $\gamma_t : \mathbb{H}_t \rightarrow \mathcal{P}(\mathbb{U})$ , where  $\mathcal{P}(\mathbb{U})$  denotes the space of probability measures on  $(\mathbb{U}, \mathcal{F}_{\mathbb{U}})$ , endowed with the weak topology. At each time  $t \geq 0$ , the decision-maker observes the realized history  $h_t := \{X_{[0,t]}, U_{[0,t-1]}\} \in \mathbb{H}_t$  and selects an action

<sup>†</sup>DEPARTMENT OF MATHEMATICS AND STATISTICS, QUEEN'S UNIVERSITY, KINGSTON, ON, CANADA

*E-mail addresses:* yichen.zhou@queensu.ca, yanglei.song@queensu.ca, yuksel@queensu.ca.

*Key words and phrases.* Markov decision processes, robustness, model estimation, sample complexity.

$U_t \in \mathbb{U}$  according to the distribution  $\gamma_t(h_t)$ . The system then incurs a cost  $c(X_t, U_t)$  and transitions to a new state  $X_{t+1}$  according to  $\mathcal{T}(\cdot \mid X_t, U_t)$ . This procedure is repeated over time.

We refer to the quadruple  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  as a discrete-time Markov decision process, abbreviated as MDP, and denote the set of admissible policies by

$$\Gamma_A := \{\{\gamma_t\}_{t=0}^\infty : \forall t \in \mathbb{N}, \gamma_t : \mathbb{H}_t \rightarrow \mathcal{P}(\mathbb{U}) \text{ measurable}\}.$$

Our objective is to minimize the accumulated cost over time by selecting a control policy  $\{\gamma_t : t \in \mathbb{N}\}$  from  $\Gamma_A$ , according to one of the performance criteria specified below.

(a) **Discounted Cost Criterion.** Given a discount factor  $\beta \in (0, 1)$ , for  $x \in \mathbb{X}$ , define

$$\text{the value function under policy } \gamma : \quad J_\beta(c, \mathcal{T}, \gamma)(x) := \mathbb{E}^{\mathcal{T}, \gamma} \left[ \sum_{t=0}^{\infty} \beta^t c(X_t, U_t) \mid X_0 = x \right],$$

$$\text{and the optimal value function :} \quad J_\beta^*(c, \mathcal{T})(x) := \inf_{\gamma \in \Gamma_A} J_\beta(c, \mathcal{T}, \gamma)(x).$$

(b) **Average Cost Criterion.** For  $x \in \mathbb{X}$ , define

$$\text{the value function under policy } \gamma \quad J_\infty(c, \mathcal{T}, \gamma)(x) := \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}^{\mathcal{T}, \gamma} \left[ \sum_{t=0}^{T-1} c(X_t, U_t) \mid X_0 = x \right],$$

$$\text{and the optimal value function:} \quad J_\infty^*(c, \mathcal{T})(x) := \inf_{\gamma \in \Gamma_A} J_\infty(c, \mathcal{T}, \gamma)(x).$$

Here,  $\mathbb{E}^{\mathcal{T}, \gamma}$  denotes the expectation under the controlled transition kernel  $\mathcal{T}$  and the policy  $\gamma$ .

**1.2. Problem Statement and Contributions.** Consider a reference MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$ , where a decision-maker seeks to perform optimal control. In practice, it is rarely the case that the decision-maker has complete knowledge of the MDP—particularly the cost function  $c$  and the transition kernel  $\mathcal{T}$ . It is therefore natural to consider a learning-based scheme in which the decision-maker first estimates an approximate MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$  using samples generated from the reference MDP, derives an optimal policy  $\gamma_{\hat{c}, \mathcal{S}}^*$  for the approximate model, and then applies this policy to the true MDP.

This gives rise to a fundamental question: how much performance is lost by applying an optimal policy derived from an approximate model to the true system? We refer to this performance degradation due to model mismatch as the *robustness error*. Specifically, under the discounted and average cost criteria, respectively, the robustness error due to model mismatch is given by

$$\|J_\beta(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*) - J_\beta^*(c, \mathcal{T})\|_\infty \quad \text{and} \quad \|J_\infty(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*) - J_\infty^*(c, \mathcal{T})\|_\infty.$$

In this paper, we derive upper bounds on the robustness error in terms of the estimation errors in the cost function and the transition kernel. We then apply these results to the context of empirical model learning. Furthermore, we examine the implications of our bounds for sample complexity, characterizing how the mismatch error scales with the number of samples in data-driven settings. Our main contributions are as follows.

- (i) In Section 2.2.1, we establish conditions under which the optimal *discounted* cost value function is Lipschitz continuous with respect to the model; specifically in the cost function and the transition kernel (Theorem 2.4). In Section 2.2.2, we present analogous conditions for the Lipschitz continuity of the optimal *average* cost value function in the model space (Theorems 2.5 and 2.6). For the average cost setting, we adopt two distinct approaches: one based on a minorization condition and the other on the vanishing discount factor method.
- (ii) In Section 2.3.1, we establish conditions under which the *robustness error* bound for the optimal *discounted* cost value function exhibits Lipschitz continuity with respect to the model (Theorem 2.7). In Section 2.3.2, we present the analogous result for the *average* cost criterion (Theorems 2.8 and 2.9, Lemma 2.1), using the two approaches discussed earlier.

- (iii) In Section 3, as a primary contribution, we establish robustness guarantees for empirical model learning, along with explicit convergence rates that yield *parametric* sample complexity bounds. Specifically, we propose data-driven learning algorithms and derive sample complexity results under two distinct data generation scenarios: (a) data collected along a controlled sample path (Theorems 3.1 and 3.2), and (b) data generated via a device that simulates the Markov kernel (Theorem 3.3).
- (iv) In Section 4, we extend our analysis to the setting where robustness is measured with respect to the distribution of the driving noise (Theorem 4.1 and 4.2). We establish convergence rates in terms of the sample size when the driving noise distribution is estimated empirically (Theorem 4.3), and show improved, *parametric* rates under additional uniform regularity conditions (Theorem 4.4). Furthermore, we generalize these results to the setting where both the model and the driving noise must be learned from data (Theorem 4.5 and the two examples that follow).

**1.3. Literature Review.** Stochastic control and reinforcement learning under model misspecification are fundamental to their applications and have been extensively studied.

**Robustness to model approximation.** We should note that the term *robustness* has various interpretations, contexts and solution methods. For example, a common approach to robustness in the literature has been to design controllers that work sufficiently well for a family of (uncertain) systems under some structured constraints, such as  $H_\infty$  norm bounded perturbations for linear time-invariant systems (see [1, 2]). For such problems, the design for robust controllers has often been developed through a game theoretic formulation where the minimizer is the controller and the maximizer is the uncertainty. In [3, 4] the authors established the connections of this formulation to risk sensitive control. Aligned with this formulation, relative entropy constraints came in to the literature to probabilistically model the uncertainties, see e.g. [5, Eqn. (4)] or [3, Eqns. (2)-(3)]. Therefore, one approach in robust stochastic control has been to consider all models which satisfy certain constraints; see e.g. [3, 5–7] among others. In order to quantify the uncertainty in the system models, other than the relative entropy pseudo-distance, several other metrics or criteria have also been used in the literature, such as total variation [8]. For the related distributionally robust stochastic optimization framework, it is assumed that the underlying probability measure of the system lies within an ambiguity set and a worst case single-stage optimization is made considering the probability measures in the ambiguity set. To construct ambiguity sets, [9–12] use the Wasserstein metric, [13] uses the Prokhorov metric which metrizes the weak topology, [14] uses the total variation distance, and [15] works with relative entropy. For fully observed finite state-action space models with uncertain transition probabilities, the authors in [16, 17] have studied robust dynamic programming approaches through a min-max formulation. Similar work with model uncertainty includes [18–20]. In the economics literature related work has been reported in seminal studies such as [21, 22].

The robustness formulation in this paper will be to consider a single approximate model, and establish upper bounds for the errors induced by applying a policy designed for the approximate model. We note that the upper bounds are readily applicable to a set of approximate models by taking the suprema of the error bounds with respect to set of models, and the difference between our formulation and the model uncertainties formulation reviewed above is that the latter has natural connections to minimax risk [12]. Our setting has been considered in [23–31] (discrete-time) and in [32] (continuous-time). To our knowledge, one of the the earliest studies in this setting is [23], where the author considered fully observed discrete-time controlled models and established continuity of the optimal value function with respect to models and gives a set convergence result for sets of optimal control actions. Related results are presented in [26, 27]. A closely related sequence of works is [24, 25], where the authors study robustness to incorrect transition kernels, focusing primarily on asymptotic convergence under weak convergence and setwise convergence,

as well as on strong uniformity results with respect to total variation convergence. Specifically, a robustness result is obtained, in an asymptotic convergence sense, for the discounted cost criterion in [24] and for the average cost criterion in [25], respectively, under *continuous* weak convergence of transition kernels; uniformity results under total variation convergence are also established. A unified perspective is given in [33], which considers quantized approximations as a special case.

Another closely related sequence of studies is [28–30], which, to our knowledge, are among the first to consider quantitative performance bounds induced by model mismatch. In [28], the authors focused on the discounted cost criterion, obtaining a weighted norm version of the first inequality in our Theorem 2.7. In [29], the authors focused on the average cost criterion, obtaining results similar to Theorem 2.8 by imposing contraction assumptions on Bellman operators of the reference and approximate models. In [30], the authors worked with the “relative stability index”, i.e. the robustness error divided by the optimal value function under the true model, and obtained an upper bound under the discounted criterion that is independent of the discount factor under an assumption that is similar to that of our Theorem 2.8. We remark that a recent work [31] independently recovers and extend the results in [28]. We acknowledge that Theorem 2.4 and 2.7 already appeared in [31] as Corollary 1, and our Theorem 2.7 is a refinement of Theorem 5 of the same paper. In the first half of this paper, our focus is to provide a unified proof strategy to obtain results mentioned above under both discounted and average cost criterion, and discuss the set of assumptions behind them, what is shared and what is different.

**Model learning from data.** In the second half of the paper, we connect our robustness error bounds to the statistical rates at which the performance of a policy, obtained under a model estimated via sampling, approaches that of the optimal policy. We will refer to this general setting as *model learning from data*.

**Model-based offline learning.** If the model estimation is done via sampling the transition kernel and cost function, the setting will fall under model-based offline learning. For finite MDPs, the statistical rates of error bounds has been well studied for both model estimation with i.i.d samples [34, 35] and model estimation from one trajectory [36]. Results for continuous MDPs often impose structural assumptions on the transition kernel, such as the kernel can be parameterized [37, 38]. In Section 3, we build on the rich literature concerning the fundamental performance loss of static discretization in MDPs [39–43] to establish statistical rates for model-based learning under MDPs with Wasserstein-1 Lipschitz transition kernels, under both i.i.d. sampling (Theorem 3.3) and one trajectory of samples (Theorem 3.1 and 3.2). To our knowledge, these results are new to both RL and control literature. We remark that our setting has been considered in recent works of regret minimization with adaptive discretization [44–46]. However, their focus is on adaptively finding the best discretization scheme, while our work focus on the statistical performance of a fixed state partition. We also remark a recent line of works [47–49] connecting static discretization to learning on partially observable MDPs, and it is of independent interest to connect our results to theirs.

**Disturbance estimation.** Another framework of model estimation is by approximating the distribution of the driving “disturbance process” [50–53]: consider a stochastic dynamical system  $X_{t+1} = f(X_t, U_t, W_t)$ , where  $\{W_t\}_{t=0}^\infty$  is an i.i.d process with distribution  $\mu$ . If an alternative distribution  $\nu$  is proposed and an approximate optimal policy is computed under  $\nu$ , the objective is to characterize the relationship between the error induced by disturbance approximation and a certain distance between  $\mu$  and  $\nu$ . In [53], a robustness result is given for the discounted cost criterion with the bounded-Lipschitz distance between  $\mu$  and  $\nu$ . In [50], similar results are obtained for the average cost criterion, using either the total variation distance or bounded-Lipschitz distance.

Results on robustness to weak convergence in [24, 25] and to Wasserstein convergence in [37, 51, 52] imply empirical consistency with i.i.d. learning of models, as noted in these studies. In [51, 52], the authors consider a learning scenario where  $\nu = \frac{1}{n+1} \sum_{i=0}^n \delta_{w_i}$  is the empirical measure of  $\mu$  under samples  $\{w_t\}_{t=0}^n$ . The chosen distance between measures is Wasserstein-1 distance, with the

discounted cost criterion considered in [51], and the average cost criterion in [52]. In Section 4, we show that the disturbance distribution approximation problem can be viewed as a special case of model approximation, and through Theorem 2.7 and 2.9, we obtain results analogous to those in [51, 52], with relaxed assumptions and improved bounds, as presented in Theorem 4.3 and 4.4. We remark a further related result on empirical consistency from a different perspective is presented in [54].

As noted above, distributionally robust stochastic control approach [10–12, 55, 56] considers a related, but distinct setting. The key difference is that under distributional robustness the disturbance distribution is allowed to vary over time. In [11], the authors study the case of the Wasserstein ball, while in [12], both Wasserstein and f-divergence balls are considered. As noted earlier, the distributionally robust formulation has a natural connection to minimax bounds on robustness error [12].

#### 1.4. Space of Probability Measures, Transition Kernels, and Wasserstein Regular MDPs.

In this subsection, we introduce notation and definitions used later. Given a complete separable metric space  $(\mathbb{X}, d_{\mathbb{X}})$  endowed with its Borel  $\sigma$ -algebra, denote by  $\mathcal{P}(\mathbb{X})$  the space of probability measures on it. We denote the discrepancy between two probability measures with respect to a function  $f : \mathbb{X} \rightarrow \mathbb{R}$  by

$$d_f(\mu, \nu) := \left| \int_{\mathbb{X}} f(x)(\mu(dx) - \nu(dx)) \right|.$$

**Definition 1.2** (Wasserstein- $p$  Distance, Definition 3.1.1, [57]). Let  $1 \leq p < \infty$ . For two probability measures  $\mu, \nu \in \mathcal{P}(\mathbb{X})$ , where  $\mathbb{X}$  is locally compact and separable, assume that for some  $x_0 \in \mathbb{X}$ ,  $\int_{\mathbb{X}} d_{\mathbb{X}}(x, x_0)^p \mu(dx) < \infty$  and  $\int_{\mathbb{X}} d_{\mathbb{X}}(x, x_0)^p \nu(dx) < \infty$ . The Wasserstein- $p$  distance is defined as follows:

$$\mathcal{W}_p(\mu, \nu) := \inf \left\{ \left( \int_{\mathbb{X} \times \mathbb{X}} d_{\mathbb{X}}(x, y)^p \pi(dx, dy) \right)^{\frac{1}{p}} : \pi \text{ is a coupling between } \mu, \nu \right\},$$

where a coupling  $\pi$  between two probability measures  $\mu, \nu \in \mathcal{P}(\mathbb{X})$  is defined as any probability measure on  $(\mathbb{X}^2, \sigma(\mathcal{F}_{\mathbb{X}} \times \mathcal{F}_{\mathbb{X}}))$  such that for any  $A \in \mathcal{F}_{\mathbb{X}}$ :  $\pi(A \times \mathbb{X}) = \mu(A)$ ,  $\pi(\mathbb{X} \times A) = \nu(A)$ .

*Remark 1.1.* For  $p = 1$ , the Wasserstein distance admits a dual formulation (Kantorovich–Rubinstein formula, Chapter 6, [58]):

$$\mathcal{W}_1(\mu, \nu) = \sup_{\|f\|_{\text{Lip}} \leq 1} d_f(\mu, \nu), \text{ where } \|f\|_{\text{Lip}} := \sup_{x \neq y \in \mathbb{X}} \frac{|f(x) - f(y)|}{d_{\mathbb{X}}(x, y)}.$$

That is,  $\|f\|_{\text{Lip}}$  is the Lipschitz norm of a function  $f : \mathbb{X} \rightarrow \mathbb{R}$ .

A key motivation for this paper is to show that if an MDP is regular in the Wasserstein sense, as detailed in Section 2.4, the MDP can be approximated via sampling, and the optimal policies obtained from the sampled MDPs are near-optimal. Specifically, the performance loss shares the same diminishing rate as the distance between the empirical and population-level distributions.

Next, we consider controlled probability kernels, which can also be viewed as measure-valued functions:  $\mathcal{T} : \mathbb{X} \times \mathbb{U} \rightarrow \mathcal{P}(\mathbb{X})$ . With a slight abuse of notation, we denote the discrepancy between two controlled transition kernels  $\mathcal{T}$  and  $\mathcal{S}$ , with respect to a function  $f : \mathbb{X} \rightarrow \mathbb{R}$  and uniformly over varying inputs, by

$$d_f(\mathcal{T}, \mathcal{S}) := \sup_{x \in \mathbb{X}, u \in \mathbb{U}} d_f(\mathcal{T}(\cdot|x, u), \mathcal{S}(\cdot|x, u)).$$

Furthermore, we extend the Wasserstein-1 distance to controlled transition kernels using the concept of uniform discrepancy defined above:

$$d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}) := \sup_{\|f\|_{\text{Lip}} \leq 1} d_f(\mathcal{T}, \mathcal{S}).$$

The following lemma follows directly from the definition above, showing that an upper bound based on  $d_{\mathcal{W}_1}$  is more conservative than one based on  $d_f$ .

**Lemma 1.1.** *Given a function  $f : \mathbb{X} \rightarrow \mathbb{R}$  that is  $\|f\|_{\text{Lip}}$ -Lipschitz, and two controlled transition kernels  $\mathcal{T}, \mathcal{S}$ , then we have*

$$d_f(\mathcal{T}, \mathcal{S}) \leq \|f\|_{\text{Lip}} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}).$$

We now define the central assumption in this paper, which will be the sufficient condition for most of the results presented:

**Assumption 1.1** (Wasserstein Regular MDPs). Consider an MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and a discount factor  $\beta < 1$ . In addition to Assumption 2.1, assume the following holds:

- (a). For any  $u \in \mathbb{U}$ ,  $c : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$  is  $\|c\|_{\text{Lip}}$ -Lipschitz continuous as a function of  $x$ .
- (b). For any  $u \in \mathbb{U}$ ,  $\mathcal{T} : \mathbb{X} \times \mathbb{U} \rightarrow \mathcal{P}(\mathbb{X})$  is  $\|\mathcal{T}\|_{\text{Lip}}$ -Lipschitz continuous as function of  $x$ , with respect to Wasserstein-1 distance, i.e.

$$\mathcal{W}_1(\mathcal{T}(\cdot|x, u), \mathcal{T}(\cdot|y, u)) \leq \|\mathcal{T}\|_{\text{Lip}} d_{\mathbb{X}}(x, y), \forall u \in \mathbb{U}, x, y \in \mathbb{X}.$$

- (c).  $\beta \|\mathcal{T}\|_{\text{Lip}} < 1$ .

*Remark 1.2.* This has been a standard assumption in related works [31, 43], and also in regret minimization with weakly continuous MDPs [44–46]. [59] is one of the first papers to study the implications of such an assumption, and it was revisited in [43, 60], in which the latter imposes more restrictive assumptions. In a recent work [61], It is shown that Lipschitz assumptions for belief MDP reduction of POMDP can be fulfilled via standard Lipschitz assumptions on original MDP and a further Dobrushin coefficient condition. Some applications of the assumptions, in addition to works we previously discussed, include [62] on policy gradient, and [63] on continuity of Q-factors.

## 2. CONTINUITY OF OPTIMAL VALUE FUNCTIONS IN MODELS AND ROBUSTNESS TO MODEL APPROXIMATION

**2.1. Optimality Equations.** In this section, we present the basic assumptions that allow us to characterize solutions to stochastic optimal control problems by fixed point equations, and furthermore, allow us to restrict attention to deterministic stationary control policies  $\gamma : \mathbb{X} \rightarrow \mathbb{U}$ . We denote the set of continuous and bounded measurable functions mapping from  $(\mathbb{X}, \mathcal{F}_{\mathbb{X}})$  to  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  by  $C_b(\mathbb{X})$ . For a function  $c : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$ , we define the supremum norm as  $\|c\|_{\infty} = \sup_{x,u} |c(x, u)|$ . This notation also extends to functions defined on other domains.

**Assumption 2.1.** Unless otherwise noted, we assume that any MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  in this paper satisfies the following:

- (a).  $\mathbb{U}$  is compact.
- (b).  $c : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$  is nonnegative, bounded ( $\|c\|_{\infty} < \infty$ ), and continuous on both  $\mathbb{X}$  and  $\mathbb{U}$ .
- (c). For any  $v \in C_b(\mathbb{X})$ ,  $\int_{\mathbb{X}} v(y) \mathcal{T}(dy|x, u)$  is a continuous on  $\mathbb{X}$  and  $\mathbb{U}$ , i.e.  $\mathcal{T}$  is weakly continuous on  $\mathbb{X} \times \mathbb{U}$ .

Under the above condition, it is well known that the optimal value function is the unique solution to the Discounted Cost Optimality Equation (DCOE), as summarized in the following theorem.

**Theorem 2.1** (DCOE, Theorem 5.2.1 and 5.5.2, [64]). *Consider an MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and a discount factor  $\beta \in (0, 1)$ . Suppose Assumption 2.1 holds.*

- (a). *The optimal discounted cost function  $J_{\beta}^*(c, \mathcal{T}) \in C_b(\mathbb{X})$  is the unique solution to the following fixed point equation in  $v \in C_b(\mathbb{X})$ , which we refer to as DCOE:*

$$v(x) = \inf_{u \in \mathbb{U}} \left\{ c(x, u) + \beta \int_{\mathbb{X}} v(y) \mathcal{T}(dy|x, u) \right\}, \forall x \in \mathbb{X}.$$



(b). There exists a deterministic stationary policy  $\gamma^* : \mathbb{X} \rightarrow \mathbb{U}$  such that

$$v(x) = c(x, \gamma^*(x)) + \beta \int_{\mathbb{X}} v(y) \mathcal{T}(dy|x, \gamma^*(x)), \quad \forall x \in \mathbb{X},$$

and  $J_\beta(c, \mathcal{T}, \gamma^*)(x) = J_\beta^*(c, \mathcal{T})(x)$ ,  $\forall x \in \mathbb{X}$ .

We note that part (a) follows from the fact that if we define the right-hand side of DCOE as an operator on  $C_b(\mathbb{X})$ , which we refer to as the Bellman operator  $\mathbb{T}$ , it can be shown that  $\mathbb{T}$  is a contraction on  $(C_b(\mathbb{X}), \|\cdot\|_\infty)$ , i.e. for any  $f_1, f_2 \in C_b(\mathbb{X})$ ,  $\mathbb{T} : C_b(\mathbb{X}) \rightarrow C_b(\mathbb{X})$  satisfies

$$\|\mathbb{T}f_1 - \mathbb{T}f_2\|_\infty \leq \beta \|f_1 - f_2\|_\infty,$$

where, and throughout,  $\beta$  denotes the discount factor in the discounted cost setting. Thus, by the Banach fixed-point theorem (see, e.g., Theorem 3.1.7 of [64]), the DCOE admits a unique solution (see p.19 of [65]). As discussed in Section 2.4, this uniqueness is important for characterizing the Lipschitz continuity of value functions under appropriate assumptions.

*Remark 2.1* (Discounted Cost Bellman Consistency Equation). Given a fixed deterministic stationary policy  $\gamma$ , we consider the equation

$$v(x) = c(x, \gamma(x)) + \beta \int_{\mathbb{X}} v(y) \mathcal{T}(dy|x, \gamma(x)), \quad \forall x \in \mathbb{X}.$$

Defining the right-hand side as an operator  $\mathbb{T}^\gamma$ , it can be shown that  $\mathbb{T}^\gamma$  is also a contraction, and thus admits a unique solution on the set of bounded measurable functions (as  $\gamma$  can be discontinuous). Furthermore, the solution is  $J_\beta(c, \mathcal{T}, \gamma)$ . We refer to this equation as the discounted cost Bellman consistency equation for  $\gamma$  (see Section 1.1.2, [66]).

For the average cost criterion, we present two additional sets of assumptions, each sufficient to guarantee the validity of the Average Cost Optimality Equation (ACOE). These assumptions are utilized in different parts of our analysis, depending on the specific setting.

