Twice regularized MDPs and the equivalence between robustness and regularization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Robust Markov decision processes (MDPs) aim to handle changing or partially known system dynamics. To solve them, one typically resorts to robust optimization methods. However, this significantly increases the computational complexity and limits scalability in both learning and planning. On the other hand, regularized MDPs show more stability in policy learning without impairing time complexity. Yet, they generally do not encompass uncertainty in the model dynamics. In this work, we aim to learn robust MDPs using regularization. We first show that regularized MDPs are a particular instance of robust MDPs with uncertain reward. We thus establish that policy iteration on reward-robust MDPs has the same time complexity as on regularized MDPs. We further extend this relationship to MDPs with uncertain transitions: this leads to a regularization term with an additional dependence on the value function. We finally generalize regularized MDPs to twice regularized MDPs (R^2 MDPs), *i.e.*, MDPs with *both* value and policy regularization. The corresponding Bellman operators enable developing policy iteration schemes with convergence and robustness guarantees. It also reduces planning and learning in robust MDPS to regularized MDPs.

1 Introduction

Markov decision processes (MDPs) provide a practical framework for solving sequential decision problems under uncertainty [30]. However, the chosen strategy can be very sensitive to sampling errors or inaccurate model estimates. This can lead to complete failure in common situations where the model parameters vary adversarially or are simply unknown [21]. Robust MDPs aim to mitigate such sensitivity by assuming that the transition and/or reward function (P, r) varies arbitrarily inside a given *uncertainty set* $\mathcal{U}$ [16, 26]. In this setting, an optimal solution maximizes a performance measure under the worst-case parameters. It can be thought of as a dynamic zero-sum game with an agent choosing the best action while Nature imposes it the most adversarial model. As such, solving robust MDPs involves max-min problems, which can be computationally challenging and limits scalability.

In recent years, several methods have been developed to alleviate the computational concerns raised by robust reinforcement learning (RL). Apart from [22, 23] that consider specific types of coupled uncertainty sets, all rely on a rectangularity assumption without which the problem can be NP-hard [1, 40]. This rectangularity assumption is key to deriving tractable solvers of robust MDPs such as robust value iteration [1, 10] or more general robust modified policy iteration (MPI) [17]. Yet, reducing time complexity in robust Bellman updates remains challenging and is still researched today [14, 10].

At the same time, the empirical success of regularization in policy search methods has motivated a wide range of algorithms with diverse motivations such as improved exploration [11, 19] or stability [34, 12]. Geist et al. [9] proposed a unified view from which many existing algorithms can be derived.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Their regularized MDP formalism enables error propagation analysis in approximate MPI [33] and leads to the same bounds as for standard MDPs. Nevertheless, as we further show in Sec. 3, policy regularization accounts for reward uncertainty only: it does not encompass uncertainty in the model dynamics. Despite a vast literature on *how* regularized policy search works and convergence rates analysis [36, 4], little attention has been given to understanding *why* it can generate strategies that are robust to external perturbations [12].

To our knowledge, the only works that relate robustness to regularization in RL are [5, 15, 8]. Derman & Mannor [5] employ a distributionally robust optimization approach to regularize an empirical value function. Unfortunately, computing this empirical value necessitates several policy evaluation procedures, which is quickly unpractical. Husain et al. [15] provide a dual relationship with robust MDPs under uncertain reward. Their duality result applies to general regularization methods and gives a robust interpretation of soft-actor-critic [12]. Although these two works justify the use of regularization for ensuring robustness, they do not enclose any algorithmic novelty. Similarly, Eysenbach & Levine [8] focus specifically on maximum entropy methods and relate them to either reward or transition robustness. We shall further detail on these most related studies in Sec. 6.