**Definition 2.1** (Minorization Condition for Kernels). The controlled transition kernel  $\mathcal{T} : \mathcal{F}_{\mathbb{X}} \times \mathbb{X} \times \mathbb{U} \rightarrow [0, 1]$  is said to satisfy the minorization condition with a probability measure  $\rho$  and a constant  $\epsilon > 0$  if for all  $x \in \mathbb{X}$ ,  $u \in \mathbb{U}$ , and  $A \in \mathcal{F}_{\mathbb{X}}$ , we have  $\mathcal{T}(A|x, u) \geq \epsilon \rho(A)$ .

**Theorem 2.2** (ACOE, Theorem 5.2.1 and 7.2.1, [64]). Consider an MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  such that  $\mathcal{T}$  satisfies the minorization condition with some probability measure  $\rho$  and constant  $\epsilon > 0$ , and suppose Assumption 2.1 holds. Then there exists a constant  $g$  and  $h \in C_b(\mathbb{X})$  such that the following fixed point equation, referred to as ACOE, holds:

$$g + h(x) = \inf_{u \in \mathbb{U}} \left\{ c(x, u) + \int_{\mathbb{X}} h(y) \mathcal{T}(dy|x, u) \right\}, \quad \forall x \in \mathbb{X},$$

where  $g = \inf_{\gamma \in \Gamma_A} J_\infty(c, \mathcal{T}, \gamma)(x)$  for all  $x \in \mathbb{X}$ . Furthermore, there exists a deterministic stationary policy  $\gamma^* : \mathbb{X} \rightarrow \mathbb{U}$  such that

$$g + h(x) = c(x, \gamma^*(x)) + \int_{\mathbb{X}} h(y) \mathcal{T}(dy|x, \gamma^*(x)), \quad \forall x \in \mathbb{X},$$

and  $J_\infty(c, \mathcal{T}, \gamma^*)(x) = g$ ,  $\forall x \in \mathbb{X}$ .

*Remark 2.2.* We refer to  $(g^*, h^*, \gamma^*)$  satisfying ACOE as a canonical triplet. To emphasize the dependence on the cost function  $c$  and kernel  $\mathcal{T}$ , we use the following notation for the canonical triplet:  $(g_{c, \mathcal{T}}^*, h_{c, \mathcal{T}}^*, \gamma_{c, \mathcal{T}}^*)$ . Note that  $g_{c, \mathcal{T}}^* = J_\infty^*(c, \mathcal{T})$ .

The function  $h_{c, \mathcal{T}}^*$  is not unique. When the minorization condition holds, we define  $h_{c, \mathcal{T}}^*$  as the unique fixed-point of the following contraction map on  $C_b(\mathbb{X})$ :

$$\mathbb{T}_{c, \mathcal{T}} v(x) := \inf_{u \in \mathbb{U}} \left\{ c(x, u) + \int_{\mathbb{X}} v(y) (\mathcal{T}(dy|x, u) - \epsilon \rho(dy)) \right\}, \quad \forall x \in \mathbb{X} \quad (2.1)$$

Then  $g_{c,\mathcal{T}}^* := \epsilon \int_{\mathbb{X}} h_{c,\mathcal{T}}^* \rho(dx)$ , and the pair  $(g_{c,\mathcal{T}}^*, h_{c,\mathcal{T}}^*)$  satisfies the ACOE equation above.

*Remark 2.3* (Average Cost Bellman Consistency Equation). Similar to the case of DCOE, under the minorization condition, for a fixed deterministic stationary policy  $\gamma$ , the operator

$$\mathbb{T}^\gamma v(x) := c(x, \gamma(x)) + \int_{\mathbb{X}} v(y) (\mathcal{T}(dy|x, \gamma(x)) - \epsilon \rho(dy)), \quad \forall x \in \mathbb{X}$$

is a contraction operator. Following the above constructive approach, we have that there exists a constant  $g^\gamma$  and a bounded measurable function  $h^\gamma$  such that

$$g^\gamma + h^\gamma(x) = c(x, \gamma(x)) + \int_{\mathbb{X}} h^\gamma(y) \mathcal{T}(dy|x, \gamma(x)),$$

where  $g^\gamma = J_\infty(c, \mathcal{T}, \gamma)$ . We refer to this equation as the average cost Bellman consistency equation.

**Assumption 2.2.** Let  $L, \tilde{L} > 0$  be a constant. Assume the following hold for an MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$ :

- (a).  $\mathbb{X}$  is compact.
- (b). There exists some  $\beta^* \in (0, 1)$  such that  $\|J_\beta^*(c, \mathcal{T})\|_{\text{Lip}} \leq L$ , for all  $\beta \in [\beta^*, 1)$ .
- (c). The map  $\mathcal{T} : \mathbb{X} \times \mathbb{U} \rightarrow \mathcal{P}(\mathbb{X})$  is  $\tilde{L}$ -Lipschitz continuous with respect to Wasserstein-1 distance, in the sense that

$$\mathcal{W}_1(\mathcal{T}(\cdot|x, u), \mathcal{T}(\cdot|y, u')) \leq \tilde{L} (d_{\mathbb{X}}(x, y) + d_{\mathbb{U}}(u, u')), \quad \forall u, u' \in \mathbb{U}, x, y \in \mathbb{X}.$$

Under Assumption 2.2, the average cost setting can be analyzed via the vanishing discount factor approximation.

**Theorem 2.3.** Suppose Assumption 2.1 and Assumption 2.2(a)-(b) hold for an MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$ . Then there exist a canonical triplet  $(g_{c,\mathcal{T}}^*, h_{c,\mathcal{T}}^*, \gamma_{c,\mathcal{T}}^*)$ , a state  $z \in \mathbb{X}$ , and an increasing subsequence  $\{\beta(n); n \geq 1\} \subset (0, 1)$  such that  $\lim_{n \rightarrow \infty} \beta(n) = 1$ , and for all  $x \in \mathbb{X}$ ,

$$\lim_{n \rightarrow \infty} (1 - \beta(n)) J_{\beta(n)}^*(c, \mathcal{T})(x) = g_{c,\mathcal{T}}^* = J_\infty^*(c, \mathcal{T}), \quad \lim_{n \rightarrow \infty} h_{c,\mathcal{T},\beta(n)}(x) = h_{c,\mathcal{T}}^*(x).$$

where  $h_{c,\mathcal{T},\beta(n)}(x) := J_{\beta(n)}^*(c, \mathcal{T})(x) - J_{\beta(n)}^*(c, \mathcal{T})(z)$ , and  $h_{c,\mathcal{T}}^*$  is  $L$ -Lipschitz.

Suppose, in addition, Assumption 2.2(c) holds. Define for each  $n \geq 1$ , and  $(x, u) \in \mathbb{X} \times \mathbb{U}$

$$\mathcal{I}_n(x, u) = \int h_{c,\mathcal{T},\beta(n)}(x') \mathcal{T}(dx'|x, u), \quad \mathcal{I}_\infty(x, u) = \int h_{c,\mathcal{T}}^*(x') \mathcal{T}(dx'|x, u).$$

Then  $\{\mathcal{I}_n : n \geq 1\}, \mathcal{I}_\infty$  are  $L \times \tilde{L}$ -Lipschitz functions on  $\mathbb{X} \times \mathbb{U}$ , and  $\lim_{n \rightarrow \infty} \mathcal{I}_n(x, u) = \mathcal{I}_\infty(x, u)$  for each  $(x, u) \in \mathbb{X} \times \mathbb{U}$ .

*Proof.* The first claim follows from the proof of Lemma 2.2 in [61]. We now focus on the second claim. Note that for each  $n \geq 1$ ,  $h_{c,\mathcal{T},\beta(n)}(z) = 0$  and  $h_{c,\mathcal{T},\beta(n)}$  is  $L$ -Lipschitz. Since  $\mathbb{X}$  is compact and  $\lim_{n \rightarrow \infty} h_{c,\mathcal{T},\beta(n)}(x) = h^*(x)$  for each  $x \in \mathbb{X}$ , by bounded convergence theorem, we have  $\lim_{n \rightarrow \infty} \mathcal{I}_n(x, u) = \mathcal{I}_\infty(x, u)$  for each  $(x, u) \in \mathbb{X} \times \mathbb{U}$ . Finally, for each  $n \geq 1$ ,  $x, y \in \mathbb{X}$  and  $u, u' \in \mathbb{U}$ , note that

$$\begin{aligned} |\mathcal{I}_n(x, u) - \mathcal{I}_n(y, u')| &= \left| \int h_{c,\mathcal{T},\beta(n)}(x') \mathcal{T}(dx'|x, u) - \int h_{c,\mathcal{T},\beta(n)}(x') \mathcal{T}(dx'|y, u') \right| \\ &\leq L \mathcal{W}_1(\mathcal{T}(\cdot|x, u), \mathcal{T}(\cdot|y, u')) \leq L \tilde{L} (d_{\mathbb{X}}(x, y) + d_{\mathbb{U}}(u, u')), \end{aligned}$$

where we use the fact that  $\|h_{c,\mathcal{T},\beta(n)}\|_{\text{Lip}}$  is bounded by  $L$  and Assumption 2.2(c). Thus,  $\mathcal{I}_n$  is Lipschitz with a constant  $L \times \tilde{L}$ . By the same argument, we have  $\mathcal{I}_\infty$  is also  $L \times \tilde{L}$ -Lipschitz. The proof is complete.  $\square$

*Remark 2.4.* By the proof of Lemma 2.2 in [61], in fact, we can show that for any increasing sequence  $\{\beta(n); n \geq 1\} \subset (0, 1)$  such that  $\lim_{n \rightarrow \infty} \beta(n) = 1$ , we can extract a further subsequence  $\{\beta(n_k); n_k \geq 1\} \subset (0, 1)$  such that the conclusions in Theorem 2.3 continue to hold for this subsequence. Further, the state  $z \in \mathbb{X}$  may be chosen arbitrarily.



**2.2. Continuity of Optimal Value Functions in Models.** We begin by establishing the continuity of the optimal value functions with respect to the model components, including the cost function and the transition kernel. These results serve as key intermediate steps in our robustness analysis in Subsection 2.3. The analysis is conducted separately for the discounted cost criterion in Subsection 2.2.1 and the average cost criterion in Subsection 2.2.2. Related continuity results can be found in [24, 25].

### 2.2.1. Discounted Cost Criterion.

**Theorem 2.4.** *Suppose that Assumption 2.1 holds for two MDPs, (reference)  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and (approximation)  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$ . Then*

$$\|J_\beta^*(c, \mathcal{T}) - J_\beta^*(\hat{c}, \mathcal{S})\|_\infty \leq \frac{1}{1-\beta} \|c - \hat{c}\|_\infty + \frac{\beta}{1-\beta} d_{J_\beta^*(c, \mathcal{T})}(\mathcal{T}, \mathcal{S}).$$

*If in addition,  $J_\beta^*(c, \mathcal{T})$  is Lipschitz continuous, then*

$$\|J_\beta^*(c, \mathcal{T}) - J_\beta^*(\hat{c}, \mathcal{S})\|_\infty \leq \frac{1}{1-\beta} \|c - \hat{c}\|_\infty + \frac{\beta}{1-\beta} \|J_\beta^*(c, \mathcal{T})\|_{\text{Lip}} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}).$$

*Proof.* For ease of notation, we use the following abbreviation:

$$v_{c, \mathcal{T}} := J_\beta^*(c, \mathcal{T}), \quad v_{\hat{c}, \mathcal{S}} := J_\beta^*(\hat{c}, \mathcal{S}). \quad (2.2)$$

By Theorem 2.1, we have that for any  $x \in \mathbb{X}$ :

$$v_{c, \mathcal{T}}(x) = \inf_{u \in \mathbb{U}} \left( c(x, u) + \beta \int v_{\mathcal{T}}(x') \mathcal{T}(dx' | x, u) \right), \quad v_{\hat{c}, \mathcal{S}}(x) = \inf_{u \in \mathbb{U}} \left( \hat{c}(x, u) + \beta \int v_{\mathcal{S}}(x') \mathcal{S}(dx' | x, u) \right).$$

Therefore, by triangular inequality, we have that for any  $x \in \mathbb{X}$ :

$$\begin{aligned} |v_{c, \mathcal{T}}(x) - v_{\hat{c}, \mathcal{S}}(x)| &\leq \sup_{u \in \mathbb{U}} |c(x, u) - \hat{c}(x, u)| + \beta \sup_{u \in \mathbb{U}} \left| \int v_{c, \mathcal{T}}(x') \mathcal{T}(dx' | x, u) - \int v_{\hat{c}, \mathcal{S}}(x') \mathcal{S}(dx' | x, u) \right| \\ &\leq \|c - \hat{c}\|_\infty + \beta \sup_{u \in \mathbb{U}} \left| \int v_{c, \mathcal{T}}(x') \mathcal{T}(dx' | x, u) - \int v_{c, \mathcal{T}}(x') \mathcal{S}(dx' | x, u) \right| \\ &\quad + \beta \sup_{u \in \mathbb{U}} \left| \int v_{c, \mathcal{T}}(x') \mathcal{S}(dx' | x, u) - \int v_{\hat{c}, \mathcal{S}}(x') \mathcal{S}(dx' | x, u) \right| \\ &\leq \|c - \hat{c}\|_\infty + \beta d_{J_\beta^*(c, \mathcal{T})}(\mathcal{T}, \mathcal{S}) + \beta \|v_{c, \mathcal{T}} - v_{\hat{c}, \mathcal{S}}\|_\infty. \end{aligned}$$

Taking the supremum over all  $x \in \mathbb{X}$ , and using the boundedness of  $v_{c, \mathcal{T}}$  and  $v_{\hat{c}, \mathcal{S}}$ , we can rearrange terms to establish the first claim. The second claim then follows from Lemma 1.1.  $\square$

**2.2.2. Average Cost Criterion.** In this subsection, we first establish an analogous result to Theorem 2.4 under the average cost criterion, assuming the minorization condition (Definition 2.1). Next, we present an alternative method—based on a vanishing discount factor and Theorem 2.4—that yields a similar result without requiring the minorization condition.

**Theorem 2.5.** *Suppose that Assumption 2.1 holds for two MDPs, (reference)  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and (approximation)  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$ . Further, assume  $\mathcal{T}$  (resp.  $\mathcal{S}$ ) satisfies the minorization condition with a probability measure  $\rho$  (resp.  $\tau$ ) and a constant  $\epsilon > 0$ . Then*

$$\|J_\infty^*(c, \mathcal{T}) - J_\infty^*(\hat{c}, \mathcal{S})\|_\infty \leq \|c - \hat{c}\|_\infty + d_{h_{c, \mathcal{T}}^*}(\mathcal{T}, \mathcal{S}),$$

*where we recall that  $h_{c, \mathcal{T}}^*$  is the unique fixed-point of  $\mathbb{T}_{c, \mathcal{T}}$  defined in (2.1). If in addition,  $h_{c, \mathcal{T}}^*$  is Lipschitz continuous, we have*

$$\|J_\infty^*(c, \mathcal{T}) - J_\infty^*(\hat{c}, \mathcal{S})\|_\infty \leq \|c - \hat{c}\|_\infty + \|h_{c, \mathcal{T}}^*\|_{\text{Lip}} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}).$$

*Proof.* Recall that  $J_\infty^*(c, \mathcal{T})(x) = g_{c, \mathcal{T}}^*$  and  $J_\infty^*(\hat{c}, \mathcal{S})(x) = g_{\hat{c}, \mathcal{S}}^*$  for  $x \in \mathbb{X}$ . By Remark 2.2, we have for  $x \in \mathbb{X}$ ,

$$\begin{aligned} h_{c, \mathcal{T}}^*(x) &= \inf_{u \in \mathbb{U}} \left( c(x, u) + \int_{\mathbb{X}} h_{c, \mathcal{T}}^*(x') (\mathcal{T}(dx'|x, u) - \epsilon \rho(dx')) \right), \quad g_{c, \mathcal{T}}^* = \epsilon \int_{\mathbb{X}} h_{c, \mathcal{T}}^*(x') \rho(dx'), \\ h_{\hat{c}, \mathcal{S}}^*(x) &= \inf_{u \in \mathbb{U}} \left( \hat{c}(x, u) + \int_{\mathbb{X}} h_{\hat{c}, \mathcal{S}}^*(x') (\mathcal{S}(dx'|x, u) - \epsilon \tau(dx')) \right), \quad g_{\hat{c}, \mathcal{S}}^* = \epsilon \int_{\mathbb{X}} h_{\hat{c}, \mathcal{S}}^*(x') \tau(dx'). \end{aligned} \quad (2.3)$$

By subtracting the second equation from the first, we have that for each  $x \in \mathbb{X}$ ,

$$|h_{c, \mathcal{T}}^*(x) - h_{\hat{c}, \mathcal{S}}^*(x)| \leq \sup_{u \in \mathbb{U}} |c(x, u) - \hat{c}(x, u)| + I_n + II_n$$

where we define

$$\begin{aligned} I_n &:= \sup_{u \in \mathbb{U}} \left| \int_{\mathbb{X}} h_{c, \mathcal{T}}^*(x') (\mathcal{T}(dx'|x, u) - \epsilon \tau(dx')) - \int_{\mathbb{X}} h_{c, \mathcal{T}}^*(x') (\mathcal{S}(dx'|x, u) - \epsilon \tau(dx')) \right|, \\ II_n &:= \sup_{u \in \mathbb{U}} \int_{\mathbb{X}} |h_{c, \mathcal{T}}^*(x') - h_{\hat{c}, \mathcal{S}}^*(x')| (\mathcal{S}(dx'|x, u) - \epsilon \tau(dx')). \end{aligned}$$

For the two terms, by definition,

$$I_n \leq d_{h_{c, \mathcal{T}}^*}(\mathcal{T}, \mathcal{S}) + \epsilon d_{h_{c, \mathcal{T}}^*}(\rho, \tau), \quad II_n \leq \|h_{c, \mathcal{T}}^* - h_{\hat{c}, \mathcal{S}}^*\|_\infty (1 - \epsilon).$$

As a result, by rearranging terms and taking the supremum over  $x \in \mathbb{X}$ , we have

$$\|h_{c, \mathcal{T}}^* - h_{\hat{c}, \mathcal{S}}^*\|_\infty \leq \frac{1}{\epsilon} \left( \|c - \hat{c}\|_\infty + d_{h_{c, \mathcal{T}}^*}(\mathcal{T}, \mathcal{S}) \right) + d_{h_{c, \mathcal{T}}^*}(\rho, \tau). \quad (2.4)$$

Finally, by the relationships between  $h_{c, \mathcal{T}}^*$  and  $g_{c, \mathcal{T}}^*$ , and between  $h_{\hat{c}, \mathcal{S}}^*$  and  $g_{\hat{c}, \mathcal{S}}^*$ , and due to the triangle inequality, we have

$$\begin{aligned} |g_{c, \mathcal{T}}^* - g_{\hat{c}, \mathcal{S}}^*| &\leq \epsilon \left| \int_{\mathbb{X}} h_{c, \mathcal{T}}^*(x') \rho(dx') - \int_{\mathbb{X}} h_{c, \mathcal{T}}^*(x') \tau(dx') \right| + \epsilon \int_{\mathbb{X}} |h_{c, \mathcal{T}}^* - h_{\hat{c}, \mathcal{S}}^*| \tau(dx') \\ &\leq \epsilon d_{h_{c, \mathcal{T}}^*}(\rho, \tau) + \epsilon \|h_{c, \mathcal{T}}^* - h_{\hat{c}, \mathcal{S}}^*\|_\infty \leq 2\epsilon d_{h_{c, \mathcal{T}}^*}(\rho, \tau) + \|c - \hat{c}\|_\infty + d_{h_{c, \mathcal{T}}^*}(\mathcal{T}, \mathcal{S}). \end{aligned}$$

Since  $\epsilon$  can be arbitrary small and  $h_{c, \mathcal{T}}^*$  is bounded, sending  $\epsilon \rightarrow 0$ , the proof of the first claim is complete. The second claim then is due to Lemma 1.1.  $\square$

Next, we present an approach based on the vanishing discounted factor approach to achieve similar results under different assumptions, utilizing Theorem 2.3.

**Theorem 2.6.** *Suppose Assumption 2.1 and Assumption 2.2(a)-(b) hold for two MDPs, (reference)  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and (approximation)  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$ . Then*

$$\|J_\infty^*(c, \mathcal{T}) - J_\infty^*(\hat{c}, \mathcal{S})\|_\infty \leq \|c - \hat{c}\|_\infty + L d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}),$$

where recall that  $L$  appears in Assumption 2.2. If additionally Assumption 2.2(c) holds for both MDPs, then

$$\|J_\infty^*(c, \mathcal{T}) - J_\infty^*(\hat{c}, \mathcal{S})\|_\infty \leq \|c - \hat{c}\|_\infty + d_{h_{c, \mathcal{T}}^*}(\mathcal{T}, \mathcal{S}),$$

where  $h_{c, \mathcal{T}}^*$  appears in Theorem 2.3.

*Proof.* By Theorem 2.4 and due to Assumption 2.2(b), for any  $\beta \in [\beta^*, 1)$ , we have

$$\begin{aligned} \|(1 - \beta) (J_\beta^*(c, \mathcal{T}) - J_\beta^*(\hat{c}, \mathcal{S}))\|_\infty &\leq \|c - \hat{c}\|_\infty + \beta L d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}), \\ \|(1 - \beta) (J_\beta^*(c, \mathcal{T}) - J_\beta^*(\hat{c}, \mathcal{S}))\|_\infty &\leq \|c - \hat{c}\|_\infty + \beta d_{J_\beta^*(c, \mathcal{T})}(\mathcal{T}, \mathcal{S}). \end{aligned} \quad (2.5)$$

By Theorem 2.3 and Remark 2.4, there exist *canonical triplets*  $(g_{c,\mathcal{T}}^*, h_{c,\mathcal{T}}^*, \gamma_{c,\mathcal{T}}^*)$  and  $(g_{\hat{c},\mathcal{S}}^*, h_{\hat{c},\mathcal{S}}^*, \gamma_{\hat{c},\mathcal{S}}^*)$ , a state  $z \in \mathbb{X}$ , and an increasing subsequence  $\{\beta(n); n \geq 1\} \subset (0, 1)$  and  $z \in \mathbb{X}$  such that  $\lim_{n \rightarrow \infty} \beta(n) = 1$ , and for all  $x \in \mathbb{X}$ ,

$$\begin{aligned} \lim_{n \rightarrow \infty} (1 - \beta(n)) J_{\beta(n)}^*(c, \mathcal{T})(x) &= g_{c,\mathcal{T}}^* = J_\infty^*(c, \mathcal{T}), & \lim_{n \rightarrow \infty} h_{c,\mathcal{T},\beta(n)}(x) &= h_{c,\mathcal{T}}^*(x), \\ \lim_{n \rightarrow \infty} (1 - \beta(n)) J_{\beta(n)}^*(\hat{c}, \mathcal{S})(x) &= g_{\hat{c},\mathcal{S}}^* = J_\infty^*(\hat{c}, \mathcal{S}), & \lim_{n \rightarrow \infty} h_{\hat{c},\mathcal{S},\beta(n)}(x) &= h_{\hat{c},\mathcal{S}}^*(x). \end{aligned}$$

where  $h_{c,\mathcal{T},\beta(n)}(x) := J_{\beta(n)}^*(c, \mathcal{T})(x) - J_{\beta(n)}^*(c, \mathcal{T})(z)$  and  $h_{\hat{c},\mathcal{S},\beta(n)}(x) := J_{\beta(n)}^*(\hat{c}, \mathcal{S})(x) - J_{\beta(n)}^*(\hat{c}, \mathcal{S})(z)$ .

Thus, by taking the limit of the first equation in (2.5) along the subsequence  $\{\beta(n); n \geq 1\}$ , we immediately obtain the first claim.

Further, by taking the limit of the second equation in (2.5) along the subsequence, we have

$$\|J_\infty^*(c, \mathcal{T}) - J_\infty^*(\hat{c}, \mathcal{S})\|_\infty \leq \|c - \hat{c}\|_\infty + \limsup_{n \rightarrow \infty} d_{J_{\beta(n)}^*}(c, \mathcal{T}) = \|c - \hat{c}\|_\infty + \limsup_{n \rightarrow \infty} \sup_{(x,u) \in \mathbb{X} \times \mathbb{U}} |\widetilde{\mathcal{I}}_n(x, u)|,$$

where we define

$$\begin{aligned} \widetilde{\mathcal{I}}_n(x, u) &:= \int h_{c,\mathcal{T},\beta(n)}(x') \mathcal{T}(dx'|x, u) - \int h_{c,\mathcal{T},\beta(n)}(x') \mathcal{S}(dx'|x, u), \\ \widetilde{\mathcal{I}}_\infty(x, u) &:= \int h_{c,\mathcal{T}}^*(x') \mathcal{T}(dx'|x, u) - \int h_{c,\mathcal{T}}^*(x') \mathcal{S}(dx'|x, u). \end{aligned}$$

By Theorem 2.3, since Assumption 2.2(c) is imposed for the second claim, we have: (i).  $\mathbb{X}$  is compact; (ii).  $\lim_{n \rightarrow \infty} \widetilde{\mathcal{I}}_n(x, u) = \widetilde{\mathcal{I}}_\infty(x, u)$  for each  $(x, u) \in \mathbb{X} \times \mathbb{U}$ ; (iii).  $\{|\widetilde{\mathcal{I}}_n|; n \geq 1\}, |\widetilde{\mathcal{I}}_\infty|$  are  $L \times \widetilde{L}$ -Lipschitz functions on  $\mathbb{X} \times \mathbb{U}$ , and thus equicontinuous. As a result, we have

$$\lim_{n \rightarrow \infty} \sup_{(x,u) \in \mathbb{X} \times \mathbb{U}} |\widetilde{\mathcal{I}}_n(x, u)| = \sup_{(x,u) \in \mathbb{X} \times \mathbb{U}} |\widetilde{\mathcal{I}}_\infty(x, u)| = d_{h_{c,\mathcal{T}}^*}(\mathcal{T}, \mathcal{S}).$$

The proof is then complete.  $\square$

Note that the first claim in Theorem 2.6 does not require Assumption 2.2(c), while the second claim imposes it for *both* MDPs.

**2.3. Bounds on Robustness Errors due to Model Misspecification.** We now use the continuity results discussed in Section 2.2 to establish our main results: an upper bound on the performance loss incurred by applying a policy that is optimal with respect to an *approximate* MDP. This bound is expressed in terms of the supremum norm between cost functions and two types of distances between the transition kernels: one based on their discrepancy with respect to optimal value functions, and the other given by the uniform Wasserstein-1 distance.

**2.3.1. Discounted Cost Criterion.** We denote by  $\gamma_{\hat{c},\mathcal{S},\beta}^*$  the optimal policy learned under an approximate MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$  under the  $\beta$ -discounted cost criterion in the subsection.

**Theorem 2.7.** *Suppose that Assumption 2.1 holds for two MDPs, (reference)  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and (approximation)  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$ . Then*

$$\begin{aligned} \|J_\beta(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S},\beta}^*) - J_\beta^*(c, \mathcal{T})\|_\infty &\leq \frac{2}{(1-\beta)^2} \|c - \hat{c}\|_\infty + \frac{2\beta}{(1-\beta)^2} d_{J_\beta^*(c,\mathcal{T})}(\mathcal{T}, \mathcal{S}), \\ \|J_\beta(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S},\beta}^*) - J_\beta^*(c, \mathcal{T})\|_\infty &\leq \frac{2}{1-\beta} \|c - \hat{c}\|_\infty + \frac{\beta}{1-\beta} \left( d_{J_\beta^*(c,\mathcal{T})}(\mathcal{T}, \mathcal{S}) + d_{J_\beta^*(\hat{c},\mathcal{S})}(\mathcal{T}, \mathcal{S}) \right), \\ \|J_\beta(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S},\beta}^*) - J_\beta^*(c, \mathcal{T})\|_\infty &\leq \frac{2}{1-\beta} \|c - \hat{c}\|_\infty + \frac{2\beta}{1-\beta} d_{J_\beta^*(\hat{c},\mathcal{S})}(\mathcal{T}, \mathcal{S}). \end{aligned}$$

*Proof.* For ease of notation, we adopt the abbreviations in (2.2), and let  $v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}}(x) := J_\beta(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S},\beta}^*)(x)$  for  $x \in \mathbb{X}$ . By the triangle inequality, we have  $\|v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - v_{c,\mathcal{T}}\|_\infty \leq \|v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - v_{\hat{c},\mathcal{S}}^{\hat{c},\mathcal{S}}\|_\infty + \|v_{\hat{c},\mathcal{S}}^{\hat{c},\mathcal{S}} - v_{c,\mathcal{T}}\|_\infty$ . Then, by Theorem 2.4 and the symmetry between  $\mathcal{T}$  and  $\mathcal{S}$ , we have

$$\begin{aligned} \|v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - v_{c,\mathcal{T}}\|_\infty &\leq \|v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - v_{\hat{c},\mathcal{S}}^{\hat{c},\mathcal{S}}\|_\infty + \frac{1}{1-\beta}\|c - \hat{c}\|_\infty + \frac{\beta}{1-\beta}d_{J_\beta^*(c,\mathcal{T})}(\mathcal{T}, \mathcal{S}) \\ \|v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - v_{c,\mathcal{T}}\|_\infty &\leq \|v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - v_{\hat{c},\mathcal{S}}^{\hat{c},\mathcal{S}}\|_\infty + \frac{1}{1-\beta}\|c - \hat{c}\|_\infty + \frac{\beta}{1-\beta}d_{J_\beta^*(\hat{c},\mathcal{S})}(\mathcal{T}, \mathcal{S}). \end{aligned} \quad (2.6)$$

Next, we bound  $\|v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - v_{\hat{c},\mathcal{S}}^{\hat{c},\mathcal{S}}\|_\infty$ . By the discounted cost Bellman consistency equation, we have:

$$\begin{aligned} v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}}(x) &= c(x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x)) + \beta \int v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}}(x') \mathcal{T}(dx'|x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x)), \text{ for } x \in \mathbb{X}, \\ v_{\hat{c},\mathcal{S}}^{\hat{c},\mathcal{S}}(x) &= \hat{c}(x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x)) + \beta \int v_{\hat{c},\mathcal{S}}^{\hat{c},\mathcal{S}}(x') \mathcal{S}(dx'|x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x)), \text{ for } x \in \mathbb{X}. \end{aligned}$$

Taking the difference between them, we obtain

$$\begin{aligned} |v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}}(x) - v_{\hat{c},\mathcal{S}}^{\hat{c},\mathcal{S}}(x)| &\leq \|c - \hat{c}\|_\infty + \beta \left| \int (v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}}(x') - v_{\hat{c},\mathcal{S}}^{\hat{c},\mathcal{S}}(x')) \mathcal{T}(dx'|x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x)) \right| + \beta I_n \\ &\leq \|c - \hat{c}\|_\infty + \beta \|v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - v_{c,\mathcal{T}}\|_\infty + \beta I_n \end{aligned}$$

where we define

$$I_n := \left| \int v_{\hat{c},\mathcal{S}}^{\hat{c},\mathcal{S}}(x') (\mathcal{T}(dx'|x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x)) - \mathcal{S}(dx'|x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x))) \right|.$$

**Approach 1** We apply the following bound:  $I_n \leq d_{J_\beta^*(\hat{c},\mathcal{S})}(\mathcal{T}, \mathcal{S})$ . By taking the supremum over  $x \in \mathbb{X}$  and rearranging the above terms, we have

$$\|v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - v_{c,\mathcal{T}}\|_\infty \leq \frac{1}{1-\beta}\|c - \hat{c}\|_\infty + \frac{\beta}{1-\beta}d_{J_\beta^*(\hat{c},\mathcal{S})}(\mathcal{T}, \mathcal{S}).$$

Then, in view of (2.6), the second and third claims follow immediately.

**Approach 2** Note that by the triangle inequality,

$$\begin{aligned} I_n &\leq \left| \int v_{c,\mathcal{T}}(x') (\mathcal{T}(dx'|x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x)) - \mathcal{S}(dx'|x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x))) \right| \\ &\quad + \left| \int (v_{\hat{c},\mathcal{S}}(x') - v_{c,\mathcal{T}}(x')) (\mathcal{T}(dx'|x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x)) - \mathcal{S}(dx'|x, \gamma_{\hat{c},\mathcal{S},\beta}^*(x))) \right| \\ &\leq d_{J_\beta^*(c,\mathcal{T})}(\mathcal{T}, \mathcal{S}) + 2\|v_{\hat{c},\mathcal{S}} - v_{c,\mathcal{T}}\|_\infty. \end{aligned}$$

Now, by applying Theorem 2.4 to the last term above, taking the supremum over  $x \in \mathbb{X}$  and rearranging the above terms, we have

$$\|v_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - v_{c,\mathcal{T}}\|_\infty \leq \frac{1+\beta}{(1-\beta)^2}\|c - \hat{c}\|_\infty + \frac{\beta+\beta^2}{(1-\beta)^2}d_{J_\beta^*(c,\mathcal{T})}(\mathcal{T}, \mathcal{S}).$$

Then, in view of (2.6), the first claim follows.  $\square$

The first result in Theorem 2.7 only depends on the optimal value function  $J_\beta^*(c, \mathcal{T})$  under the reference MDP, while the second and third results also involve the optimal value function  $J_\beta^*(\hat{c}, \mathcal{S})$  under the approximating MDP. However, the first result exhibits a worse dependence on  $\beta$  compared to the others. In particular, to apply the discounted vanishing factor approach to the average cost setting, the upper bound must scale as  $1/(1-\beta)$  rather than  $1/(1-\beta)^2$ .

*Remark 2.5.* If  $J_\beta^*(c, \mathcal{T})$  and  $J_\beta^*(\hat{c}, \mathcal{S})$  are Lipschitz, by Lemma 1.1, we have

$$d_{J_\beta^*(c, \mathcal{T})}(\mathcal{T}, \mathcal{S}) \leq \|J_\beta^*(c, \mathcal{T})\|_{\text{Lip}} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}), \quad d_{J_\beta^*(\hat{c}, \mathcal{S})}(\mathcal{T}, \mathcal{S}) \leq \|J_\beta^*(\hat{c}, \mathcal{S})\|_{\text{Lip}} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}).$$

Using this loose bound, the third inequality becomes:

$$\|J_\beta(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}, \beta}^*) - J_\beta^*(c, \mathcal{T})\|_\infty \leq \frac{2}{1-\beta} \|c - \hat{c}\|_\infty + \frac{2\beta}{1-\beta} \|J_\beta^*(\hat{c}, \mathcal{S})\|_{\text{Lip}} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S})$$

This recovers part (2) of Corollary 1 in [31].

**2.3.2. Average Cost Criterion.** In this subsection, we consider the average cost criterion and bound the performance loss incurred by applying a policy that is optimal with respect to an *approximate* MDP. As in Subsection 2.2.2, we first consider the minorization approach, and then the vanishing discount factor approach.

We denote by  $\gamma_{\hat{c}, \mathcal{S}}^*$  the optimal policy learned under an approximate MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$  under the *average* cost criterion in the subsection.

**Theorem 2.8.** *Suppose that Assumption 2.1 holds for two MDPs, (reference)  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and (approximation)  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$ . Further, assume  $\mathcal{T}$  (resp.  $\mathcal{S}$ ) satisfies the minorization condition with a probability measure  $\rho$  (resp.  $\tau$ ) and a constant  $\epsilon > 0$ . Then*

$$\|J_\infty(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*) - J_\infty^*(c, \mathcal{T})\|_\infty \leq \frac{2+\epsilon}{\epsilon} \left( \|c - \hat{c}\|_\infty + d_{h_{c, \mathcal{T}}^*}(\mathcal{T}, \mathcal{S}) \right) + (2+\epsilon) d_{h_{c, \mathcal{T}}^*}(\rho, \tau).$$

where we recall that  $h_{c, \mathcal{T}}^*$  is the unique fixed-point of  $\mathbb{T}_{c, \mathcal{T}}$  defined in (2.1).

*Proof.* By Theorem 2.2, under the minorization condition for both MDPs, there exist canonical triplets  $(g_{c, \mathcal{T}}^*, h_{c, \mathcal{T}}^*, \gamma_{c, \mathcal{T}}^*)$  and  $(g_{\hat{c}, \mathcal{S}}^*, h_{\hat{c}, \mathcal{S}}^*, \gamma_{\hat{c}, \mathcal{S}}^*)$  that satisfy (2.3), and further

$$h_{\hat{c}, \mathcal{S}}^*(x) = \hat{c}(x, \gamma_{c, \mathcal{T}}^*(x)) + \int_{\mathbb{X}} h_{\hat{c}, \mathcal{S}}^*(y) (\mathcal{S}(dy|x, \gamma_{c, \mathcal{T}}^*(x)) - \epsilon \tau(dy)), \quad \forall x \in \mathbb{X}, \quad (2.7)$$

Further, by Remark 2.3, there exists a triplet  $(g_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}}, h_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}}, \gamma_{\hat{c}, \mathcal{S}}^*)$  such that

$$\begin{aligned} h_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}}(x) &= c(x, \gamma_{\hat{c}, \mathcal{S}}^*(x)) + \int h_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}}(x') (\mathcal{T}(dx'|x, \gamma_{\hat{c}, \mathcal{S}}^*(x)) - \epsilon \rho(dx')) \quad \text{for } x \in \mathbb{X}, \\ g_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}} &= \epsilon \int_{\mathbb{X}} h_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}}(x') \rho(dx') = J_\infty(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*)(x) \quad \text{for } x \in \mathbb{X}. \end{aligned} \quad (2.8)$$

By the triangle inequality and Theorem 2.5, we have:

$$|g_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}} - g_{c, \mathcal{T}}^*| \leq |g_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}} - g_{\hat{c}, \mathcal{S}}^*| + |g_{\hat{c}, \mathcal{S}}^* - g_{c, \mathcal{T}}^*| \leq |g_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}} - g_{\hat{c}, \mathcal{S}}^*| + \|c - \hat{c}\|_\infty + d_{h_{c, \mathcal{T}}^*}(\mathcal{T}, \mathcal{S}).$$

Further, by (2.3) and (2.8), we have

$$\begin{aligned} |g_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}} - g_{\hat{c}, \mathcal{S}}^*| &= \epsilon \left| \int_{\mathbb{X}} h_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}}(x') \rho(dx') - \int_{\mathbb{X}} h_{\hat{c}, \mathcal{S}}^*(x') \tau(dx') \right| \\ &\leq \epsilon \|h_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}} - h_{\hat{c}, \mathcal{S}}^*\|_\infty + \epsilon \left| \int_{\mathbb{X}} h_{\hat{c}, \mathcal{S}}^*(x') \rho(dx') - \int_{\mathbb{X}} h_{\hat{c}, \mathcal{S}}^*(x') \tau(dx') \right| \\ &\leq \epsilon \|h_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}} - h_{\hat{c}, \mathcal{S}}^*\|_\infty + \epsilon \|h_{c, \mathcal{T}}^* - h_{\hat{c}, \mathcal{S}}^*\|_\infty + \epsilon d_{h_{c, \mathcal{T}}^*}(\rho, \tau). \end{aligned} \quad (2.9)$$

Note that we establish an upper bound for  $\|h_{c, \mathcal{T}}^* - h_{\hat{c}, \mathcal{S}}^*\|_\infty$  in (2.4). Thus it suffices to focus on the term  $\|h_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}} - h_{\hat{c}, \mathcal{S}}^*\|_\infty$ . By subtracting the fixed-point equation in (2.8) from (2.7), we obtain that for  $x \in \mathbb{X}$ ,

$$|h_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}}(x) - h_{\hat{c}, \mathcal{S}}^*(x)| \leq \|c - \hat{c}\|_\infty + I_n + II_n + III_n,$$

where we define

$$\begin{aligned} I_n &:= \left| \int \left( h_{c,\mathcal{T}}^{\hat{c},\mathcal{S}}(x') - h_{\hat{c},\mathcal{S}}^*(x') \right) \left( \mathcal{T}(dx'|x, \gamma_{\mathcal{S}}^*(x)) - \epsilon \rho(dx') \right) \right|, \\ II_n &:= \left| \int \left( h_{\hat{c},\mathcal{S}}^*(x') - h_{c,\mathcal{T}}^*(x') \right) \left( \mathcal{T}(dx'|x, \gamma_{\mathcal{S}}^*(x)) - \mathcal{S}(dx'|x, \gamma_{\mathcal{S}}^*(x)) + \epsilon \tau(dx') - \epsilon \rho(dx') \right) \right|, \\ III_n &:= \left| \int h_{c,\mathcal{T}}^*(x') \left( \mathcal{T}(dx'|x, \gamma_{\mathcal{S}}^*(x)) - \mathcal{S}(dx'|x, \gamma_{\mathcal{S}}^*(x)) + \epsilon \tau(dx') - \epsilon \rho(dx') \right) \right|. \end{aligned}$$

By definition, we have

$$I_n \leq (1 - \epsilon) \left\| h_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - h_{\hat{c},\mathcal{S}}^* \right\|_{\infty}, \quad II_n \leq (2 - 2\epsilon) \|h_{c,\mathcal{T}}^* - h_{\hat{c},\mathcal{S}}^*\|_{\infty}, \quad III_n \leq d_{h_{c,\mathcal{T}}^*}(\mathcal{T}, \mathcal{S}) + \epsilon d_{h_{c,\mathcal{T}}^*}(\rho, \tau).$$

Taking the supremum over all  $x \in \mathbb{X}$ , and rearranging terms, we have

$$\epsilon \left\| h_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - h_{\hat{c},\mathcal{S}}^* \right\|_{\infty} \leq \|c - \hat{c}\|_{\infty} + (2 - 2\epsilon) \|h_{c,\mathcal{T}}^* - h_{\hat{c},\mathcal{S}}^*\|_{\infty} + d_{h_{c,\mathcal{T}}^*}(\mathcal{T}, \mathcal{S}) + \epsilon d_{h_{c,\mathcal{T}}^*}(\rho, \tau).$$

which, due to (2.9), implies that

$$\left\| g_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} - g_{\hat{c},\mathcal{S}}^* \right\|_{\infty} \leq \|c - \hat{c}\|_{\infty} + (2 - \epsilon) \|h_{c,\mathcal{T}}^* - h_{\hat{c},\mathcal{S}}^*\|_{\infty} + d_{h_{c,\mathcal{T}}^*}(\mathcal{T}, \mathcal{S}) + 2\epsilon d_{h_{c,\mathcal{T}}^*}(\rho, \tau).$$

Applying (2.4) to bound  $\|h_{c,\mathcal{T}}^* - h_{\hat{c},\mathcal{S}}^*\|_{\infty}$  completes the proof.  $\square$

Note that in the above theorem, only the optimal relative value function  $h_{c,\mathcal{T}}^*$  of the reference MDP is involved, and not that of the approximate MDP

Next, we extend Theorem 2.7 to the average cost criterion using the vanishing discount factor approach.

**Theorem 2.9.** *Suppose Assumption 2.1 and Assumption 2.2(a)-(b) hold for two MDPs, (reference)  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and (approximation)  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$ . Further, assume that for each  $x \in \mathbb{X}$ , there exists a sequence  $\beta_n(x) \uparrow 1$  such that  $(1 - \beta_n(x))J_{\beta_n(x)}(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S}}^*)(x) \rightarrow g_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} = J_{\infty}(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S}}^*)(x)$ . Then*

$$\left\| J_{\infty}(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S}}^*) - J_{\infty}^*(c, \mathcal{T}) \right\|_{\infty} \leq 2\|c - \hat{c}\|_{\infty} + 2Ld_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}),$$

where recall that  $L$  appears in Assumption 2.2. If additionally Assumption 2.2(c) holds for both MDPs, then

$$\left\| J_{\infty}(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S}}^*) - J_{\infty}^*(c, \mathcal{T}) \right\|_{\infty} \leq 2\|c - \hat{c}\|_{\infty} + \left( d_{h_{c,\mathcal{T}}^*}(\mathcal{T}, \mathcal{S}) + d_{h_{\hat{c},\mathcal{S}}^*}(\mathcal{T}, \mathcal{S}) \right),$$

where  $h_{c,\mathcal{T}}^*$  and  $h_{\hat{c},\mathcal{S}}^*$  are defined as in Theorem 2.3.

*Proof.* The proof proceeds in a similar manner as that of Theorem 2.6, with the modification that for the discounted cost result, we use Theorem 2.7 instead of Theorem 2.4.  $\square$

We note that, in general, no assumption necessarily ensures that  $J_{\beta}(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S}}^*)$  is Lipschitz. Consequently, it is not immediately clear under what conditions there exists, for any  $x \in \mathbb{X}$ , a sequence  $\beta_n(x) \rightarrow 1$  such that  $(1 - \beta_n(x))J_{\beta_n(x)}(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S}}^*)(x) \rightarrow g_{c,\mathcal{T}}^{\hat{c},\mathcal{S}}$ . We now present an assumption concerning the ergodicity of the approximate kernel  $\mathcal{S}$ , under which this result holds.

**Assumption 2.3.** For the reference controlled Markov process  $(\mathbb{X}, \mathbb{U}, \mathcal{T})$ , assume that for any deterministic stationary policy  $\gamma$ , the induced Markov chain with transition kernel  $\mathcal{T}^{\gamma}(\cdot|x) := \mathcal{T}(\cdot|x, \gamma(x))$  for  $x \in \mathbb{X}$  is positive Harris recurrent.

**Lemma 2.1.** *Suppose Assumption 2.1 and Assumption 2.2(a)-(b) hold for two MDPs, (reference)  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and (approximation)  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$ . If, in addition, Assumption 2.3 holds, then*

$$\lim_{\beta \uparrow 1} (1 - \beta)J_{\beta}(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S}}^*)(x) = g_{c,\mathcal{T}}^{\hat{c},\mathcal{S}} = J_{\infty}(c, \mathcal{T}, \gamma_{\hat{c},\mathcal{S}}^*)(x).$$



*Proof.* Let  $\{X_n : n \geq 0\}$  be a Markov chain with the state space  $\mathbb{X}$  and transition kernel  $\mathcal{T}^{\gamma_{\hat{c}, \mathcal{S}}^*}(\cdot | x) := \mathcal{T}(\cdot | x, \gamma_{\hat{c}, \mathcal{S}}^*(x))$  for  $x \in \mathbb{X}$ . Due to Assumption 2.3 and Assumption 2.1(b), and by Theorem 17.1.7 of [67], for any  $x \in \mathbb{X}$ :

$$\mathbb{P} \left( \lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{i=0}^n c(X_i, \gamma_{\hat{c}, \mathcal{S}}^*(X_i)) = \int_{\mathbb{X}} c(y, \gamma_{\hat{c}, \mathcal{S}}^*(y)) \pi(dy) \mid X_0 = x \right) = 1,$$

where  $\pi$  is an invariant probability measure of  $\{X_n : n \geq 0\}$ . Furthermore, since  $c(\cdot)$  is bounded, by the bounded convergence theorem and the definition of the average cost criterion, we have

$$J_{\infty}(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}})(x) = \lim_{n \rightarrow \infty} \frac{1}{n+1} \mathbb{E} \left[ \sum_{i=0}^n c(X_i, \gamma_{\hat{c}, \mathcal{S}}^*(X_i)) \mid X_0 = x \right] = \int_{\mathbb{X}} c(y, \gamma_{\hat{c}, \mathcal{S}}^*(y)) \pi(dy) := g_{\hat{c}, \mathcal{T}}^{\mathcal{S}}.$$

By definition, for each  $\beta \in (0, 1)$  and  $x \in \mathbb{X}$ ,

$$(1 - \beta) J_{\beta}(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*)(x) = (1 - \beta) \sum_{t=0}^{\infty} \beta^t \mathbb{E} [c(X_t, \gamma_{\hat{c}, \mathcal{S}}^*(X_t)) \mid X_0 = x].$$

Thus, for each  $x \in \mathbb{X}$ , by the Abelian inequality in Lemma 2.2 with  $a_t = \mathbb{E} [c(X_t, \gamma_{\hat{c}, \mathcal{S}}^*(X_t)) \mid X_0 = x]$ ,

$$\lim_{\beta \uparrow 1} (1 - \beta) J_{\beta}(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*)(x) = \int_{\mathbb{X}} c(y, \gamma_{\hat{c}, \mathcal{S}}^*(y)) \pi(dy).$$

Then the proof is complete.  $\square$

**Lemma 2.2** (Abelian Inequality, Lemma 5.3.1 of [68]). *Let  $\{a_t\}_{t \in \mathbb{N}}$  be a sequence of nonnegative numbers and  $\beta \in (0, 1)$ . Then*

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} a_t \leq \liminf_{\beta \rightarrow 1} (1 - \beta) \sum_{t=0}^{\infty} \beta^t a_t \leq \limsup_{\beta \rightarrow 1} (1 - \beta) \sum_{t=0}^{\infty} \beta^t a_t \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} a_t.$$

**2.4. Lipschitz Regularity of optimal value functions.** When applying previous results, it is often necessary to establish that the optimal (relative) value functions for the reference MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$ —namely,  $J_{\beta}^*(c, \mathcal{T})$  and  $h_{c, \mathcal{T}}^*$  as defined in (2.1)—are Lipschitz continuous, and to obtain an upper bound on their Lipschitz constants. In this section, we present a sufficient condition to achieve this. To our knowledge, these results are first introduced in [59], and one of the main results is also presented in [43]. For the case of average cost, see [61].