The robustness-regularization duality is well established in statistical learning theory [41, 35, 18], as opposed to RL theory. In fact, standard setups such as classification or regression may be considered as single-stage decision-making problems, *i.e.*, one-step MDPs, a particular case of RL setting. Extending this robustness-regularization duality to RL would yield cheaper learning methods with robustness properties. As such, we propose a regularization function $\Omega_{\mathcal{U}}$ that depends on the uncertainty set $\mathcal{U}$ and is defined over both policy and value spaces (see Sec. 5), thus inducing a *twice regularized* Bellman operator. We show that this regularizer yields an equivalence of the form $v_{\pi,\mathcal{U}} = v_{\pi,\Omega_{\mathcal{U}}}$, where $v_{\pi,\mathcal{U}}$ is the robust value function for policy π and $v_{\pi,\Omega_{\mathcal{U}}}$ the regularized one. This equivalence is derived through the objective function each value optimizes. More concretely, we formulate the robust value function $v_{\pi,\mathcal{U}}$ as an optimal solution of the robust optimization problem:

$$\max_{v\in\mathbb{R}^{\mathcal{S}}} \langle v, \mu_0 \rangle \text{ s. t. } v \leq \inf_{(P,r)\in\mathcal{U}} T^{\pi}_{(P,r)} v, \tag{RO}$$

where $T^{\pi}_{(P,r)}$ is the evaluation Bellman operator. Then, we show that $v_{\pi,\mathcal{U}}$ is also an optimal solution of the convex (non-robust) optimization problem:

$$\max_{v\in\mathbb{R}^{\mathcal{S}}} \langle v, \mu_0 \rangle \text{ s. t. } v \leq T^{\pi}_{(P_0,r_0)} v - \Omega_{\mathcal{U}}(\pi, v), \tag{CO}$$

where (P_0, r_0) is the *nominal model*, establishing equivalence between the two optimization problems. Moreover, the inequality constraint of (CO) enables to derive a *twice regularized* (R^2) Bellman operator defined according to $\Omega_{\mathcal{U}}$, a policy and value regularizer. For ball-constrained uncertainty sets, $\Omega_{\mathcal{U}}$ has an explicit form and under mild conditions, the corresponding R^2 Bellman operators are contracting. The equivalence between the two problems (RO) and (CO) together with the contraction properties of R^2 Bellman operators enable to circumvent robust optimization problems at each Bellman update. As such, it alleviates robust planning and learning algorithms by reducing them to regularized ones, which are as complex as classical methods.

To summarize, we make the following contributions: (i) We show that regularized MDPs are a specific instance of robust MDPs with uncertain reward. Besides formalizing a general connection between the two settings, our result enables to explicit the uncertainty sets induced by standard regularizers. (ii) We generalize this dual relationship to MDPs with uncertain transition and provide the first regularizer that recovers robust MDPs with s-rectangular balls and arbitrary norm. (iii) We introduce twice regularized MDPs that apply both policy and value regularization to retrieve robust MDPs. We establish contraction properties of the corresponding Bellman operators. This leads us to proposing a robust MPI algorithm that opens new perspectives for practical and scalable robust RL algorithms.

Notations. We designate the extended reals by $\overline{\mathbb{R}} := \{-\infty, \infty\}$. Given a finite set $\mathcal{Z}$, the set of real-valued functions (resp. probability distributions) over $\mathcal{Z}$ is denoted by $\mathbb{R}^{\mathcal{Z}}$ (resp. $\Delta_{\mathcal{Z}}$), while the constant function equal to 1 over $\mathcal{Z}$ is denoted by $\mathbb{1}_{\mathcal{Z}}$. The inner product of two functions $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{\mathcal{Z}}$ is defined as $\langle \mathbf{a}, \mathbf{b} \rangle := \sum_{z\in\mathcal{Z}} \mathbf{a}(z)\,\mathbf{b}(z)$, which induces the ℓ_2-norm $\|\mathbf{a}\| := \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}$. The ℓ_2-norm coincides with its dual norm, *i.e.*, $\|\mathbf{a}\| = \max_{\|\mathbf{b}\|\leq 1} \langle \mathbf{a}, \mathbf{b} \rangle =: \|\mathbf{a}\|_*$. Let a function $f : \mathbb{R}^{\mathcal{Z}} \to \overline{\mathbb{R}}$. The Legendre-Fenchel transform (or convex conjugate) of f is $f^*(\mathbf{y}) := \max_{\mathbf{a}\in\mathbb{R}^z} \{\langle \mathbf{a}, \mathbf{y} \rangle - f(\mathbf{a})\}$. Given a set $\mathfrak{Z} \subseteq \mathbb{R}^{\mathcal{Z}}$, the characteristic function $\delta_{\mathfrak{Z}} : \mathbb{R}^{\mathcal{Z}} \to \overline{\mathbb{R}}$ is $\delta_{\mathfrak{Z}}(\mathbf{a}) = 0$ if $\mathbf{a} \in \mathfrak{Z}$; $+\infty$ otherwise. The Legendre-Fenchel transform of $\delta_{\mathfrak{Z}}$ is the support function $\sigma_{\mathfrak{Z}}(\mathbf{y}) = \max_{\mathbf{a}\in\mathfrak{Z}} \langle \mathbf{a}, \mathbf{y} \rangle$ [2, Ex. 1.6.1].