**Lemma 2.3** (Theorem 4.37 of [43]). *For an MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and a discount factor  $\beta < 1$  satisfying Assumption 1.1,  $J_{\beta}^*(c, \mathcal{T})$  is  $\frac{\|c\|_{Lip}}{1 - \beta \|\mathcal{T}\|_{Lip}}$ -Lipschitz.*

By combining Lemma 2.3 with Theorem 2.4 and Theorem 2.7, we immediately have the following corollary for the  $\beta$ -discounted cost case.

**Corollary 2.1.** *Consider two MDPs, (reference)  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and (approximation)  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$ . Suppose that the reference MDP satisfies Assumption 1.1, and that the approximate MDP satisfies Assumption 2.1. Then*

$$\begin{aligned} \|J_{\beta}^*(c, \mathcal{T}) - J_{\beta}^*(\hat{c}, \mathcal{S})\|_{\infty} &\leq \frac{1}{1 - \beta} \|c - \hat{c}\|_{\infty} + \frac{\beta \|c\|_{Lip}}{(1 - \beta)(1 - \beta \|\mathcal{T}\|_{Lip})} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}), \\ \|J_{\beta}(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*) - J_{\beta}^*(c, \mathcal{T})\|_{\infty} &\leq \frac{2}{(1 - \beta)^2} \|c - \hat{c}\|_{\infty} + \frac{2\beta \|c\|_{Lip}}{(1 - \beta)^2(1 - \beta \|\mathcal{T}\|_{Lip})} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}). \end{aligned}$$

If, in addition, the approximate MDP satisfies Assumption 1.1, then

$$\|J_\beta(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}, \beta}^*) - J_\beta^*(c, \mathcal{T})\|_\infty \leq \frac{2}{1-\beta} \|c - \hat{c}\|_\infty + \frac{\beta}{1-\beta} \left( \frac{\|c\|_{Lip}}{1-\beta \|\mathcal{T}\|_{Lip}} + \frac{\|\hat{c}\|_{Lip}}{1-\beta \|\mathcal{S}\|_{Lip}} \right) d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}).$$

We now present results for the average cost criterion case.

**Assumption 2.4.** In addition to Assumption 1.1, assume  $\|\mathcal{T}\|_{Lip} < 1$ .

The following is an analog of Lemma 2.3 for the average cost criterion; see [61, Lemma 2.2] or [69, Theorem 3.5].

**Lemma 2.4.** Consider an MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  satisfying Assumption 2.4. Assume that  $\mathcal{T}$  satisfies the minorization condition with a probability measure  $\rho$  and a constant  $\epsilon > 0$ . Then  $h_{c, \mathcal{T}}^*$  as defined in (2.1) is  $\frac{\|c\|_{Lip}}{1-\|\mathcal{T}\|_{Lip}}$ -Lipschitz.

By combining Lemma 2.4 with Theorem 2.5 and Theorem 2.8, we immediately have the following corollaries for the average cost case under minorization conditions.

**Corollary 2.2.** Consider two MDPs, (reference)  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and (approximation)  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$ . Suppose that the reference satisfies Assumption 2.4, and the approximation satisfies Assumption 2.1. Further, assume  $\mathcal{T}$  (resp.  $\mathcal{S}$ ) satisfies the minorization condition with a probability measure  $\rho$  (resp.  $\tau$ ) and a constant  $\epsilon > 0$ . Then

$$\begin{aligned} \|J_\infty^*(c, \mathcal{T}) - J_\infty^*(\hat{c}, \mathcal{S})\|_\infty &\leq \|c - \hat{c}\|_\infty + \frac{\|c\|_{Lip}}{1-\|\mathcal{T}\|_{Lip}} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}). \\ \|J_\infty(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*) - J_\infty^*(c, \mathcal{T})\|_\infty &\leq \frac{2+\epsilon}{\epsilon} \left( \|c - \hat{c}\|_\infty + \frac{\|c\|_{Lip}}{1-\|\mathcal{T}\|_{Lip}} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}) + \frac{\epsilon \|c\|_{Lip}}{1-\|\mathcal{T}\|_{Lip}} d_{\mathcal{W}_1}(\rho, \tau) \right). \end{aligned}$$

Now, we consider the vanishing discounted factor approach under the average cost criterion. By Lemma 2.3 and due to Assumption 2.4, the Assumption 2.2(b) holds with  $L = \frac{\|c\|_{Lip}}{1-\|\mathcal{T}\|_{Lip}}$ . Then, due to Theorem 2.6 and Theorem 2.9, we have the following corollary.

**Corollary 2.3.** Suppose Assumptions 2.4 hold for two MDPs, (reference)  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  and (approximation)  $(\mathbb{X}, \mathbb{U}, \mathcal{S}, \hat{c})$ . Further, assume  $\mathbb{X}$  is compact. Then

$$\|J_\infty^*(c, \mathcal{T}) - J_\infty^*(\hat{c}, \mathcal{S})\|_\infty \leq \|c - \hat{c}\|_\infty + \frac{\|c\|_{Lip}}{1-\|\mathcal{T}\|_{Lip}} d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}).$$

If, in addition, for each  $x \in \mathbb{X}$ , there exists a sequence  $\beta_n(x) \uparrow 1$  such that  $(1-\beta_n(x))J_{\beta_n(x)}(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*)(x) \rightarrow g_{c, \mathcal{T}}^{\hat{c}, \mathcal{S}} = J_\infty(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*)(x)$ , then

$$\|J_\infty(c, \mathcal{T}, \gamma_{\hat{c}, \mathcal{S}}^*) - J_\infty^*(c, \mathcal{T})\|_\infty \leq 2\|c - \hat{c}\|_\infty + \left( \frac{\|c\|_{Lip}}{1-\|\mathcal{T}\|_{Lip}} + \frac{\|\hat{c}\|_{Lip}}{1-\|\mathcal{S}\|_{Lip}} \right) d_{\mathcal{W}_1}(\mathcal{T}, \mathcal{S}).$$

### 3. APPLICATION TO MODEL LEARNING FROM DATA AND SAMPLE COMPLEXITY

In this section, we introduce a more general learning framework in which the model itself is learned from data. We begin by reviewing model approximation with finite model representations, noting that some cases fall within the scope of our general robustness results. We then propose learning algorithms to estimate the quantized model from data, establishing sample complexity bounds. The generality of these results appears to be novel in the literature.

**3.1. Quantized Approximations.** In this section, we briefly review the state quantization scheme, first introduced, to the best of our knowledge, in [70, Section 3] (see also [48, Section 2.3]). These results will also be useful when developing two general learning algorithms in the following section. We demonstrate that the near-optimality of such quantization, as shown for the discounted cost problem in Theorem 6 of [48] (which slightly refines [43, Theorem 4.38]) and for the average cost problem in Theorem 3.5 of [69], can be considered as special cases of Corollary 2.1 and Corollary 2.2, respectively. For simplicity, we assume that  $\mathbb{U}$  is a finite set, which does not restrict generality if we consider a compact action space (see Chapter 3 of [43]).

Consider a state space  $\mathbb{X}$  and an  $M$ -partition of it,  $\{B_i\}_{i=1}^M$ . For each partition  $B_i$ , we pick some representative element  $y_i \in B_i$ . A quantizer is then a map  $q : \mathbb{X} \rightarrow \{y_i\}_{i=1}^M$ , where

$$q(x) = y_i \quad \text{if and only if} \quad x \in B_i.$$

Given a weighting measure  $\pi \in \mathcal{P}(\mathbb{X})$ , assuming  $\pi(B_i) > 0$  for  $i \in [M] = \{1, \dots, M\}$ ,

we define a new finite state MDP,  $(\{y_i\}_{i=1}^M, \mathbb{U}, \mathcal{T}^{M,\pi}, c^{\pi,M})$ , where  $c^{\pi,M}$  and  $\mathcal{T}^{M,\pi}$  are defined as follows: for  $i, j \in [M]$  and  $u \in \mathbb{U}$ ,

$$c^{\pi,M}(y_i, u) := \frac{\int_{B_i} c(x, u) \pi(dx)}{\int_{B_i} \pi(dx)}, \quad \mathcal{T}^{\pi,M}(y_j | y_i, u) := \frac{\int_{B_i} \mathcal{T}(B_j | x, u) \pi_i(dx)}{\int_{B_i} \pi(dx)}. \quad (3.1)$$

We denote by  $\gamma_\beta^{\pi,M}$  and  $\gamma_\infty^{\pi,M}$  the optimal policy for the  $\beta$ -discounted and the average cost MDP  $(\{y_i\}_{i=1}^M, \mathbb{U}, \mathcal{T}^{\pi,M}, c^{\pi,M})$ , respectively. Next, we define the piecewise constant extension of such a finite model to the original state space as in [70, Section 3].

**Definition 3.1** (piecewise constant extension I). For  $1 \leq i \leq M$ ,  $x \in B_i$ ,  $u \in \mathbb{U}$ , and  $A \in \mathcal{F}_\mathbb{X}$ , we define

$$\overline{c^{\pi,M}}(x) := c^{\pi,M}(y_i, u), \quad \overline{\mathcal{T}^{\pi,M}}(A | x, u) := \int_{B_i} \mathcal{T}(A | z, u) \pi_i(dz) / \pi_i(B_i),$$

and further  $\overline{\gamma_\beta^{\pi,M}}(x) := \gamma_\beta^{\pi,M}(y_i)$ ,  $\overline{\gamma_\infty^{\pi,M}}(x) := \gamma_\infty^{\pi,M}(y_i)$ .

As discussed in [70, Section 3],  $\overline{\gamma_\beta^{\pi,M}}$  and  $\overline{\gamma_\infty^{\pi,M}}$  are the optimal policies for the  $\beta$ -discounted and the average cost MDP  $(\mathbb{X}, \mathbb{U}, \overline{\mathcal{T}^{\pi,M}}, \overline{c^{\pi,M}})$ , respectively.

We note that the transition kernel under piecewise constant extension is, in general, not weakly continuous, so Assumption 2.1 does not apply to the approximate model. As a remedy, for the discounted problem, we adopt Assumption 2.1 of [65], which automatically holds for the piecewise constant model under Assumption 2.1 with a finite action set  $\mathbb{U}$ . For the average cost problem, in addition to Assumption 2.1 of [65], Assumption 3.1(4) of [65] also holds if the original kernel  $\mathcal{T}$  is minorized as in Definition 2.1. Therefore, we can invoke optimality equations for the piecewise constant model for both problems, which means that all our results hold when using the piecewise constant model as an approximation. We further define the quantization error:

$$\delta_M := \max_{1 \leq i \leq M} \sup_{x, x' \in B_i} d_\mathbb{X}(x, x'). \quad (3.2)$$

The following approximation error is noted in [69]:

**Lemma 3.1** (Lemma 3.3, [69]). *Under Assumption 1.1(a) and (b),*

$$\|c - \overline{c^{\pi,M}}\|_\infty \leq \|c\|_{Lip} \delta_M, \quad d_{\mathcal{W}_1}(\mathcal{T}, \overline{\mathcal{T}^{\pi,M}}) \leq \|\mathcal{T}\|_{Lip} \delta_M.$$

Combining Corollary 2.1 and Lemma 3.1, we immediately recover the following result for the  $\beta$ -discounted cost case, as stated in Theorem 6 of [48].

**Corollary 3.1** (Theorem 6, [48]). *Under Assumption 1.1,*

$$\left\| J_\beta(c, \mathcal{T}, \overline{\gamma_\beta^{\pi, M}}) - J_\beta^*(c, \mathcal{T}) \right\|_\infty \leq \frac{2\|c\|_{Lip}}{(1-\beta)^2(1-\beta\|\mathcal{T}\|_{Lip})} \delta_M.$$

Further, combining Corollary 2.2 and Lemma 3.1, we also recover the following result for the average cost case, as stated in Theorem 3.5 of [69].

**Corollary 3.2** (Theorem 3.5, [69]). *Suppose that  $\mathcal{T}$  satisfies the minorization condition with a probability measure  $\rho$  and a constant  $\epsilon > 0$  as in Definition 2.1. Under Assumption 2.4,*

$$\left\| J_\infty^*(c, \mathcal{T}, \overline{\gamma_\infty^{\pi, M}}) - J_\infty^*(c, \mathcal{T}) \right\|_\infty \leq \left(1 + \frac{2}{\epsilon}\right) \left(\frac{\|c\|_{Lip}}{1 - \|\mathcal{T}\|_{Lip}}\right) \delta_M.$$

*Proof.* Due to the construction in Definition 3.1,  $\overline{\mathcal{T}^{\pi, M}}$  also satisfies the minorization condition with the probability measure  $\rho$  and constant  $\epsilon$ . Consequently, the proof is complete by Corollary 2.2.  $\square$

Note that in dealing with the approximate MDP  $(\mathbb{X}, \mathbb{U}, \overline{\mathcal{T}^{\pi, M}}, \overline{c^{\pi, M}})$  above, we work with the  $\mathcal{W}_1$ -distance between  $\mathcal{T}$  and  $\overline{\mathcal{T}^{\pi, M}}$ , and use Corollary 2.1 and Corollary 2.2. As shown in the Section 4, this approach is not tight. On the other hand, in Lemma 3.1, the estimation errors,  $\|c - \overline{c^{\pi, M}}\|_\infty$  and  $d_{\mathcal{W}_1}(\mathcal{T}, \overline{\mathcal{T}^{\pi, M}})$ , are of the same order, so there is no benefit in applying tighter results in Theorem 2.7 and Theorem 2.8. However, as we shall see in the next subsection, those tighter results will be important in finite model learning.

In the next subsections, we adopt a different extension for the kernel for the purpose of technical analysis. Denote by  $\mathbb{Q}_M := \{y_i\}_{i=1}^M$ .

**Definition 3.2** (piecewise constant extension II). Let  $\mathcal{S} : \mathcal{F}_{\mathbb{Q}_M} \times \mathbb{Q}_M \times \mathbb{U} \rightarrow [0, 1]$  be a controlled transition kernel. Define its piecewise constant extension  $\tilde{\mathcal{S}} : \mathcal{F}_{\mathbb{X}} \times \mathbb{X} \times \mathbb{U} \rightarrow [0, 1]$  as follows: for  $i = 1, \dots, M$ ,  $x \in B_i$ ,  $u \in \mathbb{U}$ , and  $A \in \mathcal{F}_{\mathbb{X}}$ ,  $\tilde{\mathcal{S}}(A|x, u) := \sum_{j=1}^M \mathcal{S}(y_j|y_i, u) \mathbb{1}\{y_j \in A\}$ .

We can bound the approximation error in a manner similar to Lemma 3.1. Since the proof closely parallels that of Lemma 3.1, we omit it here.

**Lemma 3.2.** *Under Assumption 1.1(b),  $d_{\mathcal{W}_1}(\widetilde{\mathcal{T}^{\pi, M}}, \mathcal{T}) \leq (1 + \|\mathcal{T}\|_{Lip})\delta_M$ .*

**3.2. Simultaneous Finite Model Approximation and Finite Model Learning.** In this section, we focus on the case where the quantized model in (3.1) is unknown and needs to be learned from data. We consider two scenarios regarding data availability: (i) a single trajectory of a controlled Markov process and (ii) independent transitions where a simulation device is available for each initial state and action.

For either case, our goal is to show the robustness of the optimal policies derived from *simultaneous* finite model approximation and the learning of these approximate finite models obtained via empirical estimates. We provide explicit sample complexity bounds that relate performance loss to the number of samples.

As noted earlier, a related approach is presented in [37], which involves a different construction for scenarios where the model is known and satisfies an absolute continuity condition with respect to a reference measure. This construction aims to obtain a finite model by utilizing empirical data collected from the reference measure. Together with the known density, the empirical measure is used to construct a finite model.

**Algorithm 1** Sampling a Wasserstein Regular MDP along a Single Trajectory**Require:** Simulation Length  $N$  and ergodic exploration policy  $\gamma$ 

- 1: **for all**  $n = 0, \dots, N - 1$  **do**
- 2:   Observe  $(X_n, U_n, C_n, X_{n+1})$  that evolves according to the true kernel  $\mathcal{T}$  and policy  $\gamma$
- 3: **end for**
- 4: Set  $\hat{\mathcal{T}}_N(y_j|y_i, u) := \frac{\sum_{n=0}^{N-1} \mathbb{1}\{X_{n+1} \in B_j, X_n \in B_i, U_n = u\}}{\sum_{n=0}^{N-1} \mathbb{1}\{X_n \in B_i, U_n = u\}}$
- 5: Set  $\hat{c}_N(y_i, u) := \frac{\sum_{n=0}^{N-1} C_n \mathbb{1}\{X_n \in B_i, U_n = u\}}{\sum_{n=0}^{N-1} \mathbb{1}\{X_n \in B_i, U_n = u\}}$
- 6: **return**  $(\hat{\mathcal{T}}_N, \hat{c}_N)$

3.2.1. *Empirical model learning from a single trajectory.* Let  $\gamma \in \mathcal{P}(\mathbb{U})$  be a probability mass function on  $\mathbb{U}$  such that for  $u \in \mathbb{U}$ ,  $\gamma_u > 0$ . We view  $\gamma$  as a state-independent control policy. We observe a Markov chain  $(X_n, U_n, C_n), 0 \leq n \leq N$ , whose dynamics are determined by the kernel  $\mathcal{T}$  and policy  $\gamma$ , as shown in Algorithm 1. Specifically, for each  $x \in \mathbb{X}$ ,  $u, u' \in \mathbb{U}$ , and  $A \in \mathcal{F}_{\mathbb{X}}$ ,

$$\mathbb{P}(X_{n+1} \in A, U_{n+1} = u' | X_n = x, U_n = u) = \gamma_u \int_A \mathcal{T}(dx' | x, u), \quad C_n = c(X_n, U_n). \quad (3.3)$$

Define the following estimators for  $i, j \in [M]$  and  $u \in \mathbb{U}$ ,

$$\hat{\mathcal{T}}_N(y_j|y_i, u) := \frac{\sum_{n=0}^{N-1} \mathbb{1}\{X_{n+1} \in B_j, X_n \in B_i, U_n = u\}}{\sum_{n=0}^{N-1} \mathbb{1}\{X_n \in B_i, U_n = u\}}, \quad \hat{c}_N(y_i, u) := \frac{\sum_{n=0}^{N-1} C_n \mathbb{1}\{X_n \in B_i, U_n = u\}}{\sum_{n=0}^{N-1} \mathbb{1}\{X_n \in B_i, U_n = u\}},$$

where if the denominator is zero, we use the convention that  $\hat{c}_N(y_i, u) = 0$ , and that  $\hat{\mathcal{T}}_N(y_j|y_i, u)$  equal 1 if  $j = i$  and 0 if  $j \neq i$ .

Given the learned MDP  $(\hat{\mathcal{T}}_N, \hat{c}_N)$ , we can solve the finite state space MDP for the optimal  $\beta$ -discounted cost policy  $\hat{\gamma}_{N,\beta}$  and the optimal average cost policy  $\hat{\gamma}_{N,\infty}$ . We then extend them to the original state space as in Definition 3.1, and denote the extensions by  $\hat{\gamma}_{N,\beta}$  and  $\hat{\gamma}_{N,\infty}$ . Next, we show that the extended policies are robust under the true dynamic  $(c, \mathcal{T})$ .

Consider the augmented Markov chain  $\{Z_n := (X_n, U_n, X_{n+1}) : n \geq 0\}$  with the state space  $\mathbb{Z} := \mathbb{X} \times \mathbb{U} \times \mathbb{X}$  and the transition kernel  $\mathcal{P}^{\mathcal{T}, \gamma}$ , which does not depend on the partitions  $\{B_i : i \in [M]\}$ .

**Assumption 3.1. (a).** Assume that  $\{Z_n : n \geq 0\}$  is  $\psi$ -irreducible with a unique invariant distribution  $\tilde{\pi}$  on  $\mathbb{Z}$ , which by definition must satisfy:  $\tilde{\pi}(A, u, A') = \int_{\mathbb{X}} \left( \int_{A'} \mathcal{T}(dx' | x, u) \right) \gamma_u \pi(dx)$  for  $A, A' \in \mathcal{F}_{\mathbb{X}}$  and  $u \in \mathbb{U}$ . Assume that  $\pi(B_i) > 0$  for  $1 \leq i \leq M$ .

**(b).** There exist some absolute constants  $c_0, C_0$  such that for any measurable function  $f : \mathbb{X} \times \mathbb{U} \times \mathbb{X} \rightarrow [0, 1]$ ,  $\epsilon > 0$  and  $n \geq 1$ , we have

$$\mathbb{P} \left( \left| n^{-1} \sum_{t=0}^{n-1} f(X_t, U_t, X_{t+1}) - \int_{\mathbb{X}} \sum_{u \in \mathbb{U}} \int_{\mathbb{X}} f(x, u, x') \mathcal{T}(dx' | x, u) \gamma_u \pi(dx) \right| \geq \epsilon \right) \leq C_0 \exp(-c_0 n \epsilon^2).$$

*Remark 3.1.* Assumption 3.1(b) has been extensively studied in the literature, primarily through three closely interrelated approaches: (i) Spectral methods for both reversible and non-reversible Markov chains [71, 72], (ii) Concentration inequalities and martingale difference methods [73, Chapter 23] and in particular [73, Theorem 23.3.1 and Corollary 23.2.4], and (iii) coupling based methods via stochastic drift conditions (see e.g. [74, Theorem 12] via drift conditions [67] to a *small* set).

**Example 3.1.** As an explicit example, Assumption 3.1(b) holds if the Markov chain  $\{Z_n : n \geq 0\}$  has a spectral gap and the distribution of  $X_0$  is not too far away from  $\pi$ . Specifically, denote by  $\mathcal{L}_2^0(\tilde{\pi}) := \{f : \mathbb{Z} \rightarrow \mathbb{R} : \tilde{\pi}(f^2) < \infty, \tilde{\pi}(f) = 0\}$  the space of zero mean, square integrable functions on  $\mathbb{Z}$  with respect to  $\tilde{\pi}$ , where  $\tilde{\pi}(f) := \int_{\mathbb{Z}} f(z) \tilde{\pi}(z)$ . Let  $\lambda \in [0, 1]$  be the operator norm of  $\mathcal{P}^{\mathcal{T}, \gamma}$  acting on  $\mathcal{L}_2^0(\tilde{\pi})$ . Further, denote by  $\nu$  the distribution of  $X_0$ . Assume that  $\lambda < 1$ , that

$\nu$  is absolutely continuous with respect to  $\pi$  with the Radon–Nikodym derive  $d\nu/d\pi$ , and that  $C'_0 := \left( \int_{\mathbb{X}} \left( \frac{d\nu}{d\pi}(x) \right)^2 \pi(dx) \right)^{1/2} < \infty$ . Then by Corollary 3.11 in [72], Assumption 3.1(b) holds with  $c_0 = (1 - \lambda)/(1 + \lambda)$ ,  $C_0 = C'_0$ .

Note that the above assumption is only imposed on the Markov chain  $\{Z_n : n \geq 0\}$ , and not on the finite approximation. Further, we do not need to know  $\pi$  above (as the model is unknown); this is *just used for analysis purposes*.

Recall the definitions of  $c^{\pi, M}$  and  $\mathcal{T}^{\pi, M}$  in (3.1). Define

$$\kappa_{\pi, M} := \min \left\{ \gamma_u \times \int_{B_i} \pi(dx) : i \in [M], u \in \mathbb{U} \right\}, \quad \kappa_{\mathcal{T}, M} := \min \left\{ \mathcal{T}^{\pi, M}(y_j | y_i, u) : (i, j, u) \in \mathcal{I}_M \right\}, \quad (3.4)$$

where  $\mathcal{I}_M := \{i, j \in [M], u \in \mathbb{U} : \mathcal{T}^{\pi, M}(y_j | y_i, u) > 0\}$ . In particular,  $\kappa_{\mathcal{T}, M}$  is the smallest non-zero element in the matrix  $\mathcal{T}^{\pi, M}$ . Further, define the event

$$\mathcal{E}_{N, M} := \bigcup_{1 \leq i, j \leq M, u \in \mathbb{U}} \left\{ \hat{T}_N(y_j | y_i, u) < \frac{1}{2} \mathcal{T}^{\pi, M}(y_j | y_i, u) \right\}. \quad (3.5)$$

Now, we bound the difference between the estimated  $(\hat{T}_N, \hat{c}_N)$  and their limits  $(c^{\pi, M}, \mathcal{T}^{\pi, M})$ .

**Lemma 3.3.** *Suppose that  $c$  is non-negative with  $\|c\|_{\infty} < \infty$ , and that Assumption 3.1 holds. Then, there exists an absolute constant  $C > 0$  such that*

$$\mathbb{E} \left[ |\hat{c}_N - c^{\pi, M}|_{\infty} \right] \leq CK_1 \|c\|_{\infty} \frac{1}{\kappa_{\pi, M}} \sqrt{\frac{\log(M|\mathbb{U}|)}{N}} + 2C_0 \|c\|_{\infty} M |\mathbb{U}| \exp(-K_2 \kappa_{\pi, M}^2 N),$$

where we denote  $K_1 := \sqrt{(1 + 2C_0)/c_0}$  and  $K_2 := c_0/4$ . Further, for any bounded function  $g : \{y_j : j \in [M]\} \rightarrow [-L, L]$  with  $L > 0$ , there exists an absolute constant  $C > 0$  such that

$$\mathbb{E} \left[ d_g(\hat{T}_N, \mathcal{T}^{\pi, M}) \right] \leq CK_1 L \frac{1}{\kappa_{\pi, M}} \sqrt{\frac{\log(M|\mathbb{U}|)}{N}} + 4C_0 LM |\mathbb{U}| \exp(-K_2 \kappa_{\pi, M}^2 N).$$

Finally, the following upper bound holds:  $\mathbb{P}(\mathcal{E}_{N, M}) \leq 6C_0 M^2 |\mathbb{U}| \exp(-K_2 \kappa_{\pi, M}^2 \kappa_{\mathcal{T}, M}^2 N/16)$ .

*Proof.* The proof is presented in Appendix A.2.  $\square$

Next, we extend the cost function  $\hat{c}_N$  to the original state space as in Definition 3.1, and denote it as  $\overline{\hat{c}_N}$ . Further, we extend  $\hat{T}_N$  to  $\widetilde{\hat{T}_N}$  as in Definition 3.2. It is clear that  $\overline{\hat{\gamma}_{N, \beta}}$  and  $\overline{\hat{\gamma}_{N, \infty}}$ , which are extensions of  $\hat{\gamma}_{N, \beta}$  and  $\hat{\gamma}_{N, \infty}$  as in Definition 3.1, are the optimal discounted cost and average cost MDP  $(\mathbb{X}, \mathbb{U}, \widetilde{\hat{T}_N}, \overline{\hat{c}_N})$ . We compare the original and learned MDP, i.e.,  $(\mathcal{T}, c)$  and  $(\widetilde{\hat{T}_N}, \overline{\hat{c}_N})$ , through the finite model approximation  $(\widetilde{\mathcal{T}^{\pi, M}}, \overline{c^{\pi, M}})$ ; see (3.1). We start with the discounted problem. Recall  $\delta_M$  from (3.2).

**Theorem 3.1.** *Suppose Assumptions 1.1 and 3.1 hold. There exists a constant  $C > 0$ , depending only on  $C_0, c_0, \|c\|_{Lip}, \|\mathcal{T}\|_{Lip}, \beta, \|c\|_{\infty}$ , such that*

$$\mathbb{E} \left[ \|J_{\beta}(c, \mathcal{T}, \overline{\hat{\gamma}_{N, \beta}}) - J_{\beta}^*(c, \mathcal{T})\|_{\infty} \right] \leq C \left( \delta_M + \frac{1}{\kappa_{\pi, M}} \sqrt{\frac{\log(M|\mathbb{U}|)}{N}} \right) + CM |\mathbb{U}| \exp(-C^{-1} \kappa_{\pi, M}^2 N).$$

*Proof.* By Lemma 3.1 and the triangle inequality, we have  $\|\overline{\hat{c}_N} - c\|_{\infty} \leq \|c\|_{Lip} \delta_M + \|\hat{c}_N - c^{\pi, M}\|_{\infty}$ .



Clearly,  $\|J_\beta^*(c, \mathcal{T})\|_\infty \leq \|c\|_\infty / (1 - \beta)$ . By Lemma 2.3,  $J_\beta^*(c, \mathcal{T})$  is  $\|c\|_{\text{Lip}} / (1 - \beta \|\mathcal{T}\|_{\text{Lip}})$ -Lipschitz. Thus, by Lemma 3.2 and Lemma 1.1, due to the triangle inequality, we have

$$d_{J_\beta^*(c, \mathcal{T})}(\mathcal{T}, \tilde{\mathcal{T}}_N) \leq \frac{\|c\|_{\text{Lip}}(1 + \|\mathcal{T}\|_{\text{Lip}})}{1 - \beta \|\mathcal{T}\|_{\text{Lip}}} \delta_M + d_{J_\beta^*(c, \mathcal{T})}(\mathcal{T}^{\pi, M}, \hat{\mathcal{T}}_N).$$

The result then follows by combining Lemma 3.3 with the first inequality in Theorem 2.7.  $\square$

*Remark 3.2.* Note that in Theorem 3.1, the constant  $C$  is independent of the partition.

Finally, we focus on the average cost problem. Assume the reference kernel  $\mathcal{T}$  satisfies the minorization condition in Definition 2.1 with some probability measure  $\rho$  and scaling constant  $\epsilon$ . We first find a minorization measure  $\tau^M$  for the approximating kernel  $\widetilde{\mathcal{T}^{\pi, M}}$ , and then bound the Wasserstein-1 distance between  $\rho$  and  $\tau^M$ .

**Lemma 3.4.** *Assume there exist a constant  $\epsilon > 0$  and a probability measures  $\rho \in \mathcal{P}(\mathbb{X})$ , such that  $\forall x \in \mathbb{X}, u \in \mathbb{U}$ , and  $A \in \mathcal{F}_{\mathbb{X}}, \mathcal{T}(A|x, u) \geq \epsilon \rho(A)$ . Define the following probability measure  $\tau^M$ :*

$$\tau^M(A) := \sum_{j=1}^M \rho(B_j) \mathbb{1}\{y_j \in A\}, \quad \text{for any } A \in \mathcal{F}_{\mathbb{X}}.$$

*Then for any  $x \in \mathbb{X}, u \in \mathbb{U}$  and  $A \in \mathcal{F}_{\mathbb{X}}, \widetilde{\mathcal{T}^{\pi, M}}(A|x, u) \geq \epsilon \tau^M(A)$  and  $\mathcal{W}_1(\rho, \tau^M) \leq \delta_M$ .*

*Proof.* For any  $A \in \mathcal{F}_{\mathbb{X}}, u \in \mathbb{U}$  and  $x \in B_i$  with  $1 \leq i \leq M$ , by Definition 3.2,

$$\widetilde{\mathcal{T}^{\pi, M}}(A|x, u) = \sum_{j=1}^M \frac{1}{\pi_i(B)} \int_{B_i} \mathcal{T}(B_j|x, u) \pi_i(dx) \mathbb{1}\{y_j \in A\} \geq \sum_{j=1}^M \frac{1}{\pi_i(B)} \int_{B_i} \rho(B_j) \pi_i(dx) \mathbb{1}\{y_j \in A\},$$

which implies the first claim.

Further, for any  $f : \mathbb{X} \rightarrow \mathbb{R}$  with  $\|f\|_{\text{Lip}} \leq 1$ , by the triangle inequality,

$$\left| \int f(x) \rho(dx) - \int f(x) \tau^M(dx) \right| \leq \sum_{j=1}^M \left| \int_{B_j} f(x) \rho(dx) - f(y_j) \rho(B_j) \right| \leq \sum_{j=1}^M \int_{B_j} |f(x) - f(y_j)| \rho(dx),$$

which can be further bounded by  $\delta_M$ . The proof is complete.  $\square$

Recall the event  $\mathcal{E}_{N, M}$  defined in (3.5). By definition, on its complete  $\mathcal{E}_{N, M}^c$ , the approximation kernel  $\tilde{\mathcal{T}}_N$  is minorized by the probability measure  $\tau^M$  and constant  $\epsilon/2$ . Then we can apply the minorization approach in Theorem 2.8 on  $\mathcal{E}_{N, M}^c$ .

**Theorem 3.2.** *Suppose that Assumptions 2.4 and 3.1 hold, and that  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  satisfies the minorization condition with a probability measure  $\rho$  and a constant  $\epsilon > 0$ . There exists a constant  $C > 0$ , depending only on  $C_0, c_0, \|c\|_{\text{Lip}}, \|\mathcal{T}\|_{\text{Lip}}, \|c\|_\infty, \epsilon$ , such that*

$$\begin{aligned} \mathbb{E} [\|J_\infty(c, \mathcal{T}, \tilde{\gamma}_{N, \infty}) - J_\infty^*(c, \mathcal{T})\|_\infty] &\leq C \left( \delta_M + \frac{1}{\kappa_{\pi, M}} \sqrt{\frac{\log(M|\mathbb{U}|)}{N}} \right) + CM|\mathbb{U}| \exp(-C^{-1} \kappa_{\pi, M}^2 N) \\ &\quad + CM^2 |\mathbb{U}| \exp(-C^{-1} \kappa_{\pi, M}^2 \kappa_{\mathcal{T}, M}^2 N). \end{aligned}$$

*Proof.* Since  $c$  is non-negative and bounded, and by Lemma 3.3, we have

$$\mathbb{E} [\|J_\infty(c, \mathcal{T}, \tilde{\gamma}_{N, \infty}) - J_\infty^*(c, \mathcal{T})\|_\infty; \mathcal{E}_{N, M}^c] \leq 6C_0 \|c\|_\infty M^2 |\mathbb{U}| \exp(-K_2 \kappa_{\pi, M}^2 \kappa_{\mathcal{T}, M}^2 N / 16).$$

Now, we focus on the event  $\mathcal{E}_{N, M}$ , on which  $\tilde{\mathcal{T}}_N$  satisfies the minorization condition with  $\tau^M$  and  $\epsilon/2$ . Note that  $\mathcal{T}$  satisfies the minorization condition with  $\rho$  and  $\epsilon/2$ .

In the proof of Theorem 3.1, we have shown  $\|\hat{c}_N - c\|_\infty \leq \|c\|_{\text{Lip}} \delta_M + \|\hat{c}_N - c^{\pi, M}\|_\infty$ . By Lemma 2.3,  $J_\beta^*(c, \mathcal{T})$  is  $\|c\|_{\text{Lip}}/(1 - \beta \|\mathcal{T}\|_{\text{Lip}})$ -Lipschitz, which, due to Theorem 2.3, implies that  $h_{c, \mathcal{T}}^*$  is  $\|c\|_{\text{Lip}}/(1 - \|\mathcal{T}\|_{\text{Lip}})$ -Lipschitz. Thus, by Lemma 3.2 and Lemma 1.1, due to the triangle inequality, we have

$$d_{h_{c, \mathcal{T}}^*}(\mathcal{T}, \tilde{\mathcal{T}}_N) \leq \frac{\|c\|_{\text{Lip}}(1 + \|\mathcal{T}\|_{\text{Lip}})}{1 - \|\mathcal{T}\|_{\text{Lip}}} \delta_M + d_{h_{c, \mathcal{T}}^*}(\mathcal{T}^{\pi, M}, \hat{\mathcal{T}}_N).$$

Finally, by Theorem 2.3,  $\|h\|_\infty \leq \|c\|_\infty$ . Then the proof is complete by combining Lemma 3.3 and Lemma 3.4 with Theorem 2.8.  $\square$

A few remarks are in order. First, for a fixed partition—and thus a fixed  $M$ —the upper bounds in Theorem 3.1 and 3.2 achieve the optimal parametric rate, i.e.,  $O(N^{-1/2})$ .

Second, assume that for each  $x \in \mathbb{X}$ ,  $\|x\| \leq L$ . Recall that  $\delta_M$  is the quantization error in (3.2). By a volume argument, there exists a quantization scheme such that

$$\delta_M \leq (C/M)^{1/d}, \quad (3.6)$$

for some constant  $C$ , depending only on  $d$  and  $L$ . Under this assumption, we choose  $N$  so that the approximation error  $\delta_M$ , arising from finite model approximation, matches the statistical estimation error from model learning. Specifically, define

$$N_{\text{disc}} = C \frac{1}{\kappa_{\pi, M}^2} M^{2/d} \log(M|\mathbb{U}|), \quad N_{\text{ave}} = C \frac{1}{\kappa_{\pi, M}^2} M^{2/d} \log(M|\mathbb{U}|) + C \frac{1}{\kappa_{\pi, M}^2 \kappa_{\mathcal{T}, M}^2} \log(M|\mathbb{U}|).$$

If we let  $N = N_{\text{disc}}$  in the  $\beta$ -discounted cost case and  $N = N_{\text{ave}}$  in the average cost case, then the upper bounds in Theorem 3.1 and 3.2 on the overall performance loss, that include finite model approximation and model learning, are both of order  $M^{-1/d}$ .

Recall the definition of  $\kappa_{\pi, M}$  and  $\kappa_{\mathcal{T}, M}$  in (3.4). Assume that

$$\kappa_{\pi, M} \geq 1/(CM|\mathbb{U}|), \quad \kappa_{\mathcal{T}, M} \geq 1/(CM). \quad (3.7)$$

Then  $N_{\text{disc}}$  is of order  $M^{2+2/d}|\mathbb{U}|^2 \log(M|\mathbb{U}|)$ , while  $N_{\text{ave}}$  is of order  $M^4|\mathbb{U}|^2 \log(M|\mathbb{U}|)$ .

*Remark 3.3.* We also note that the empirically estimated model is closely related to Quantized Q-Learning, as discussed in [48, Theorem 8] and [49, Theorem 2.1]. Specifically, the Q-learning algorithm for the quantized model converges to the fixed-point equation corresponding to the approximate finite model learned in this section, with the exploration policy guiding the empirical estimation of the model. An important operational implication of this connection is that, in many applications, it may be more efficient to learn the model directly rather than running Q-learning.

**3.2.2. Empirical model learning from independently generated transition data.** Let  $\pi \in \mathcal{P}(\mathbb{X})$  be a given weighting measure such that  $\pi(B_i) > 0$  for  $i \in [M]$ . We denote by  $\hat{\pi}_i$  the restriction of  $\pi$  on the bin  $B_i$ , that is,  $\hat{\pi}_i(A) := \pi(A \cap B_i)/\pi(B_i)$  for  $i \in [M]$ . We consider the data collected under the scheme described in Algorithm 2. Specifically, for each bin  $i \in [M]$  and action  $u \in \mathbb{U}$ , we have  $N_0$  independent triplets of observations: for  $1 \leq k \leq N_0$ ,

$$(X_{k,i,u}, Y_{k,i,u}, c(X_{k,i,u}, u)), \quad \text{where } X_{k,i,u} \sim \hat{\pi}_i, \quad Y_{k,i,u} | X_{k,i,u} \sim \mathcal{T}(\cdot | X_{k,i,u}, u).$$

Note that in the above  $\hat{\pi}_i$  has its full measure on  $B_i$  and that the total number of triplets is  $N := M \times |\mathbb{U}| \times N_0$ . Given the data, the estimated controlled kernel and cost function  $(\hat{\mathcal{T}}_N, \hat{c}_N)$  on the quantized space  $\{y_i\}_{i=1}^M$  are defined as follows. For each  $1 \leq i \leq M$  and action  $u \in \mathbb{U}$ ,

$$\hat{c}_N(y_i, u) := \frac{1}{N_0} \sum_{k=1}^{N_0} c(X_{k,i,u}, u), \quad \hat{\mathcal{T}}_N(y_j | y_i, u) := \frac{1}{N_0} \sum_{k=1}^{N_0} \mathbb{1}\{Y_{k,i,u} \in B_j\}, \quad \text{for } 1 \leq j \leq M. \quad (3.8)$$

As in the previous section, we can solve the finite state space MDP  $(\{y_i\}_{i=1}^M, \mathbb{U}, \hat{\mathcal{T}}_N, \hat{c}_N)$  for the optimal  $\beta$ -discounted cost policy  $\hat{\gamma}_{N, \beta}$  and the optimal average cost policy  $\hat{\gamma}_{N, \infty}$ . We then extend

**Algorithm 2** Sampling a Wasserstein Regular MDP with Restart**Require:** number of repetitions  $N_0$  for each state and action pair

---

```

1: for all  $i = 1, \dots, M$  do
2:   for all  $u \in \mathbb{U}$  do
3:     for all  $k = 1, \dots, N_0$  do
4:       Sample i.i.d  $X_{k,i,u} \sim \hat{\pi}_i$  and  $Y_{k,i,u} \sim \mathcal{T}(\cdot | X_{k,i,u}, u)$ 
5:       Obtain cost  $c(X_{k,i,u}, u)$ 
6:     end for
7:     Set  $\hat{c}_N(y_i, u) := \frac{1}{N_0} \sum_{k=1}^{N_0} c(X_{k,i,u}, u)$ 
8:     Set  $\hat{\mathcal{T}}_N(y_j | y_i, u) := \frac{1}{N_0} \sum_{k=1}^{N_0} \mathbb{1}\{Y_{k,i,u} \in B_j\}$  for  $j \in [M]$ 
9:   end for
10: end for
11: return  $(\hat{\mathcal{T}}_N, \hat{c}_N)$ 

```

---

them to the original state space, denoted as  $\overline{\hat{\gamma}_{N,\beta}}$  and  $\overline{\hat{\gamma}_{N,\infty}}$ , which are the optimal policies for  $(\mathbb{X}, \mathbb{U}, \widetilde{\mathcal{T}}_N, \overline{\hat{c}_N})$ , where we recall the extension of kernel  $\widetilde{\mathcal{T}}_N$  in Definition 3.2.

We note two key differences compared to the single-trajectory setup in Subsection 3.2.1 when independently generated transition data is used. First, the weighting measure  $\pi$  is provided in this subsection and therefore does not need to be estimated. As a result, we can remove the terms involving  $\kappa_{\mathcal{T},M}$  in Theorems 3.1 and 3.2. Second, due to independence, Assumption (3.1) holds automatically; in particular, Assumption (3.1)(b) holds due to the Hoeffding inequality for i.i.d. random variables (see, e.g., Theorem 2.8 in [75]).

We now present upper bounds on the performance loss incurred when applying  $\overline{\hat{\gamma}_{N,\beta}}$  and  $\overline{\hat{\gamma}_{N,\infty}}$ , optimized under the estimated MDP  $(\mathbb{X}, \mathbb{U}, \widetilde{\mathcal{T}}_N, \overline{\hat{c}_N})$  to the true MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$ . The analysis follows similarly to that in Subsection 3.2.1, except that we replace Assumption (3.1)(b) with Hoeffding inequality for i.i.d. random variables (see, e.g., Theorem 2.8 in [75]), and that in bounding  $\mathbb{P}(\mathcal{E}_{N,M})$  (see (3.5)), we apply Bernstein's inequality. Therefore, we omit the detailed arguments for brevity.

**Theorem 3.3.** *Under Assumption 1.1, there exists a constant  $C > 0$ , depending only on  $\|c\|_{Lip}$ ,  $\|\mathcal{T}\|_{Lip}$ ,  $\|c\|_\infty$ ,  $\beta$ , such that*

$$\mathbb{E} \left[ \|J_\beta(c, \mathcal{T}, \overline{\hat{\gamma}_{N,\beta}}) - J_\beta^*(c, \mathcal{T})\|_\infty \right] \leq C \left( \delta_M + \sqrt{\frac{\log(M|\mathbb{U}|)}{N_0}} \right).$$

Suppose that Assumption 2.4 holds, and that  $(\mathbb{X}, \mathbb{U}, \mathcal{T}, c)$  satisfies the minorization condition with a probability measure  $\rho$  and a constant  $\epsilon > 0$ . Then, there exists a constant  $C > 0$ , depending only on  $\|c\|_{Lip}$ ,  $\|\mathcal{T}\|_{Lip}$ ,  $\|c\|_\infty$ ,  $\epsilon$ , such that

$$\mathbb{E} \left[ \|J_\infty(c, \mathcal{T}, \overline{\hat{\gamma}_{N,\infty}}) - J_\infty^*(c, \mathcal{T})\|_\infty \right] \leq C \left( \delta_M + \sqrt{\frac{\log(M|\mathbb{U}|)}{N_0}} \right) + CM^2 |\mathbb{U}| e^{-C^{-1} \kappa_{\mathcal{T},M} N_0},$$

where we recall that  $\kappa_{\mathcal{T},M}$  is defined in (3.4).

For a fixed partition—hence a fixed  $M$ —the upper bounds in Theorem 3.3 achieve the optimal  $O(N^{-1/2})$  parametric rate for finite model learning.

Further, recall that  $\delta_M$  is the quantization error in (3.2), and that the total number of observations is  $N = M|\mathbb{U}|N_0$ . Assume that (3.6) holds for  $\delta_M$  and that (3.7) holds for  $\kappa_{\mathcal{T},M}$ . As before, we choose  $N$  so that the order of finite model approximation error matches that of the model learning

error. Specifically, define

$$N'_{\text{disc}} = CM^{2/d+1}|\mathbb{U}|\log(M|\mathbb{U}|), \quad N'_{\text{ave}} = CM^{\max\{2, 2/d+1\}}|\mathbb{U}|\log(M|\mathbb{U}|).$$

If we let  $N = N'_{\text{disc}}$  in the  $\beta$ -discounted cost case and  $N = N'_{\text{ave}}$  in the average cost case, then the upper bounds in Theorem 3.3 are both of order  $M^{-1/d}$ . Thus, with i.i.d. transition data, the sample complexity is significantly improved compared to the single trajectory case. As discussed earlier, there are two main reasons: first, we do not need to estimate  $\pi$ , which is given; second, we use Bernstein's inequality to bound  $P(\mathcal{E}_{N,M})$ , instead of relying on the Hoeffding inequality in the proof of Lemma 3.3. Finally, we note that the sample complexity improves as the dimension  $d$  of  $\mathbb{X}$  increases. This is because, as  $d$  grows, the finite model approximation error becomes larger, which results in a less stringent requirement on the finite model learning error.

#### 4. APPLICATION TO ROBUSTNESS TO NOISE DISTRIBUTION MISSPECIFICATION AND EMPIRICAL NOISE ESTIMATION

In this section, we consider the disturbance approximation and the associated robustness properties as discussed in [37, 51–53]. Specifically, we consider the following stochastic dynamical system:

$$X_{t+1} = f(X_t, U_t, W_t) \text{ for } t \geq 0, \quad \text{where } \{W_t\}_{t=0}^{\infty} \text{ is i.i.d. with some distribution } \mu. \quad (4.1)$$

Here,  $\{W_t\}_{t=0}^{\infty}$  is referred to as the disturbance process. Consider a decision maker who knows  $f$  and the true cost function  $c$ , but has no knowledge of the distribution  $\mu$ . However, we assume  $\mu$  can be estimated, say, from realized samples  $\{w_t\}_{t=0}^n$ . The decision maker then computes an optimal policy using an estimated distribution  $\nu$ , which may depend on the observed samples. Our goal is to upper bound the robustness error due to this misspecification by the distance between  $\mu$  and  $\nu$ . We show that this estimation procedure can be viewed as a controlled kernel approximation, which allows us to leverage the results from the previous section. Additional details are given in Section 4.1.