2 Preliminaries

This section describes the background material that we use throughout our work. Firstly, we recall useful properties in convex analysis. Secondly, we address classical discounted MDPs and their linear program (LP) formulation. Thirdly, we briefly detail on regularized MDPs and the associated operators and lastly, we focus on the robust MDP setting.

2.1 Convex Analysis

Let $\Omega : \Delta_{\mathcal{Z}} \to \mathbb{R}$ be a strongly convex function. Throughout this work, the function Ω plays the role of a policy and/or value regularization function. Its Legendre-Fenchel transform Ω^* satisfies several smoothness properties, hence its alternative name "smoothed max operator" [24]. Our work will make use of the following result [13, 24].

Proposition 2.1. *Let $\Omega : \Delta_{\mathcal{Z}} \to \mathbb{R}$ be a strongly convex function. The following properties hold:*
(i) $\nabla\Omega^*$ is Lipschitz and satisfies $\nabla\Omega^*(\mathbf{y}) = \arg\max_{\mathbf{a}\in\Delta_{\mathcal{Z}}} \langle \mathbf{a}, q\rangle - \Omega(\mathbf{a}), \forall\, \mathbf{y} \in \mathbb{R}^{\mathcal{Z}}$.
(ii) For any $c \in \mathbb{R}, \mathbf{y} \in \mathbb{R}^{\mathcal{Z}}$, $\Omega^*(\mathbf{y} + c\mathbb{1}_{\mathcal{Z}}) = \Omega^*(\mathbf{y}) + c$.
(iii) The Legendre-Fenchel transform Ω^* is non-decreasing.

2.2 Discounted MDPs and LP formulation

Consider an infinite horizon MDP $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mu_0, \gamma, P, r)$ with $\mathcal{S}$ and $\mathcal{A}$ finite state and action spaces respectively, $0 < \mu_0 \in \Delta_{\mathcal{S}}$ an initial state distribution and $\gamma \in (0, 1)$ a discount factor. Denoting $\mathcal{X} := \mathcal{S} \times \mathcal{A}$, $P \in \Delta_{\mathcal{S}}^{\mathcal{X}}$ is a transition kernel mapping each state-action pair to a probability distribution over $\mathcal{S}$ and $r \in \mathbb{R}^{\mathcal{X}}$ a reward function. A policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ maps any state $s \in \mathcal{S}$ to an action distribution $\pi_s \in \Delta_{\mathcal{A}}$, and we evaluate its performance through the following measure:

$$\rho(\pi) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \mu_0, \pi, P\right] = \langle v^{\pi}_{(P,r)}, \mu_0\rangle, \tag{1}$$

where the expectation is conditioned on the process distribution determined by μ_0, π and P, and for all $s \in \mathcal{S}$, $v^{\pi}_{(P,r)}(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, \pi, P]$ is the *value function* at state s. Maximizing (1) defines the standard RL objective, which can be solved thanks to the Bellman operators:

$$\begin{aligned} T^{\pi}_{(P,r)} v &:= r^{\pi} + \gamma P^{\pi} v \quad \forall v \in \mathbb{R}^{\mathcal{S}}, \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}, \\ T_{(P,r)} v &:= \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} T^{\pi}_{(P,r)} v \quad \forall v \in \mathbb{R}^{\mathcal{S}}, \\ \mathcal{G}(v) &:= \{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}} : T^{\pi}_{(P,r)} v = T_{(P,r)} v\} \quad \forall v \in \mathbb{R}^{\mathcal{S}}, \end{aligned}$$