**4.1. Robustness to Noise Distribution Approximations.** To reduce the problem to controlled kernel approximation, we first establish a relationship between the distance between disturbance distributions and the distance between their corresponding controlled kernels, along with their Lipschitz continuity.

Let  $\mu, \nu \in \mathcal{P}(\mathbb{W})$  be two probability measures. For  $x \in \mathbb{X}$ ,  $u \in \mathbb{U}$  and  $A \subset \mathcal{F}_{\mathbb{X}}$ , define two controlled kernels:

$$\mathcal{T}_{\mu}(A|x, u) := \mu(f_{x,u}^{-1}(A)), \quad \text{and} \quad \mathcal{T}_{\nu}(A|x, u) := \nu(f_{x,u}^{-1}(A))$$

where  $f_{x,u}^{-1}(A) := \{w \in \mathbb{W} : f(x, u, w) \in A\}$ . For any continuous and bounded function  $g \in C_b(\mathbb{X})$ , by definition,

$$d_g(\mathcal{T}_{\mu}, \mathcal{T}_{\nu}) = \sup_{(x,u) \in \mathbb{X} \times \mathbb{U}} \left| \int_{\mathbb{W}} g(f(x, u, w)) \mu(dw) - \int_{\mathbb{W}} g(f(x, u, w)) \nu(dw) \right|.$$

We start with the continuity and Lipschitz continuity property of the controlled kernels  $\mathcal{T}_{\mu}$  and  $\mathcal{T}_{\nu}$ . The following Lemma is stated for  $\mathcal{T}_{\mu}$ , but it also applies to  $\mathcal{T}_{\nu}$ .

We say that  $f$  is  $\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})}$ -Lipschitz continuous (resp. continuous) in  $(x, u)$  if for each  $w \in \mathbb{W}$ ,  $f(\cdot, \cdot, w) : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{X}$  is  $\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})}$ -Lipschitz continuous (resp. continuous). Further, we say  $f$  is  $\|f\|_{\text{Lip}(\mathbb{X})}$ -Lipschitz continuous in  $x$  if for each  $(u, w) \in \mathbb{U} \times \mathbb{W}$ ,  $f(\cdot, u, w) : \mathbb{X} \rightarrow \mathbb{X}$  is  $\|f\|_{\text{Lip}(\mathbb{X})}$ -Lipschitz continuous. Finally, we say  $f$  is  $\|f\|_{\text{Lip}(\mathbb{W})}$ -Lipschitz continuous in  $w$  if for each  $(x, u) \in \mathbb{X} \times \mathbb{U}$ ,  $f(x, u, \cdot)$  is  $\|f\|_{\text{Lip}(\mathbb{W})}$ -Lipschitz continuous.

**Lemma 4.1.** *i) If  $f$  is continuous function in  $(x, u)$ , then the transition kernel  $\mathcal{T}_{\mu}$  is weakly continuous on  $\mathbb{X} \times \mathbb{U}$ , that is, for any  $v \in C_b(\mathbb{X})$ ,  $\int_{\mathbb{X}} v(x') \mathcal{T}_{\mu}(dx'|x, u)$  is a continuous on  $\mathbb{X} \times \mathbb{U}$ .*

ii) Assume that  $f$  is  $\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})}$ -Lipschitz continuous in  $(x, u)$ . Let  $v : \mathbb{X} \rightarrow \mathbb{R}$  be a bounded, Lipschitz function with a constant  $\|v\|_{\text{Lip}}$ . Then for each  $x, y \in \mathbb{X}$  and  $u, u' \in \mathbb{U}$ , we have

$$\left| \int_{\mathbb{X}} v(x') \mathcal{T}_\mu(x'|x, u) - \int_{\mathbb{X}} v(x') \mathcal{T}_\mu(x'|y, u') \right| \leq \|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})} \|v\|_{\text{Lip}} (d_{\mathbb{X}}(x, y) + d_{\mathbb{U}}(u, u')).$$

iii) Assume that  $f$  is  $\|f\|_{\text{Lip}(\mathbb{X})}$ -Lipschitz continuous in  $x$ . Then for any  $u \in \mathbb{U}$ , and  $x, y \in \mathbb{X}$ ,

$$\mathcal{W}_1(\mathcal{T}_\mu(\cdot|x, u), \mathcal{T}_\mu(\cdot|y, u)) \leq \|f\|_{\text{Lip}(\mathbb{X})} d_{\mathbb{X}}(x, y).$$

*Proof.* Note that for any  $\nu \in C_b(\mathbb{X})$ , by definition,  $\int_{\mathbb{X}} v(x') \mathcal{T}_\mu(dx'|x, u) = \int_{\mathbb{W}} v(f(x, u, w)) \mu(dw)$ .

Under the assumption of claim (i), for each  $w \in \mathbb{W}$ ,  $v(f(\cdot, \cdot, w)) : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$  is a continuous function bounded by  $\|v\|_\infty$ . Then claim (i) follows from the bounded convergence theorem.

Under the assumption of claim (ii), for each  $w \in \mathbb{W}$ ,  $v(f(\cdot, \cdot, w)) : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$  is Lipschitz continuous with a constant  $\|v\|_{\text{Lip}} \|f\|_{\text{Lip}}$ . Thus, for each  $x, y \in \mathbb{X}$  and  $u, u' \in \mathbb{U}$ , we have

$$\begin{aligned} \left| \int_{\mathbb{X}} v(x') \mathcal{T}_\mu(x'|x, u) - \int_{\mathbb{X}} v(x') \mathcal{T}_\mu(x'|y, u') \right| &\leq \int_{\mathbb{W}} |v(f(x, u, w)) - v(f(y, u', w))| \mu(dw) \\ &\leq \|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})} \|v\|_{\text{Lip}} (d_{\mathbb{X}}(x, y) + d_{\mathbb{U}}(u, u')), \end{aligned}$$

which proves the claim (ii). Finally, we focus on claim (iii). By the dual formulation of  $\mathcal{W}_1$ ,

$$\begin{aligned} \mathcal{W}_1(\mathcal{T}(\cdot|x, u), \mathcal{T}(\cdot|y, u)) &= \sup_{\|g\|_{\text{Lip}} \leq 1} \left( \int g(f(x, u, w)) \mu(dw) - \int g(f(y, u, w)) \mu(dw) \right) \\ &\leq \sup_{\|g\|_{\text{Lip}} \leq 1} \left( \int \|f\|_{\text{Lip}(\mathbb{X})} d_{\mathbb{X}}(x, y) \mu(dw) \right) = \|f\|_{\text{Lip}(\mathbb{X})} d_{\mathbb{X}}(x, y), \end{aligned}$$

where the supremum is taken over all  $g : \mathbb{X} \rightarrow \mathbb{R}$  such that  $\|g\|_{\text{Lip}} \leq 1$ . The proof is complete.  $\square$

We denote by  $\gamma_{\nu, \beta}^*$  and  $\gamma_\nu^*$  the optimal policy for the MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}_\nu, c)$  under the  $\beta$ -discounted and average cost criterion respectively; that is, these are the optimal policies when the common distribution of the noise sequence  $\{W_t\}_{t \geq 0}$  in (4.1) is given by  $\nu$ . Our goal is to bound the performance loss incurred when these policies, optimized under  $\mathcal{T}_\nu$ , are applied to an MDP governed by the true dynamics  $\mathcal{T}_\mu$ . Define the following real-values functions: for each  $(x, u, w) \in \mathbb{X} \times \mathbb{U} \times \mathbb{W}$ ,

$$\begin{aligned} \tilde{f}_{\beta, \mu}(x, u, w) &:= (J_\beta^*(c, \mathcal{T}_\mu) \circ f)(x, u, w), & \tilde{f}_{\beta, \nu}(x, u, w) &:= (J_\beta^*(c, \mathcal{T}_\nu) \circ f)(x, u, w), \\ \tilde{h}_\mu(x, u, w) &:= (h_{c, \mathcal{T}_\mu}^* \circ f)(x, u, w), & \tilde{h}_\nu(x, u, w) &:= (h_{c, \mathcal{T}_\nu}^* \circ f)(x, u, w), \end{aligned} \quad (4.2)$$

where  $h_{c, \mathcal{T}_\mu}^*$  and  $h_{c, \mathcal{T}_\nu}^*$  are defined in Theorem 2.3.

**Theorem 4.1.** Suppose that Assumption 2.1 holds, and that  $f$  is continuous in  $(x, u)$ . Then

$$\|J_\beta(c, \mathcal{T}_\mu, \gamma_{\nu, \beta}^*) - J_\beta^*(c, \mathcal{T}_\mu)\|_\infty \leq \frac{2\beta}{1-\beta} d_{\tilde{f}_{\beta, \mu}}(\mu, \nu).$$

Further, let  $\mathbb{X}$  be compact and that for each  $x \in \mathbb{X}$ , there be a sequence  $\beta_n(x) \uparrow 1$  such that

$$(1 - \beta_n(x)) J_{\beta_n(x)}(c, \mathcal{T}_\mu, \gamma_\nu^*)(x) \rightarrow g_{c, \mathcal{T}_\mu}^{c, \mathcal{T}_\nu} = J_\infty(c, \mathcal{T}_\mu, \gamma_\nu^*)(x). \quad (4.3)$$

Furthermore, assume that  $c$  is  $\|c\|_{\text{Lip}}$ -Lipschitz in  $x$ , and that  $f$  is  $\|f\|_{\text{Lip}(\mathbb{X})}$ -Lipschitz continuous in  $x$  and  $\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})}$ -Lipschitz continuous in  $(x, u)$ . If  $\|f\|_{\text{Lip}(\mathbb{X})} < 1$ , then

$$\|J_\infty(c, \mathcal{T}_\mu, \gamma_\nu^*) - J_\infty^*(c, \mathcal{T}_\mu)\|_\infty \leq d_{\tilde{h}_\mu}(\mu, \nu) + d_{\tilde{h}_\nu}(\mu, \nu).$$

*Proof.* Due to Lemma 4.1(i), Assumption 2.1 holds for both claims. The first claim then follows immediately from the third inequality in Theorem 2.7.

Now, we focus on the second claim and apply Theorem 2.9. Specifically, we only need to verify Assumptions 2.2(b) and (c) for both MDPs,  $\mathcal{T}_\mu$  and  $\mathcal{T}_\nu$ . Due to Lemma 4.1(iii) and Lemma 2.3, for each  $\beta \in (0, 1)$ ,  $J_\beta^*(c, \mathcal{T})$  is Lipschitz with a constant  $L := \|c\|_{\text{Lip}}/(1 - \|f\|_{\text{Lip}(\mathbb{X})})$ , which implies that Assumption 2.2(b) holds for both MDPs. Finally, by Lemma 4.1(ii), Assumption 2.2(c) holds with  $\tilde{L} := \|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})}$  for both MDPs. The proof then is complete by Theorem 2.9.  $\square$

*Remark 4.1.* As discussed before, a sufficient condition for (4.3) is that Assumption 2.3 holds for the true MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}_\mu, c)$ .

Next, we present results that involve the  $\mathcal{W}_1$ -distance between  $\mu$  and  $\nu$ . We start with a lemma that bounds the  $\mathcal{W}_1$ -distance between controlled kernels  $T_\mu$  and  $T_\nu$ .

**Lemma 4.2.** *Assume that for each  $(x, u) \in \mathbb{X} \times \mathbb{U}$ ,  $f(x, u, \cdot)$  is  $\|f\|_{\text{Lip}(\mathbb{W})}$ -Lipschitz continuous on  $\mathbb{W}$ . Then  $d_{\mathcal{W}_1}(\mathcal{T}_\mu, \mathcal{T}_\nu) \leq \|f\|_{\text{Lip}(\mathbb{W})} \mathcal{W}_1(\mu, \nu)$ .*

*Proof.* By the dual formulation of Wasserstein-1 distance,

$$\sup_{x, u} \mathcal{W}_1(\mathcal{T}_\mu(\cdot|x, u), \mathcal{T}_\nu(\cdot|x, u)) = \sup_{x, u} \sup_{\|g\|_{\text{Lip}} \leq 1} \left| \int g(f(x, u, w)) \mu(dw) - \int g(f(x, u, w)) \nu(dw) \right|,$$

where the supremum is taken over all  $(x, u) \in \mathbb{X} \times \mathbb{U}$  and  $g : \mathbb{X} \rightarrow \mathbb{R}$  such that  $\|g\|_{\text{Lip}} \leq 1$ . Clearly, for each fixed  $x, u$ , the function  $g(f(x, u, \cdot)) : \mathbb{W} \rightarrow \mathbb{R}$  is  $\|f\|_{\text{Lip}(\mathbb{W})}$ -Lipschitz. Then the proof is complete due to the definition of Wasserstein-1 distance.  $\square$

**Theorem 4.2.** *Assume that  $\mathbb{U}$  is compact and that  $c : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$  is nonnegative, bounded, continuous, and  $\|c\|_{\text{Lip}}$ -Lipschitz continuous in  $x$ . Further, assume that  $f$  is continuous in  $(x, u)$  and that  $\|f\|_{\text{Lip}(\mathbb{X})}$ -Lipschitz continuous in  $x$  and  $\|f\|_{\text{Lip}(\mathbb{W})}$ -Lipschitz continuous in  $w$ .*

i) *If  $\beta\|f\|_{\text{Lip}(\mathbb{X})} < 1$ , then*

$$\|J_\beta(c, \mathcal{T}_\mu, \gamma_{\nu, \beta}^*) - J_\beta^*(c, \mathcal{T}_\mu)\|_\infty \leq \frac{2\beta\|c\|_{\text{Lip}}\|f\|_{\text{Lip}(\mathbb{W})}}{(1 - \beta)(1 - \beta\|f\|_{\text{Lip}(\mathbb{X})})} \mathcal{W}_1(\mu, \nu).$$

ii) *If  $\mathbb{X}$  is compact,  $\|f\|_{\text{Lip}(\mathbb{X})} < 1$  and condition (4.3) holds, then*

$$\|J_\infty(c, \mathcal{T}_\mu, \gamma_\nu^*) - J_\infty^*(c, \mathcal{T}_\mu)\|_\infty \leq \frac{2\|c\|_{\text{Lip}}\|f\|_{\text{Lip}(\mathbb{W})}}{1 - \|f\|_{\text{Lip}(\mathbb{X})}} \mathcal{W}_1(\mu, \nu).$$

*Proof.* Due to Lemma 4.1(i), Assumption 2.1 holds. Due to Lemma 4.1(iii) and Lemma 2.3, for each  $\beta \in (0, 1)$ ,  $J_\beta^*(c, \mathcal{T})$  is Lipschitz with a constant  $\|c\|_{\text{Lip}}/(1 - \|f\|_{\text{Lip}(\mathbb{X})})$ . Then, the first follows immediately from Lemma 4.2, Lemma 1.1 and Theorem 2.7. The second claim follows from Theorem 2.9, similarly to the proof of Theorem 4.1.  $\square$

*Remark 4.2.* Note that in Theorem 4.2, the upper bounds involve the  $\mathcal{W}_1$ -distance between  $\mu$  and  $\nu$ . In contrast, Theorem 4.1 requires bounding only the difference between  $\mu$  and  $\nu$  integrated against a *single* function. For this reason, as we shall see in Subsection 4.2, we can achieve the *optimal* statistical rates under the conditions of Theorem 4.1, when we estimate  $\mu$  by empirical distributions, which is *not* the case if we use Theorem 4.2. The following two subsections will highlight this distinction.

Further, Theorem 4.1 requires that  $f$  is  $\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})}$ -Lipschitz in  $(x, u)$ , which is not required for Theorem 4.2. On the other hand, Theorem 4.2 requires that  $f$  is  $\|f\|_{\text{Lip}(\mathbb{W})}$ -Lipschitz in  $w$ , which is not required for Theorem 4.1.



## 4.2. General Sample Complexity Bounds.

**4.2.1. Sample Complexity Bounds on Robustness to Empirical Noise Distribution Estimation under Regularity in Noise.** Recall the stochastic dynamic system in (4.1), where the function  $f$  is known but the disturbance distribution  $\mu$  is *unknown*. Denote by  $\mathcal{T}_\mu$  the associated controlled transition kernel.

Assume that we can observe an i.i.d sequence  $\{W_t\}_{t=0}^\infty$  such that for any  $t \geq 0$ ,  $W_t \sim \mu$ , and that we use the empirical measures  $\{\mu_n : n \geq 0\}$  to estimate  $\mu$ :

$$\mu_n(\cdot) := \frac{1}{n+1} \sum_{t=0}^n \delta_{W_t}(\cdot), \text{ for } n \geq 0, \quad (4.4)$$

For  $n \geq 0$ , given the estimate  $\mu_n$ , we consider the following stochastic dynamic system:

$$\tilde{X}_{t+1} = f(\tilde{X}_t, \tilde{U}_t, \tilde{W}_t) \text{ for } t \geq 0, \quad \text{where } \{\tilde{W}_t\}_{t=0}^\infty \text{ is i.i.d. with distribution } \mu_n,$$

and denote by  $\mathcal{T}_{\mu_n}$  the associated controlled transition kernel. Since the cost function  $c$  is assumed known, we can solve the  $\beta$ -discounted cost and the average cost MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}_{\mu_n}, c)$ , and denote the solutions by  $\gamma_{\mu_n, \beta}^*$  and  $\gamma_{\mu_n}^*$  respectively.

By applying the result from [76] concerning the Wasserstein-1 distance between  $\mu$  and its empirical counterpart  $\mu_n$ , Theorem 4.2 immediately yields the following result. For  $q \geq 1$ , denote the  $q$ -th moment of  $\mu$  by  $M_q(\mu) := \int_{\mathbb{R}^d} |x|^q \mu(dx)$ .

**Theorem 4.3.** *Assume  $f$  in (4.1) is  $\|f\|_{Lip(\mathbb{X})}$ -Lipschitz continuous on  $\mathbb{X}$ , continuous on  $\mathbb{U}$ , and  $\|f\|_{Lip(\mathbb{W})}$ -Lipschitz continuous on  $\mathbb{W}$ . Let  $\mathbb{U}$  be compact and  $c : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$  be a known cost function such that  $c$  is  $\|c\|_{Lip}$ -Lipschitz continuous as a function of  $\mathbb{X}$ , and continuous as a function on  $\mathbb{U}$ . Let  $\beta \in (0, 1)$  be given such that  $\beta\|f\|_{Lip(\mathbb{X})} < 1$ .*

*i) If there exists  $q > 1$  such that  $M_q(\mu) < \infty$ , then there exists a positive constant  $C$  depending on  $d, q, \mu, \beta, \|c\|_{Lip}, \|f\|_{Lip(\mathbb{X})}$ , and  $\|f\|_{Lip(\mathbb{W})}$ , such that for all  $n \geq 1$ ,*

$$\mathbb{E} \left[ \|J_\beta(c, \mathcal{T}_\mu, \gamma_{\mu_n, \beta}^*) - J_\beta^*(c, \mathcal{T}_\mu)\|_\infty \right] \leq C \begin{cases} n^{-1/2} + n^{-(q-1)/q}, & \text{if } d < 2, q \neq 2, \\ n^{-1/2} \log(1+n) + n^{-(q-1)/q}, & \text{if } d = 2, q \neq 2, \\ n^{-1/d} + n^{-(q-1)/q}, & \text{if } d > 2, q \neq d/(d-1). \end{cases}$$

*ii) Additionally, assume  $\|f\|_{Lip(\mathbb{X})} < 1$ . Then the above result holds under the average cost criterion if we replace  $J_\beta, J_\beta^*$ , and  $\gamma_{\mu_n, \beta}^*$  by  $J_\infty, J_\infty^*$ , and  $\gamma_{\mu_n}^*$ , respectively; in both cases,  $C$  no longer depends on  $\beta$ .*

Assume that  $M_q(\mu) < \infty$  for  $q \geq \max\{2, d\}$ . Then the convergence rate is  $n^{-1/2}$ ,  $n^{-1/2} \log(n)$ , and  $n^{-1/d}$  for  $d = 1, 2$ , and  $d > 2$ , respectively. For  $d = 1$ , the rate is parametric and thus optimal. For  $d \geq 2$ , the rate is generally suboptimal. However, we emphasize that the above theorem does not assume  $f$  to be jointly Lipschitz in both  $x$  and  $u$ . Under this additional regularity condition, we derive improved rates in the next subsection.

**4.2.2. Improved Sample Complexity Bounds on Robustness to Empirical Noise Distribution Estimation under Regularity in State and Action.** As earlier in this section, suppose that we do not know the probability measure  $\mu \in \mathcal{P}(\mathbb{W})$ , but we compute empirical estimates as in (4.4).

In view of Theorem 4.1, it suffices to bound  $d_{\tilde{f}_{\beta, \mu}}(\mu, \mu_n)$ ,  $d_{\tilde{h}_\mu}(\mu, \mu_n)$  and  $d_{\tilde{h}_{\mu_n}}(\mu, \mu_n)$ , for which we can use tools from empirical processes [77]. The key step is to show that the involved function  $\tilde{f}_{\beta, \mu}$ ,  $\tilde{h}_\mu$  and  $\tilde{h}_{\mu_n}$  are Lipschitz continuous. Denote by  $\|c\|_\infty := \sup_{x, u} |c(x, u)|$ , and by  $\text{diam}(\mathbb{X})$  and  $\text{diam}(\mathbb{U})$  the diameters of  $\mathbb{X}$  and  $\mathbb{U}$  respectively.

**Theorem 4.4.** Assume that both  $\mathbb{X} \subset \mathbb{R}^{d_1}$  and  $\mathbb{U} \subset \mathbb{R}^{d_2}$  are compact, and that  $c : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$  is nonnegative, bounded, continuous, and  $\|c\|_{\text{Lip}}$ -Lipschitz in  $x$ . Further, assume that  $f$  is  $\|f\|_{\text{Lip}(\mathbb{X})}$ -Lipschitz continuous in  $x$  and  $\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})}$ -Lipschitz continuous in  $(x, u)$ . If  $\|f\|_{\text{Lip}(\mathbb{X})} < \beta^{-1}$ , then there exists a constant  $C > 0$ , depending only on  $d_1, d_2, \text{diam}(\mathbb{X})$  and  $\text{diam}(\mathbb{U})$ , such that

$$\mathbb{E} \left[ \|J_\beta(c, \mathcal{T}_\mu, \gamma_{\mu_n, \beta}^*) - J_\beta^*(c, \mathcal{T}_\mu)\|_\infty \right] \leq C \frac{\beta}{1 - \beta} \left( \frac{\|c\|_\infty}{1 - \beta} + \frac{\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})} \|c\|_{\text{Lip}}}{1 - \beta \|f\|_{\text{Lip}(\mathbb{X})}} \right) n^{-1/2}.$$

If, in addition,  $\|f\|_{\text{Lip}(\mathbb{X})} < 1$  and condition (4.3) holds, then

$$\mathbb{E} \left[ \|J_\infty(c, \mathcal{T}_\mu, \gamma_{\mu_n}^*) - J_\infty^*(c, \mathcal{T}_\mu)\|_\infty \right] \leq C \left( \|c\|_\infty + \frac{\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})} \|c\|_{\text{Lip}}}{1 - \|f\|_{\text{Lip}(\mathbb{X})}} \right) n^{-1/2}.$$

*Proof.* Define the following constants

$$m_\beta := \frac{\|c\|_{\text{Lip}}}{1 - \beta \|f\|_{\text{Lip}(\mathbb{X})}}, \quad M_\beta := m_\beta \|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})}, \quad m := \frac{\|c\|_{\text{Lip}}}{1 - \|f\|_{\text{Lip}(\mathbb{X})}}, \quad M := m \|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})}. \quad (4.5)$$

For the first claim, due to Theorem 4.1, it suffices to bound

$$\mathbb{E} \left[ d_{\tilde{f}_{\beta, \mu}}(\mu, \mu_n) \right] = \sup_{(x, u) \in \mathbb{X} \times \mathbb{U}} \left| \frac{1}{n+1} \sum_{t=0}^n \left( \tilde{f}_{\beta, \mu}(x, u, W_t) - \int_{\mathbb{W}} \tilde{f}_{\beta, \mu}(x, u, w) \mu(dw) \right) \right|.$$

Recall the definition of  $\tilde{f}_{\beta, \mu}$  in (4.2). Clearly,  $\tilde{f}_{\beta, \mu}(x, u, w) \leq \|c\|_\infty / (1 - \beta) := B_\beta$  for each  $(x, u, w) \in \mathbb{X} \times \mathbb{U} \times \mathbb{W}$ . Further, due to Lemma 4.1(iii) and Lemma 2.3, for each  $\beta \in (0, 1)$ ,  $J_\beta^*(c, \mathcal{T}_\mu)$  is Lipschitz with a constant  $m_\beta$ . Thus, for each  $\beta \in (0, 1)$ ,  $\tilde{f}_{\beta, \mu}$  is  $M_\beta$ -Lipschitz continuous in  $(x, u)$ . By applying Lemma A.1 in Appendix A.1 with constants  $B_\beta$  and  $M_\beta$  (see also Remark A.1), we have

$$\mathbb{E} \left[ d_{\tilde{f}_{\beta, \mu}}(\mu, \mu_n) \right] \leq C \left( \frac{\|c\|_\infty}{1 - \beta} + \frac{\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})} \|c\|_{\text{Lip}}}{1 - \beta \|f\|_{\text{Lip}(\mathbb{X})}} \right) n^{-1/2}. \quad (4.6)$$

which complete the proof for the  $\beta$ -discounted case in view of Theorem 4.1.