where $r^{\pi} := [\langle \pi_s, r(s, \cdot)\rangle]_{s\in\mathcal{S}}$ and $P^{\pi} = [P^{\pi}(s'|s)]_{s',s\in\mathcal{S}}$ with $P^{\pi}(s'|s) := \langle \pi_s, P(s'|s, \cdot)\rangle$. Both $T^{\pi}_{(P,r)}$ and $T_{(P,r)}$ are γ-contractions with respect to (w.r.t.) the supremum norm, so each admits a unique fixed point $v^{\pi}_{(P,r)}$ and $v^{*}_{(P,r)}$, respectively. The set of greedy policies w.r.t. value v defines $\mathcal{G}_{(P,r)}(v)$, and any policy $\pi \in \mathcal{G}(v^{*}_{(P,r)})$ is optimal [30]. For all $v \in \mathbb{R}^{\mathcal{S}}$, the associated function $q \in \mathbb{R}^{\mathcal{X}}$ is given by $q(s, a) = r(s, a) + \gamma\langle P(\cdot|s, a), v\rangle, \forall (s, a) \in \mathcal{X}$. In particular, the fixed point $v^{\pi}_{(P,r)}$ satisfies $v^{\pi}_{(P,r)} = \langle \pi_s, q^{\pi}_{(P,r)}\rangle$ where $q^{\pi}_{(P,r)}$ is its associated q-function.

The problem in (1) can also be formulated as an LP [30]. Given a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, we characterize its performance $\rho(\pi)$ by the following v-LP [30, 25]:

$$\min_{v\in\mathbb{R}^{\mathcal{S}}} \langle v, \mu_0\rangle \text{ subject to (s.t.) } v \geq r^{\pi} + \gamma P^{\pi} v. \tag{P$^{\pi}$}$$

This primal objective provides a policy view of the problem. Alternatively, one may take a state visitation perspective by studying the dual objective instead:

$$\max_{\mu\in\mathbb{R}^{\mathcal{S}}} \langle r^{\pi}, \mu\rangle \text{ s. t. } \mu \geq 0 \text{ and } (\mathbf{Id}_{\mathbb{R}^{\mathcal{S}}} - \gamma P^{\pi}_{*})\mu = \mu_0, \tag{D$^{\pi}$}$$

where P^{π}_{*} is the *adjoint policy transition operator*[1]: $[P^{\pi}_{*}\mu](s) := \sum_{\bar{s}\in\mathcal{S}} P^{\pi}(s|\bar{s})\mu(\bar{s}), \forall \mu \in \mathbb{R}^{\mathcal{S}}$, and $\mathbf{Id}_{\mathcal{S}}$ is the identity function in $\mathbb{R}^{\mathcal{S}}$. Let $\mathbf{I}(s'|s, a) := \delta_{s'=s}, \forall (s, a) \in \mathcal{X}, s' \in \mathcal{S}$ the trivial transition

[1] It is the adjoint operator of P^{π} in the sense that $\langle P^{\pi} v, v'\rangle = \langle v, P^{\pi}_{*} v'\rangle \quad \forall v, v' \in \mathbb{R}^{\mathcal{S}}$.

matrix, and define its *adjoint transition operator* as $\mathbf{I}_*\mu(s) := \sum_{(\bar{s},\bar{a})\in\mathcal{X}} \mathbf{I}(s|\bar{s},\bar{a})\mu(\bar{s},\bar{a}), \forall s \in \mathcal{S}$. The correspondence between either the occupancy measure or the policy view lies in the one-to-one mapping $\mu \mapsto \frac{\mu(\cdot,\cdot)}{\mathbf{I}_*\mu(\cdot)} =: \pi_\mu$ and its inverse $\pi \mapsto \mu_\pi$ given by

$$\mu_\pi(s,a) := \sum_{t=0}^{\infty} \gamma^t \mathbb{P}\left(s_t = s, a_t = a \middle| \mu_0, \pi, P\right), \forall (s,a) \in \mathcal{X}\,.$$

As such, one can interchangeably work with the primal LP (P^π) or the dual (D^π).