Next, we focus on the average cost case. Due to Theorem 4.1, it suffices to bound  $d_{\tilde{h}_\mu}(\mu, \mu_n)$  and  $d_{\tilde{h}_{\mu_n}}(\mu, \mu_n)$ . Recall the definition of  $\tilde{h}_\mu$  and  $\tilde{h}_\nu$  (with  $\nu = \mu_n$ ) in (4.2). Due to Theorem 2.3, in particular the approximation by vanishing discounted factor approach, we have  $\|h_{c, \mathcal{T}_\mu}^*\|_\infty \leq \|c\|_\infty$  and  $\|h_{c, \mathcal{T}_{\mu_n}}^*\|_\infty \leq \|c\|_\infty$ , which implies  $\|\tilde{h}_\mu\|_\infty \leq \|c\|_\infty$  and  $\|\tilde{h}_{\mu_n}\|_\infty \leq \|c\|_\infty$ .

Due to Lemma 4.1(iii) and Lemma 2.3, for each  $\beta \in (0, 1)$ , we have  $\|J_\beta^*(c, \mathcal{T}_\mu)\|_{\text{Lip}} \leq m$  and  $\|J_\beta^*(c, \mathcal{T}_{\mu_n})\|_{\text{Lip}} \leq m$ . Then by Theorem 2.3, we have  $\|h_{c, \mathcal{T}_\mu}^*\|_{\text{Lip}} \leq m$  and  $\|h_{c, \mathcal{T}_{\mu_n}}^*\|_{\text{Lip}} \leq m$ , which implies that both  $\tilde{h}_\mu$  and  $\tilde{h}_{\mu_n}$  are  $M$ -Lipschitz. Then applying Lemma A.1 in Appendix A.1 with constants  $\|c\|_\infty$  and  $M$ , and due to Theorem 4.1, the proof is complete for the average cost case.  $\square$

In Theorem 4.4, we achieve the parametric statistic rate  $O(n^{-1/2})$ , which in general cannot be improved. In [12], related results are derived for the  $\beta$ -discounted criterion under similar conditions. Specifically, in [12], it is assumed that both  $\mathbb{X}$  and  $\mathbb{U}$  are compact, and that  $J_\beta^*(c, \mathcal{T}_\mu) \circ f$  is assumed to be  $L$ -Lipschitz in  $(x, u)$  (part (3) of their Assumption 2). It is not a priori clear under what conditions the composition  $J_\beta^*(c, \mathcal{T}_\mu) \circ f$  is Lipschitz continuous in  $(x, u)$ . In this work, we provide concrete sufficient conditions—specifically, that  $\|f\|_{\text{Lip}} < \beta$ —under which the Lipschitz constant  $L$  can be chosen as  $\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})} \|c\|_{\text{Lip}} / (1 - \beta \|f\|_{\text{Lip}(\mathbb{X})})$ ; see the proof above.

More importantly, to the best of our knowledge, the sample complexity results under the average cost criterion are novel. This setting is more challenging than the discounted cost case. Our analysis proceeds via an approximation approach based on a vanishing discount factor.

**4.3. Case where Model and Noise are Learned Simultaneously.** In this subsection, we consider the following additive dynamic system, which is a special case of (4.1):

$$X_{t+1} = f(X_t, U_t, W_t) = r(X_t, U_t) + W_t, \quad \text{for } t \geq 0.$$

where  $\{W_t\}_{t=0}^\infty$  are i.i.d. with some distribution  $\mu$  on  $\mathbb{W} \subset \mathbb{R}^{d_1}$ . In this subsection, we assume that  $\mathbb{X} \subset \mathbb{R}^{d_1}$  and  $\mathbb{U} \subset \mathbb{R}^{d_2}$  compact, and that  $r : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}^{d_1}$  is continuous.

In contrast to the previous subsection, we now consider the case in which both the function  $r(\cdot)$  and the distribution  $\mu$  are *unknown* and must be learned simultaneously from data. Specifically, we have access to the following data  $(\tilde{Y}_i, \tilde{X}_i, \tilde{U}_i)$  for  $1 \leq i \leq n$ , where

$$\tilde{Y}_i = r(\tilde{X}_i, \tilde{U}_i) + \tilde{W}_i, \quad \text{with } \{\tilde{W}_i\}_{i=0}^\infty \text{ are i.i.d. with distribution.} \quad (4.7)$$

Note that the variables  $\{\tilde{W}_i\}_{1 \leq i \leq n}$  are *unobserved*, and thus the empirical measure in (4.4) is not available. Further, the pair  $\{(\tilde{X}_i, \tilde{U}_i)\}_{1 \leq i \leq n}$  may not originate from a single trajectory of a controlled MDP—for instance, they could be generated independently. Based on this data, we denote by  $\tilde{r}_n(\cdot)$  an estimator of the function  $r(\cdot)$ , and define the following estimator for the distribution  $\mu$ :

$$\tilde{\mu}_n(\cdot) := \frac{1}{n+1} \sum_{i=0}^n \delta_{\hat{W}_i}(\cdot), \quad \text{for } n \geq 0, \quad \text{where } \hat{W}_i := Y_i - \tilde{r}_n(\tilde{X}_i, \tilde{U}_i), \quad (4.8)$$

where we recall that  $\delta_{\hat{W}_i}(\cdot)$  is the Dirac measure at  $\hat{W}_i$ . Concrete examples of the estimator  $\tilde{r}_n$  will be presented later. For  $x \in \mathbb{X}$ ,  $u \in \mathbb{U}$  and  $A \subset \mathcal{F}_{\mathbb{X}}$ , define the following controlled kernels:

$$\mathcal{T}_{r,\mu}(A|x, u) := \mu(A - r(x, u)), \quad \text{and} \quad \mathcal{T}_{\tilde{r}_n, \tilde{\mu}_n}(A|x, u) := \tilde{\mu}_n(A - \tilde{r}_n(x, u)).$$

Thus,  $(\mathbb{X}, \mathbb{U}, \mathcal{T}_{r,\mu}, c)$  is the reference MDP, while  $(\mathbb{X}, \mathbb{U}, \mathcal{T}_{\tilde{r}_n, \tilde{\mu}_n}, c)$  is an approximation. We denote by  $\gamma_{n,\beta}^*$  and  $\gamma_n^*$  the optimal policy for the approximate MDP  $(\mathbb{X}, \mathbb{U}, \mathcal{T}_{\tilde{r}_n, \tilde{\mu}_n}, c)$  under the  $\beta$ -discounted and average cost criterion respectively. Our goal is to bound the performance loss incurred when they are applied under the true dynamics  $\mathcal{T}_{r,\mu}$ . We note that the asymptotic convergence for problems of this type has been considered [24, Section 1.3.2(iv)]. Here, we obtain explicit and quantitative bounds.

**Theorem 4.5.** *Assume that  $c : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}$  is nonnegative, bounded, continuous, and  $\|c\|_{Lip}$ -Lipschitz in  $x$ . Further, assume that  $r$  is  $\|r\|_{Lip(\mathbb{X})}$ -Lipschitz continuous in  $x$  and  $\|r\|_{Lip(\mathbb{X} \times \mathbb{U})}$ -Lipschitz continuous in  $(x, u)$ . If  $\|r\|_{Lip(\mathbb{X})} < \beta^{-1}$ , then there exists a constant  $C > 0$ , that only depends on  $d_1, d_2$ ,  $\text{diam}(\mathbb{X})$  and  $\text{diam}(\mathbb{U})$ , such that  $\mathbb{E} \left[ \left\| J_\beta(c, \mathcal{T}_{r,\mu}, \gamma_{n,\beta}^*) - J_\beta^*(c, \mathcal{T}_{r,\mu}) \right\|_\infty \right]$  is upper bounded by*

$$C \frac{\beta}{1-\beta} \left[ \left( \frac{\|c\|_\infty}{1-\beta} + \frac{\|r\|_{Lip(\mathbb{X} \times \mathbb{U})} \|c\|_{Lip}}{1-\beta \|r\|_{Lip(\mathbb{X})}} \right) n^{-1/2} + \frac{\|c\|_{Lip}}{1-\beta \|f\|_{Lip(\mathbb{X})}} \mathbb{E} [\|\tilde{r}_n - r\|_\infty] \right].$$

*If, in addition,  $\|r\|_{Lip(\mathbb{X})} < 1$  and condition (4.3) holds for  $\mathcal{T}_{r,\mu}$  and  $\gamma_n^*$ , then we have the following upper bound for  $\mathbb{E} [\|J_\infty(c, \mathcal{T}_{r,\mu}, \gamma_n^*) - J_\infty^*(c, \mathcal{T}_{r,\mu})\|_\infty]$ :*

$$C \left( \|c\|_\infty + \frac{\|r\|_{Lip(\mathbb{X} \times \mathbb{U})} \|c\|_{Lip}}{1 - \|r\|_{Lip(\mathbb{X})}} \right) n^{-1/2} + C \frac{\|c\|_{Lip}}{1 - \|f\|_{Lip(\mathbb{X})}} \mathbb{E} [\|\tilde{r}_n - r\|_\infty].$$

*Proof.* We first focus on the  $\beta$ -discounted case, and denote by  $J(\cdot) := J_\beta^*(c, \mathcal{T}_{r,\mu})(\cdot)$  and  $\Delta_\beta := \left\| J_\beta(c, \mathcal{T}_{r,\mu}, \gamma_{n,\beta}^*) - J_\beta^*(c, \mathcal{T}_{r,\mu}) \right\|_\infty$ . Further, let  $\mu_n := (n+1)^{-1} \sum_{i=0}^n \delta_{\tilde{W}_i}$  be the empirical measure.

By a similar argument as in Theorem 4.1, due to Theorem 2.7, we have

$$\Delta_\beta \leq \frac{\beta}{1-\beta} \sup_{x,u} \left| \int_{\mathbb{W}} J(r(x, u) + w) \mu(dw) - \int_{\mathbb{W}} J(\tilde{r}_n(x, u) + w) \tilde{\mu}_n(dw) \right| \leq I_n + II_n + III_n,$$

where we define

$$I_n := d_J(\mu, \mu_n), \quad II_n := \sup_{x,u} \left| \int_{\mathbb{W}} J(r(x,u) + w) \mu_n(dw) - \int_{\mathbb{W}} J(\tilde{r}_n(x,u) + w) \mu_n(dw) \right|,$$

$$III_n := \sup_{x,u} \left| \int_{\mathbb{W}} J(\tilde{r}_n(x,u) + w) \mu_n(dw) - \int_{\mathbb{W}} J(\tilde{r}_n(x,u) + w) \tilde{\mu}_n(dw) \right|.$$

We have bounded  $\mathbb{E}[I_n]$  in (4.6), where  $\|f\|_{\text{Lip}(\mathbb{X})} = \|r\|_{\text{Lip}(\mathbb{X})}$  and  $\|f\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})} = \|r\|_{\text{Lip}(\mathbb{X} \times \mathbb{U})}$ . Further, in the proof of Theorem 4.4, we have shown that  $J(\cdot)$  is  $m_\beta$ -Lipschitz, where  $m_\beta$  is defined in (4.5); thus,  $II_n \leq m_\beta \|\hat{r}_n - r\|_\infty$ . Finally, by definition,

$$III_n = \sup_{x,u} \left| \frac{1}{n+1} \sum_{i=0}^n \left( J(\tilde{r}_n(x,u) + \tilde{W}_i) - J(\tilde{r}_n(x,u) + \hat{W}_i) \right) \right|,$$

where  $\hat{W}_i$  is defined in (4.8). By definition,  $|\tilde{W}_i - \hat{W}_i| \leq \|\hat{r}_n - r\|_\infty$ . As a result,  $III_n \leq m_\beta \|\hat{r}_n - r\|_\infty$ . The proof for the average cost case is similar and thus omitted.  $\square$

Note that compared to Theorem 4.4, we have an extra term involving  $\mathbb{E}[\|\tilde{r}_n - r\|_\infty]$ . Next, we provide two examples where this sup-norm estimation error can be controlled.

**Example 4.1.** Assume that  $\mathbb{X} \subset \mathbb{R}$  and  $\mathbb{U} \subset \mathbb{R}^{d_2}$  are compact, and that  $r(x, u) = \alpha_0 x + \theta'_0 u$  for  $x \in \mathbb{X}, u \in \mathbb{U}$ , where  $\alpha_0 \in \mathbb{R}$  and  $\theta_0 \in \mathbb{R}^{d_2}$  are both unknown. Assume that  $\{(\tilde{X}_i, \tilde{U}_i) : 0 \leq i \leq n\}$  are i.i.d., independent from  $\{\tilde{W}_i : 0 \leq i \leq n\}$ . We can estimate the coefficients by the least squares method:  $(\tilde{\alpha}_n, \tilde{\theta}_n) := \text{argmin}_{(\alpha, \theta) \in \mathbb{R}^{1+d_2}} \sum_{i=0}^n \left( \tilde{Y}_i - \alpha \tilde{X}_i - \theta' \tilde{U}_i \right)^2$ . It is well known that if the covariance matrix of  $(1, \tilde{X}_0, \tilde{U}_0)$  is non-singular, then  $\mathbb{E} \left[ |\tilde{\alpha}_n - \alpha_0| + \|\tilde{\theta}_n - \theta_0\|_2 \right] \leq Cn^{-1/2}$  for some constant  $C$ . Since  $\mathbb{X}$  and  $\mathbb{U}$  are compact, it implies that  $\mathbb{E}[\|\tilde{r}_n - r\|_\infty] \leq Cn^{-1/2}$ , where  $\tilde{r}_n(x, u) := \tilde{\alpha}_n x + \tilde{\theta}'_n u$  for  $x \in \mathbb{X}, u \in \mathbb{U}$ . Thus, overall, the upper bounds in Theorem 4.5 decay at the parametric rate  $n^{-1/2}$ .

**Example 4.2.** Assume that  $\mathbb{X} = [0, 1]$  and  $\mathbb{U}$  is finite. Let  $m \geq 1$  be an integer. Suppose for each  $u \in \mathbb{U}$  and  $0 \leq j \leq m$ , we observe  $\tilde{Y}_i = r(j/m, u) + \tilde{W}_i$ . Thus, the dataset is  $\{(\tilde{Y}_{u,i}, j/m, u) : 0 \leq j \leq m, u \in \mathbb{U}\}$ , which has a sample size of  $n+1 := (m+1) \times |\mathbb{U}|$ . Assume that there exist some constants  $\beta, L > 0$  such that for each  $u \in \mathbb{U}$ ,  $r(\cdot, u)$  is in Hölder class with parameter  $\beta, L$ , that is,  $r(\cdot, u)$  is  $\ell = \lfloor \beta \rfloor$  times differentiable and  $|r^{(\ell)}(x, u) - r^{(\ell)}(y, u)| \leq L|x - y|^{\beta - \ell}$  for  $x, y \in \mathbb{X}$ , where  $\lfloor \beta \rfloor$  denotes the greatest integer strictly less than  $\beta$  and  $r^{(\ell)}(x, u)$  is the  $\ell$ -derivative of  $r(\cdot, u)$  for a fixed  $u$ . Further, assume  $\tilde{W}_1$  is sub-Gaussian, in the sense that for some constant  $a, b > 0$ ,  $\mathbb{E}[\exp(a(\tilde{W}_1 - \mathbb{E}[\tilde{W}_1])^2)] \leq b$ . For each  $u \in \mathbb{U}$ , let  $\tilde{r}_n(\cdot, u)$  be a local polynomial estimator of order  $\ell$  for  $r(\cdot, u)$  based on  $\{(\tilde{Y}_{u,i}, j/m) : 0 \leq j \leq m\}$ , as defined in Section 1.6 of [78]. Then as shown in Theorem 1.8 of [78], with a proper choice of bandwidth and kernel, we have  $\mathbb{E}[\sup_{x \in \mathbb{X}} |\tilde{r}_n(x, u) - r(x, u)|] \leq (\log(m)/m)^{-\beta/(1+2\beta)}$ . Since  $\mathbb{U}$  is finite, the upper bounds in Theorem 4.5 decay at the rate  $(\log(n)/n)^{-\beta/(1+2\beta)}$ .

## 5. CONCLUDING REMARKS

In this paper, we consider the Wasserstein model approximation problem, where we upper bound the performance loss of applying an optimal policy from an approximate model to the true dynamics. This loss is bounded by the sup-norm-induced metric between the approximate and true costs, as well as the Wasserstein-1 distance between the approximate and true transition kernels. We study both the discounted cost and average cost criteria. Based on these results, we develop empirical model learning algorithms, establish their empirical consistency, and obtain sample complexity

bounds. Additionally, we recover and generalize several existing results on continuous dependence, robustness and approximations.

An extension of the present work would be robustness of finite step value iteration, which generalizes [79], as clarified in the following observations. Consider applying Corollary 2.3 to the setup where  $c = \hat{c}$  and both the true and approximate system are control-free, i.e.

$$c(x, u) = c(x), \quad \mathcal{T}(\cdot|x, u) = \mathcal{T}(\cdot|x), \quad \mathcal{S}(\cdot|x, u) = \mathcal{S}(\cdot|x).$$

Further, suppose both probability kernels admit invariant probability measures, and denote them by  $\rho_{\mathcal{T}}$  and  $\rho_{\mathcal{S}}$ , respectively. By the result in Corollary 2.3, we have the following bound

$$\left| \int_{\mathbb{X}} c(x) \rho_{\mathcal{T}}(dx) - \int_{\mathbb{X}} c(x) \rho_{\mathcal{S}}(dx) \right| \leq \frac{\|c\|_{\text{Lip}}}{1 - \|\mathcal{T}\|_{\text{Lip}}} \sup_{x \in \mathbb{X}} (\mathcal{W}_1(\mathcal{T}(\cdot|x), \mathcal{S}(\cdot|x))).$$

We note that the choice of the cost function is arbitrary as long as it is Lipschitz. Restricting our attention to those with Lipschitz constant less than or equal to 1, and taking supremum over such cost functions on both sides, we have

$$\mathcal{W}_1(\rho_{\mathcal{T}}, \rho_{\mathcal{S}}) \leq \frac{1}{1 - \|\mathcal{T}\|_{\text{Lip}}} \sup_{x \in \mathbb{X}} (\mathcal{W}_1(\mathcal{T}(\cdot|x), \mathcal{S}(\cdot|x))).$$

This recovers Corollary 3.1 of [79] in an unweighted form. It is therefore natural to consider that a performance bound for finite-step approximate value iteration would be analogous to Theorem 3.1 in [79].

## REFERENCES

- [1] T. Başar and P. Bernhard, *H-infinity optimal control and related minimax design problems: A dynamic game approach*, Birkhäuser, Boston, MA, 1995.
- [2] K. Zhou, J. C. Doyle, and K. Glover, *Robust and optimal control*, Vol. 40, Prentice-Hall, 1996.
- [3] P. Dupuis, M. R. James, and I. Petersen, *Robust properties of risk-sensitive control*, Mathematics of Control, Signals and Systems **13** (2000), no. 4, 318–332.
- [4] D. Jacobson, *Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games*, IEEE Transactions on Automatic control **18** (1973), no. 2, 124–131.
- [5] P. D. Pra, L. Meneghini, and W. J. Runggaldier, *Connections between stochastic control and dynamic games*, Mathematics of Control, Signals and Systems **9** (1996), no. 4, 303–326.
- [6] I. Petersen, M. R. James, and P. Dupuis, *Minimax optimal control of stochastic uncertain systems with relative entropy constraints*, IEEE Transactions on Automatic Control **45** (2000), no. 3, 398–412.
- [7] R. K. Boel, M. R. James, and I. R. Petersen, *Robustness and risk-sensitive filtering*, IEEE Transactions on Automatic Control **47** (2002), no. 3, 451–461.
- [8] I. I. Tzortzis, C. D. Charalambous, and T. Charalambous, *Dynamic programming subject to total variation distance ambiguity*, SIAM Journal on Control and Optimization **53** (2015), no. 4, 2040–2075.
- [9] J. Blanchet and K. Murthy, *Quantifying distributional model risk via optimal transport*, SSRN Electronic Journal (201604).
- [10] P. Mohajerin Esfahani and D. Kuhn, *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations*, Mathematical Programming **171** (July 2017), no. 1–2, 115–166.
- [11] I. Yang, *Wasserstein distributionally robust stochastic control: A data-driven approach*, IEEE Transactions on Automatic Control **66** (2021), no. 8, 3863–3870.
- [12] S. Wang, N. Si, J. Blanchet, and Z. Zhou, *Statistical learning of distributionally robust stochastic control in continuous state spaces*, arXiv preprint arXiv:2406.11281 (2024), available at [2406.11281](#).
- [13] E. Erdoğan and G. N. Iyengar, *Ambiguous chance constrained problems and robust optimization*, Mathematical Programming **107** (2005), no. 1–2, 37–61.
- [14] H. Sun and H. Xu, *Convergence analysis for distributionally robust optimization and equilibrium problems*, Mathematics of Operations Research **41** (201507), 377–401.
- [15] H. Lam, *Robust sensitivity analysis for stochastic systems*, Mathematics of Operations Research **41** (2016), no. 4, 1248–1275.
- [16] G. N. Iyengar, *Robust dynamic programming*, Mathematics of Operations Research **30** (2005), no. 2, 257–280.