2.3 Regularized MDPs

A regularized MDP is a tuple $\mathcal{M}_\Omega := (\mathcal{S}, \mathcal{A}, \mu_0, \gamma, P, r, \Omega)$ with $(\mathcal{S}, \mathcal{A}, \mu_0, \gamma, P, r)$ an infinite horizon MDP as defined above, and $\Omega := (\Omega_s)_{s\in\mathcal{S}}$ a finite set of functions such that for all $s \in \mathcal{S}$, $\Omega_s : \Delta_\mathcal{A} \to \mathbb{R}$ is strongly convex. Each function Ω_s plays the role of a policy regularizer $\Omega_s(\pi_s)$. With a slight abuse of notation, we shall denote by $\Omega(\pi) := (\Omega_s(\pi_s))_{s\in\mathcal{S}}$ the family of state-dependent regularizers.[2] The regularized Bellman evaluation operator is given by

$$[T^{\pi,\Omega}_{(P,r)} v](s) := T^\pi_{(P,r)} v(s) - \Omega_s(\pi_s), \quad \forall v \in \mathbb{R}^\mathcal{S}, s \in \mathcal{S},$$

and the regularized Bellman optimality operator is $T^{*,\Omega}_{(P,r)} v := \max_{\pi\in\Delta^\mathcal{S}_\mathcal{A}} T^{\pi,\Omega}_{(P,r)} v, \forall v \in \mathbb{R}^\mathcal{S}$ [9]. The unique fixed point of $T^{\pi,\Omega}_{(P,r)}$ (respectively $T^{*,\Omega}_{(P,r)}$) is denoted by $v^{\pi,\Omega}_{(P,r)}$ (resp. $v^{*,\Omega}_{(P,r)}$) and called *regularized value function* (resp. *regularized optimal value function*). Although the regularized MDP formalism starts from the aforementioned Bellman operators in [9], it turns out that regularized MDPs are MDPs with modified reward function. Indeed, for any policy $\pi \in \Delta^\mathcal{S}_\mathcal{A}$, the regularized value function is $v^{\pi,\Omega}_{(P,r)} = (\mathbf{I}_\mathcal{S} - \gamma P^\pi)^{-1}(r^\pi - \Omega(\pi))$, which corresponds to a non-regularized value function with expected reward $\tilde{r}^\pi := r^\pi - \Omega(\pi)$. Note that the modified reward $\tilde{r}^\pi(s)$ is no longer linear in π_s because of the strong convexity of Ω_s.

2.4 Robust MDPs

In general, the MDP model is not explicitly known but rather estimated from sampled trajectories. As this may result in over-sensitive outcome [21], robust MDPs aim to reduce such performance variation. Formally, a robust MDP $(\mathcal{S}, \mathcal{A}, \mu_0, \gamma, \mathcal{U})$ is an MDP with uncertain model belonging to $\mathcal{U} := \mathcal{P} \times \mathcal{R}$, *i.e.*, uncertain transition kernel $P \in \mathcal{P} \subseteq \Delta^\mathcal{X}_\mathcal{S}$ and reward function $r \in \mathcal{R} \subseteq \mathbb{R}^\mathcal{X}$ [16, 40]. The uncertainty set $\mathcal{U}$ typically controls the confidence level for the model estimate, which in turn determines the agent's level of robustness. It is given to the agent which seeks to maximize performance under the worst-case model. Although untractable in general, this problem can be solved in polynomial time for *rectangular* uncertainty sets, *i.e.*, for $\mathcal{U} = \times_{s\in\mathcal{S}} \mathcal{U}_s = \times_{s\in\mathcal{S}} (\mathcal{P}_s \times \mathcal{R}_s)$ [40, 22]. For any policy $\pi \in \Delta^\mathcal{S}_\mathcal{A}$ and state $s \in \mathcal{S}$, the *robust value function* at s is $v^{\pi,\mathcal{U}}(s) := \min_{(P,r)\in\mathcal{U}} v^\pi_{(P,r)}(s)$ and the *robust optimal value function* $v^{*,\mathcal{U}}(s) := \max_{\pi\in\Delta^\mathcal{S}_\mathcal{A}} v^{\pi,\mathcal{U}}(s)$. Both objectives can be solved using the robust Bellman operators:

$$[T^{\pi,\mathcal{U}} v](s) := \min_{(P,r)\in\mathcal{U}} T^\pi_{(P,r)} v(s) \quad \forall v \in \mathbb{R}^\mathcal{S}, s \in \mathcal{S}, \pi \in \Delta^\mathcal{A}_\mathcal{S},$$

$$[T^{*,\mathcal{U}} v](s) := \max_{\pi\in\Delta^\mathcal{S}_\mathcal{A}} [T^{\pi,\mathcal{U}} v](s), \quad \forall v \in \mathbb{R}^\mathcal{S},$$

which are γ-contractions, so each admits a unique fixed point $v^{\pi,\mathcal{U}}$ and $v^{*,\mathcal{U}}$, respectively. For all $v \in \mathbb{R}^\mathcal{S}$, the associated robust q-function is given by $q(s,a) = \min_{(P,r)\in\mathcal{U}}\{r(s,a) + \gamma\langle P(\cdot|s,a), v\rangle\}, \forall (s,a) \in \mathcal{X}$, so that $v_{\pi,\mathcal{U}} = \langle \pi_s, q_{\pi,\mathcal{U}}\rangle$ where $q_{\pi,\mathcal{U}}$ is the robust q-function associated to $v_{\pi,\mathcal{U}}$.

3 Reward-robust MDPs

This section focuses on reward-robust MDPs, *i.e.*, robust MDPs with uncertain reward function but known transition model. We show that the regularized MDP formalism introduced in [9] represents a

[2] In the formalism of Geist et al. [9], Ω_s is initially constant over $\mathcal{S}$. However, later in the paper (Sec. 5), it changes according to policy iterates. Here, we alternatively define a family Ω of state-dependent regularizers, in order to account for state-dependent uncertainty sets (see Sec. 5).

particular instance of reward-robust MDP, as both solve the same optimization problem. This equivalence provides a general robustness motivation for the heuristic success of policy regularization. Then, we explicit the uncertainty set underlying some standard regularization functions, thus suggesting an interpretable explanation of their empirical robustness.

We first show the following general result that applies to transition and reward-robust MDPs with random policies. It slightly generalizes [16], as Lemma 3.2 there focuses on uncertain-transition MDPs with deterministic policies. For completeness, we provide a proof of Prop. 3.1 in Appx. A.1.

Proposition 3.1. *For any policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of the robust optimization problem:*

$$\max_{v \in \mathbb{R}^{\mathcal{S}}} \langle v, \mu_0 \rangle \text{ s. t. } v \leq T^{\pi}_{(P,r)} v \text{ for all } (P,r) \in \mathcal{U}. \tag{$\mathrm{P}_{\mathcal{U}}$}$$

In the robust optimization problem ($\mathrm{P}_{\mathcal{U}}$), the inequality constraint must hold over the whole uncertainty set $\mathcal{U}$. As such, a function $v \in \mathbb{R}^{\mathcal{S}}$ is said to be *robust feasible* for ($\mathrm{P}_{\mathcal{U}}$) if $v \leq T^{\pi}_{(P,r)} v$ for all $(P,r) \in \mathcal{U}$ or equivalently, if $\max_{(P,r)\in\mathcal{U}} \{v(s) - T^{\pi}_{(P,r)} v(s)\} \leq 0$ for all $s \in \mathcal{S}$. Therefore, checking robust feasibility requires to solve a maximization problem. For properly structured uncertainty sets, a closed form solution can be derived, which we shall see in the sequel. As standard in the robust RL literature [32, 14, 27], the remaining of this work focuses on uncertainty sets that are centered around a known *nominal model*. Formally, given P_0 (resp. r_0) the nominal transition kernel (resp. reward function), we consider uncertainty sets of the form $(P_0 + \mathcal{P}) \times (r_0 + \mathcal{R})$. Here, the size of $\mathcal{P} \times \mathcal{R}$ quantifies the level of uncertainty or alternatively, the degree of robustness.