- [17] A. Nilim and L. E. Ghaoui, *Robust control of Markov decision processes with uncertain transition matrices*, Oper. Res. **53** (2005), 780–798.
- [18] B. Øksendal and A. Sulem, *Forward–backward stochastic differential games and stochastic control under model uncertainty*, Journal of Optimization Theory and Applications **161** (2014), no. 1, 22–55.
- [19] A. Benavoli and L. Chisci, *Robust stochastic control based on imprecise probabilities*, IFAC Proceedings Volumes **44** (2011), no. 1, 4606–4613.
- [20] H. Xu and S. Mannor, *Distributionally robust Markov decision processes*, Advances in neural information processing systems **23**, 2010, pp. 2505–2513.
- [21] L. P. Hansen and T. J. Sargent, *Robust control and model uncertainty*, American Economic Review **91** (2001), no. 2, 60–66.
- [22] O. Gossner and T. Tomala, *Entropy bounds on bayesian learning*, Journal of Mathematical Economics **44** (2008), no. 1, 24–32.
- [23] H. J. Langen, *Convergence of dynamic programming models*, Math. Oper. Res. **6** (1981Nov.), no. 4, 493–512.
- [24] A. D. Kara and S. Yüksel, *Robustness to incorrect system models in stochastic control*, SIAM Journal on Control and Optimization **58** (2020), no. 2, 1144–1182.
- [25] A. D. Kara, M. Raginsky, and S. Yüksel, *Robustness to incorrect models and data-driven learning in average-cost optimal stochastic control*, Automatica **139** (2022), 110179.
- [26] A. Müller, *How does the value function of a Markov decision process depend on the transition probabilities?*, Mathematics of Operations Research **22** (1997), no. 4, 872–885.
- [27] O. Hernández-Lerma, *Approximation and adaptive control of Markov processes: Average reward criterion*, Kybernetika **23** (1987), no. 4, 265–288.
- [28] E. I. Gordienko and F. S. Salem, *Robustness inequality for Markov control processes with unbounded costs*, Systems and Control Letters **33** (Feb. 1998), no. 2, 125–130.
- [29] E. I. Gordienko, *An estimate of the stability of optimal control of certain stochastic and deterministic systems*, Journal of Soviet Mathematics **59** (Apr. 1992), no. 4, 891–899.
- [30] E. I. Gordienko and F. Salem-Silva, *Estimates of stability of Markov control processes with unbounded costs*, Kybernetika **36** (2000), no. 2, [195]–210 (eng).
- [31] B. Bozkurt, A. Mahajan, A. Nayyar, and Y. Ouyang, *Model approximation in mdps with unbounded per-step cost*, IEEE Transactions on Automatic Control (2025), 1–16.
- [32] S. Pradhan and S. Yüksel, *Robustness of stochastic optimal control to approximate diffusion models under several cost evaluation criteria*, Mathematics of Operations Research (2023).
- [33] A. D. Kara and S. Yüksel, *Robustness to approximations and model learning in MDPs and POMDPs*, Modern trends in controlled stochastic processes: Theory and applications, volume iii, 2021.
- [34] M. Gheshlaghi Azar, R. Munos, and H. J. Kappen, *Minimax pac bounds on the sample complexity of reinforcement learning with a generative model*, Machine Learning **91** (May 2013), no. 3, 325–349.
- [35] A. Agarwal, S. Kakade, and L. F. Yang, *Model-based reinforcement learning with a generative model is minimax optimal*, Proceedings of thirty third conference on learning theory, 202009, pp. 67–83.
- [36] M. O. Karabag and U. Topcu, *On the sample complexity of vanilla model-based offline reinforcement learning with dependent samples*, Proceedings of the thirty-seventh aaai conference on artificial intelligence and thirty-fifth conference on innovative applications of artificial intelligence and thirteenth symposium on educational advances in artificial intelligence, 2023.
- [37] F. Dufour and T. Prieto-Rumeau, *Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities*, Stochastics **87** (Mar. 2015), no. 2, 273–307.
- [38] B. Wang, Y. Yan, and J. Fan, *Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model*, Advances in neural information processing systems, 2021, pp. 23009–23022.
- [39] B. Van Roy, *Performance loss bounds for approximate value iteration with state aggregation*, Mathematics of Operations Research **31** (May 2006), no. 2, 234–244.
- [40] H. Yu and D. Bertsekas, *Discretized approximations for POMDP with average cost*, arXiv preprint arXiv:1207.4154 (2012), available at [1207.4154](https://arxiv.org/abs/1207.4154).
- [41] N. Saldi, T. Linder, and S. Yüksel, *Finite state approximations of Markov decision processes with general state and action spaces*, 2015 american control conference (acc), July 2015.
- [42] N. Saldi, S. Yüksel, and T. Linder, *Finite-state approximation of Markov decision processes with unbounded costs and Borel spaces*, 2015 54th IEEE conference on decision and control (cdc), Dec. 2015.
- [43] N. Saldi, T. Linder, and S. Yüksel, *Finite approximations in discrete-time stochastic control: Quantized models and asymptotic optimality*, Springer, Cham, 2018.
- [44] S. R. Sinclair, S. Banerjee, and C. L. Yu, *Adaptive discretization in online reinforcement learning*, Operations Research **71** (Sep. 2023), no. 5, 1636–1652.
- [45] A. Kar and R. Singh, *Policy zooming: Adaptive discretization-based infinite-horizon average-reward reinforcement learning*, 2025.



- [46] D. Maran, A. M. Metelli, M. Papini, and M. Restell, *No-regret reinforcement learning in smooth MDPs*, 2024.
- [47] A. Kara and S. Yüksel, *Near optimality of finite memory feedback policies in partially observed Markov decision processes*, Journal of Machine Learning Research **23** (2022), no. 11, 1–46.
- [48] A. D. Kara, N. Saldi, and S. Yüksel, *Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity*, Journal of Machine Learning Research (2023), 1–34.
- [49] A. D. Kara and S. Yüksel, *Q-learning for stochastic control under general information structures and non-Markovian environments*, Transactions on Machine Learning Research (arXiv:2311.00123) (2024).
- [50] E. I. Gordienko, *Stability estimates for controlled Markov chains with a minorant*, Journal of Soviet Mathematics **40** (Feb. 1988), no. 4, 481–486.
- [51] E. Gordienko, E. Lemus-Rodríguez, and R. Montes-de Oca, *Discounted cost optimality problem: stability with respect to weak metrics*, Mathematical Methods of Operations Research **68** (July 2007), no. 1, 77–96.
- [52] E. Gordienko, E. Lemus-Rodríguez, and R. Montes-de Oca, *Average cost Markov control processes: stability with respect to the Kantorovich metric*, Mathematical Methods of Operations Research **70** (June 2008), no. 1, 13–33.
- [53] E. Gordienko and J. R. d. Chavez, *Stability estimation of some Markov controlled processes*, Open Mathematics **20** (Jan. 2022), no. 1, 1509–1520.
- [54] J. Hanson, M. Raginsky, and E. Sontag, *Learning recurrent neural net models of nonlinear systems*, Learning for dynamics and control, 2021, pp. 425–435.
- [55] H. Namkoong and J. C. Duchi, *Stochastic gradient methods for distributionally robust optimization with  $f$ -divergences*, Advances in neural information processing systems, 2016.
- [56] R. Gao and A. Kleywegt, *Distributionally robust stochastic optimization with Wasserstein distance*, Math. Oper. Res. **48** (2023may), no. 2, 603–655.
- [57] A. Figalli and F. Glaudo, *An invitation to optimal transport, Wasserstein distances, and gradient flows*, EMS Press, 2021.
- [58] C. Villani, *Optimal transport*, Springer Berlin Heidelberg, 2009.
- [59] K. Hinderer, *Lipschitz continuity of value functions in Markovian decision processes*, Mathematical Methods of Operations Research **62** (Sep. 2005), no. 1, 3–22.
- [60] K. Asadi, D. Misra, and M. Littman, *Lipschitz continuity in model-based reinforcement learning*, Proceedings of the 35th international conference on machine learning, 201810, pp. 264–273.
- [61] Y. E. Demirci, A. D. Kara, and S. Yüksel, *Average cost optimality of partially observed MDPs: Contraction of nonlinear filters and existence of optimal solutions*, SIAM Journal on Control and Optimization, arXiv:2312.14111 (2024).
- [62] M. Pirodda, M. Restelli, and L. Bascetta, *Policy gradient in Lipschitz Markov decision processes*, Machine Learning **100** (Mar. 2015), no. 2–3, 255–283.
- [63] E. Rachelson and M. G. Lagoudakis, *On the locality of action domination in sequential decision making*, International symposium on artificial intelligence and mathematics, 2010.
- [64] S. Yüksel, *Optimization and control of stochastic systems*, 2024.
- [65] O. Hernández-Lerma, *Adaptive Markov control processes*, Springer New York, 1989.
- [66] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, *Reinforcement learning: Theory and algorithms*, 2021.
- [67] S. P. Meyn and R. Tweedie, *Markov chains and stochastic stability*, Springer-Verlag, London, 1993.
- [68] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov control processes*, Springer New York, 1996.
- [69] A. D. Kara and S. Yüksel, *Q-learning for continuous state and action MDPs under average cost criteria*, arXiv preprint arXiv:2308.07591 (2023).
- [70] N. Saldi, S. Yüksel, and T. Linder, *On the asymptotic optimality of finite approximations to Markov decision processes with Borel spaces*, Mathematics of Operations Research **42** (Nov. 2017), no. 4, 945–978.
- [71] I. Kontoyiannis and S. P. Meyn, *Spectral theory and limit theorems for geometrically ergodic Markov processes*, The Annals of Applied Probability **13** (2003), no. 1, 304–362.
- [72] B. Miasojedow, *Hoeffding’s inequalities for geometrically ergodic Markov chains on general state space*, Statistics & Probability Letters **87** (2014), 115–120.
- [73] R. Douc, E. Moulines, P. Priouret, and P. Soulier, *Markov chains*, Springer, 2018.
- [74] G. O. Roberts and J. S. Rosenthal, *General state space Markov chains and MCMC algorithms*, Prob. Surveys **1** (2004), 20–71.
- [75] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.
- [76] N. Fournier and A. Guillin, *On the rate of convergence in Wasserstein distance of the empirical measure*, Probability Theory and Related Fields **162** (Oct. 2014), no. 3–4, 707–738.
- [77] J. Wellner et al., *Weak convergence and empirical processes: with applications to statistics*, Springer Science & Business Media, 2013.
- [78] A. B. Tsybakov, *Introduction to nonparametric estimation*, Springer New York, 2009.

[79] D. Rudolf and N. Schweizer, *Perturbation theory for Markov chains via Wasserstein distance*, Bernoulli **24** (Nov. 2018), no. 4A.

## APPENDIX A. STATISTICAL ANALYSIS

**A.1. A concentration result.** Let  $\mu$  be a probability measure on  $\mathbb{W}$ . Let  $W, W_0, W_1, \dots$  be i.i.d random variables following the distribution  $\mu$ . For each  $\epsilon > 0$ , denote by  $N(\epsilon, \mathbb{X} \times \mathbb{U}, d_{\mathbb{X}} \times d_{\mathbb{U}})$  the  $\epsilon$ -covering number of  $\mathbb{X} \times \mathbb{U}$ , that is, it is the smallest  $M$  such that there exists  $\{(x_i, u_i), 1 \leq i \leq M\} \subset \mathbb{X} \times \mathbb{U}$  with the property that

$$\mathbb{X} \times \mathbb{U} = \bigcup_{i=1}^M \{(x, u) \in \mathbb{X} \times \mathbb{U} : d_{\mathbb{X}}(x, x_i) + d_{\mathbb{U}}(u, u_i) < \epsilon\}.$$

Further, for a measurable function  $h : \mathbb{W} \rightarrow \mathbb{R}$ , denote by  $\|h\|_{L_2(\mu)} := (\int h^2(w) \mu(dw))^{1/2}$  the  $L_2(\mu)$  norm with respect to  $\mu$ .

**Lemma A.1.** *Let  $B, M > 0$ , and consider a bounded, measurable function  $g : \mathbb{X} \times \mathbb{U} \times \mathbb{W} \rightarrow [B, B]$ .*

*i) Assume that for some constant  $K, L \geq 1$ , we have that for each  $\epsilon > 0$ ,*

$$N(\epsilon, \mathbb{X} \times \mathbb{U}, d_{\mathbb{X}} \times d_{\mathbb{U}}) \leq \left(\frac{K}{\epsilon}\right)^L.$$

*ii) For  $\mu$ -almost sure  $w$ , we have*

$$|g(x, u, w) - g(x', u', w)| \leq M (d_{\mathbb{X}}(x, x') + d_{\mathbb{U}}(u, u')), \text{ for any } (x, u), (x', u') \in \mathbb{X} \times \mathbb{U}.$$

*Then, there exists some absolute constant  $C > 0$  such that*

$$\mathbb{E} \left[ \sup_{(x, u) \in \mathbb{X} \times \mathbb{U}} \left| \frac{1}{n+1} \sum_{i=0}^n g(x, u, W_i) - E[g(x, u, W)] \right| \right] \leq C(B + M)n^{-1/2}.$$

*Proof.* For each  $(x, u) \in \mathbb{X} \times \mathbb{U}$ , define  $g_{x,u}(w) = g(x, u, w)$  for  $w \in \mathbb{W}$ . Let  $\mathcal{G} := \{g_{x,u} : (x, u) \in \mathbb{X} \times \mathbb{U}\}$ . Then by [77, Theorem 2.7.11], for each  $\epsilon > 0$ ,

$$N_{[]} (2\epsilon M, \mathcal{G}, \|\cdot\|_{L_2(\mu)}) \leq N(\epsilon, \mathbb{X} \times \mathbb{U}, d_{\mathbb{X}} \times d_{\mathbb{U}}) \leq \left(\frac{K}{\epsilon}\right)^L, \quad (\text{A.1})$$

where  $N_{[]}(\epsilon, \mathcal{G}, \|\cdot\|_{L_2(\mu)})$  is the  $\epsilon$ -bracketing number under the  $L_2(\mu)$  norm, that is, it is the smallest  $M > 0$  such that there exists  $\{g_1, \dots, g_M\} \subset \mathcal{G}$  with the property that for any  $(x, u) \in \mathbb{X} \times \mathbb{U}$ , we can find some  $1 \leq i^* \leq M$  such that  $\sup_{w \in \mathbb{W}} |g_{x,u}(w) - g_{i^*}(w)| < \epsilon$ .

By Theorem 2.14.2 in [77], for some absolute constant  $C > 0$ , we have

$$\begin{aligned} \Delta &:= \mathbb{E} \left[ \sup_{(x, u) \in \mathbb{X} \times \mathbb{U}} \left| \frac{1}{n+1} \sum_{i=0}^n g(x, u, W_i) - E[g(x, u, W)] \right| \right] \\ &\leq Cn^{-1/2} \int_0^1 \sqrt{1 + \log(N_{[]}(\epsilon B, \mathcal{G}, \|\cdot\|_{L_2(\mu)}))} d\epsilon B, \end{aligned}$$

where we note that the constant function  $B$  is an envelope function for  $\mathcal{G}$ . By a change-of-variable,

$$\begin{aligned} \Delta &\leq Cn^{-1/2} \frac{2M}{B} \int_0^{B/(2M)} \sqrt{1 + \log(N_{[]} (2\epsilon M, \mathcal{G}, \|\cdot\|_{L_2(\mu)}))} d\epsilon B \\ &\leq 2Cn^{-1/2} M(B/(2M) + 1) \int_0^1 \sqrt{1 + K \log(L/\epsilon)} d\epsilon, \end{aligned}$$

where we apply the upper bound (A.1) on the bracketing number in the last step. Since  $\int_0^1 \sqrt{\log(1/\epsilon)} d\epsilon < \infty$ , there exists an absolute constant  $C > 0$  such that  $\Delta \leq Cn^{-1/2}(M + B)$ .  $\square$

*Remark A.1.* The condition i) above holds if  $\mathbb{X}$  and  $\mathbb{U}$  are compact Euclidean subsets, and the constants  $K$  and  $L$  depend on the dimensions and diameters of  $\mathbb{X}$  and  $\mathbb{U}$ .

**A.2. Proof of Lemma 3.3.** In this subsection, we prove Lemma 3.3. By definition, we have that  $\|\hat{c}_N\|_\infty \leq \|c\|_\infty$  and  $\|c^{\pi,M}\|_\infty \leq \|c\|_\infty$ . For each  $i \in [M]$ ,  $u \in \mathbb{U}$ , define

$$A_{i,u} := \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{1}\{X_n \in B_i, U_n = u\}, \quad a_{i,u} := \gamma_u \int_{B_i} \pi(dx),$$

$$\Xi_{i,u} := \frac{1}{N} \sum_{n=0}^{N-1} C_n \mathbb{1}\{X_n \in B_i, U_n = u\}, \quad \xi_{i,u} := \gamma_u \int_{B_i} c(x, u) \pi(dx)$$

By definition of  $\hat{c}_N$  following equation (3.3) and  $c^{\pi,M}$  in (3.1), and the triangle inequality,

$$\|\hat{c}_N - c^{\pi,M}\|_\infty \leq \max_{i \in [M], u \in \mathbb{U}} \frac{1}{A_{i,u}} |\Xi_{i,u} - \xi_{i,u}| + \frac{|\xi_{i,u}|}{A_{i,u} a_{i,u}} |A_{i,u} - a_{i,u}|.$$

By definition,  $\frac{|\xi_{i,u}|}{a_{i,u}} \leq \|c\|_\infty$ . Recall the definition of  $\kappa_{\pi,M}$  prior to Lemma 3.3 and define the event

$$\Sigma_1 := \bigcap_{i \in [M], u \in \mathbb{U}} \{A_{i,u} \geq \kappa_{\pi,M}/2\}. \quad (\text{A.2})$$

By conditioning on the event  $\Sigma_1$  and its complement  $\Sigma_1^c$ , we have

$$\mathbb{E} [\|\hat{c}_N - c^{\pi,M}\|_\infty] \leq \frac{2}{\kappa_{\pi,M}} \mathbb{E} [\|\Xi - \xi\|_\infty] + \frac{2\|c\|_\infty}{\kappa_{\pi,M}} \mathbb{E} [\|A - a\|_\infty] + \|c\|_\infty \mathbb{P}(\Sigma_1^c).$$

Applying Assumption 3.1(b) with the function  $f_{i,u}(x, u', x') := \mathbb{1}\{x \in B_i, u' = u\}$  and  $f_{i,u}(x, u', x') := c(x, u') \mathbb{1}\{x \in B_i, u' = u\}$ , respectively, to the Markov chain  $\{(X_n, U_n, X_{n+1}) : n \geq 0\}$ , and obtain that for each  $i \in [M]$ ,  $u \in \mathbb{U}$ , and  $\epsilon > 0$ ,

$$\max\{\mathbb{P}(|A_{i,u} - a_{i,u}| \geq \epsilon), \mathbb{P}(|\Xi_{i,u} - \xi_{i,u}| \geq \|c\|_\infty \epsilon)\} \leq 2C_0 \exp(-c_0 N \epsilon^2).$$

By the union bound with  $\epsilon = \kappa_{\pi,M}/2$ , we have

$$\mathbb{P}(\Sigma_1^c) \leq 2C_0 M |\mathbb{U}| \exp(-c_0 \kappa_{\pi,M}^2 N/4).$$

Further, by Lemma 2.2.1 in [77], we have  $\left\| \sqrt{N}(A_{i,u} - a_{i,u}) \right\|_{\psi_2}^2 \leq (1 + 2C_0)/c_0$  for  $i \in [M], u \in \mathbb{U}$ , where for a random variable  $Z$ , its  $\psi_2$  norm is defined as follows:  $\|Z\|_{\psi_2} := \inf\{C > 0 : \exp(|Z/C|^2) \leq 2\}$ . Then by Lemma 2.2 in [77], for some absolute constant  $C > 0$ ,

$$\mathbb{E} [\|A - a\|_\infty] = \mathbb{E} \left[ \max_{i \in [M], u \in \mathbb{U}} \|A - a\|_\infty \right] \leq C \sqrt{\frac{\log(M|\mathbb{U}|)}{N}} \sqrt{\frac{1 + 2C_0}{c_0}}.$$

By a similar argument, we have

$$\mathbb{E} [\|\Xi - \xi\|_\infty] \leq C \|c\|_\infty \sqrt{\frac{\log(M|\mathbb{U}|)}{N}} \sqrt{\frac{1 + 2C_0}{c_0}}.$$

The proof for the first claim then is complete by combining the above inequalities.

Now, we focus on the second claim. For each  $i \in [M]$ ,  $u \in \mathbb{U}$ , define

$$\hat{\Xi}_{i,u} := \frac{1}{N} \sum_{n=0}^{N-1} \left( \sum_{j \in [M]} g(y_j) \mathbb{1}\{X_{n+1} \in B_j, X_n \in B_i, U_n = u\} \right),$$

$$\hat{\xi}_{i,u} := \gamma_u \times \sum_{j \in [M]} g(y_j) \int_{B_i} \mathcal{T}(B_j | x, u) \pi(dx).$$

By definition,  $d_g(\hat{\mathcal{T}}_N, \mathcal{T}^{\pi, M}) = \sup_{i \in [M], u \in \mathbb{U}} \left| \hat{\Xi}_{i,u}/A_{i,u} - \hat{\xi}_{i,u}/a_{i,u} \right|$ . Then, since  $\|g\|_\infty \leq L$ , by the same argument as above,

$$\mathbb{E} \left[ d_g(\hat{\mathcal{T}}_N, \mathcal{T}^{\pi, M}) \right] \leq \frac{2}{\kappa_{\pi, M}} \mathbb{E} \left[ \left\| \hat{\Xi} - \hat{\xi} \right\|_\infty \right] + \frac{2L}{\kappa_{\pi, M}} \mathbb{E} [\|A - a\|_\infty] + 2L\mathbb{P}(\Sigma_1^c).$$

Further, again applying Assumption 3.1(b) with the function  $f_{i,u}(x, u', x') := L + \sum_{j \in [M]} g(y_j) \mathbb{1}\{x' \in B_j, x \in B_i, u' = u\}$ , we have

$$\mathbb{P} \left( \left| \hat{\Xi}_{i,u} - \hat{\xi}_{i,u} \right| \geq 2L\epsilon \right) \leq 2C_0 \exp(-c_0 N \epsilon^2),$$

which implies that  $\mathbb{E} \left[ \left\| \hat{\Xi} - \hat{\xi} \right\|_\infty \right] \leq CL \sqrt{\frac{\log(M|\mathbb{U}|)}{N}} \sqrt{\frac{1+2C_0}{c_0}}$ . The proof is then complete.

Finally, we focus on the last claim. For  $i, j \in [M]$  and  $u \in \mathbb{U}$ , define

$$\tilde{\Xi}_{i,j,u} := \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{1}\{X_{n+1} \in B_j, X_n \in B_i, U_n = u\}, \quad \tilde{\xi}_{i,j,u} := \gamma_u \times \int_{B_i} \mathcal{T}(B_j|x, u) \pi(dx).$$

Applying Assumption 3.1(b) with the function  $f_{i,j,u}(x, u', x') := \mathbb{1}\{x \in B_i, u' = u, x' \in B_j\}$  to the Markov chain  $\{(X_n, U_n, X_{n+1}) : n \geq 0\}$ , and obtain that for each  $i, j \in [M]$ ,  $u \in \mathbb{U}$ , and  $\epsilon > 0$ ,

$$\mathbb{P} \left( \left| \tilde{\Xi}_{i,j,u} - \tilde{\xi}_{i,j,u} \right| \geq \epsilon \right) \leq 2C_0 \exp(-c_0 N \epsilon^2).$$

By the definition of the event  $\Sigma_1$  in (A.2), on the event  $\Sigma_1$ , for  $i, j \in [M]$  and  $u \in \mathbb{U}$ ,

$$\left| \hat{\mathcal{T}}_N(y_j|y_i, u) - \mathcal{T}^{\pi, M}(y_j|y_i, u) \right| \leq \frac{2}{\kappa_{\pi, M}} \left| \tilde{\Xi}_{i,j,u} - \tilde{\xi}_{i,j,u} \right| + \frac{2}{\kappa_{\pi, M}} |A_{i,u} - a_{i,u}|.$$

By the definition of  $\mathcal{E}_{N, M}$  in (3.5) and  $\kappa_{\mathcal{T}, M}$  in (3.4), we have

$$\begin{aligned} \mathcal{E}_{N, M} \cap \Sigma_1 &\subset \bigcup_{i, j \in [M], u \in \mathbb{U}} \left\{ \left| \hat{\mathcal{T}}_N(y_j|y_i, u) - \mathcal{T}^{\pi, M}(y_j|y_i, u) \right| > \kappa_{\mathcal{T}, M}/2 \right\} \cap \Sigma_1 \\ &\subset \bigcup_{i, j \in [M], u \in \mathbb{U}} \left( \left\{ \left| \tilde{\Xi}_{i,j,u} - \tilde{\xi}_{i,j,u} \right| > \kappa_{\pi, M} \kappa_{\mathcal{T}, M}/8 \right\} \bigcup \left\{ |A_{i,u} - a_{i,u}| > \kappa_{\pi, M} \kappa_{\mathcal{T}, M}/8 \right\} \right). \end{aligned}$$

As a result, by the union bound, we have

$$\begin{aligned} \mathbb{P}(\mathcal{E}_{N, M}) &\leq \mathbb{P}(\mathcal{E}_{N, M} \cap \Sigma_1) + \mathbb{P}(\Sigma_1^c) \leq \sum_{i, j \in [M], u \in \mathbb{U}} \mathbb{P} \left( \left| \tilde{\Xi}_{i,j,u} - \tilde{\xi}_{i,j,u} \right| > \kappa_{\pi, M} \kappa_{\mathcal{T}, M}/8 \right) \\ &\quad + \sum_{i \in [M], u \in \mathbb{U}} \mathbb{P}(|A_{i,u} - a_{i,u}| > \kappa_{\pi, M} \kappa_{\mathcal{T}, M}/8) + \mathbb{P}(\Sigma_1^c). \end{aligned}$$

Then the proof is complete by combining the previous upper bounds on these probabilities.