3.1 Reward-robust and regularized MDPs: an equivalence

We now focus on reward-robust MDPs, *i.e.*, robust MDPs with $\mathcal{U} = \{P_0\} \times (r_0 + \mathcal{R})$. Thm. 3.1 establishes that reward-robust MDPs are in fact regularized MDPs whose regularizer is given by a support function. Its proof can be found in Appx. A.2. This result brings two take-home messages: (i) policy regularization is equivalent to reward uncertainty; (ii) policy iteration on reward-robust MDPs has the same convergence rate as regularized MDPs, which in turn is the same as standard MDPs [9].

Theorem 3.1 (Reward-robust MDP)**.** *Assume that $\mathcal{U} = \{P_0\} \times (r_0 + \mathcal{R})$. Then, for any policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of the convex optimization problem:*

$$\max_{v \in \mathbb{R}^{\mathcal{S}}} \langle v, \mu_0 \rangle \text{ s. t. } v(s) \leq T^{\pi}_{(P_0,r_0)} v(s) - \sigma_{\mathcal{R}_s}(-\pi_s) \text{ for all } s \in \mathcal{S}\,.$$

Thm. 3.1 clearly highlights a convex regularizer $\Omega_s(\pi_s) := \sigma_{\mathcal{R}_s}(-\pi_s), \forall s \in \mathcal{S}$. We thus recover a regularized MDP by setting $[T^{\pi,\Omega} v](s) = T^{\pi}_{(P_0,r_0)} v(s) - \sigma_{\mathcal{R}_s}(-\pi_s), \forall s \in \mathcal{S}$. In particular, when $\mathcal{R}_s$ is a ball of radius α_s^r, the support function (or regularizer) can be written in closed form as $\Omega_s(\pi_s) := \alpha_s^r \|\pi_s\|$, which is strongly convex. We formalize this below (see proof in Appx. A.3).

Corollary 3.1. *Let $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$. Further assume that for all $s \in \mathcal{S}$, the reward uncertainty set is $\mathcal{R}_s := \{r_s \in \mathbb{R}^{\mathcal{A}} : \|r_s\| \leq \alpha_s^r\}$ and $\mathcal{U} = \{P_0\} \times (r_0 + \mathcal{R})$. Then, the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of the convex optimization problem:*

$$\max_{v \in \mathbb{R}^{\mathcal{S}}} \langle v, \mu_0 \rangle \text{ s. t. } v(s) \leq T^{\pi}_{(P_0,r_0)} v(s) - \alpha_s^r \|\pi_s\| \text{ for all } s \in \mathcal{S}\,.$$

While regularization induces reward-robustness, Thm. 3.1 and Cor. 3.1 suggest that, on the other hand, specific reward-robust MDPs recover well-known policy regularization methods. We explicit the reward-uncertainty sets underlying some of these regularizers in the following.

3.2 Related Algorithms

Consider uncertainty sets of the type $\mathcal{R} := \times_{(s,a)\in\mathcal{X}} \mathcal{R}_{s,a}$. This defines an (s,a)-rectangular $\mathcal{R}$ (a particular type of s-rectangular $\mathcal{R}$) whose rectangles $\mathcal{R}_{s,a}$ are independently defined for all state-action pairs. For each of the regularizers below, we derive appropriate $\mathcal{R}_{s,a}$ that recover the same regularized value function. Note that these reward uncertainty sets depend on the policy. Detailed proofs are in Appx. A.4. There, we also include a summary table that reviews properties of some RL regularizers, as well as our R^2 function which we shall introduce later in Sec. 5.

Negative Shannon entropy. Let $\mathcal{R}^{\mathrm{NS}}_{s,a}(\pi) := (-\infty, \ln(1/\pi_s(a))], \forall (s,a) \in \mathcal{X}$. The associated support function enables to write:

$$\sigma_{\mathcal{R}^{\mathrm{NS}}_s(\pi)}(-\pi_s) = \max_{r(s,a'):r(s,a')\in\mathcal{R}^{\mathrm{NS}}_{s,a'}(\pi),a'\in\mathcal{A}} \sum_{a\in\mathcal{A}} -r(s,a)\pi_s(a) = \sum_{a\in\mathcal{A}} \pi_s(a)\ln(\pi_s(a)),$$

where the last equality results from maximizing $-r(s,a)$ over $(-\infty, \ln(1/\pi_s(a))]$ for each $a \in \mathcal{A}$. We thus recover the negative Shannon entropy $\Omega(\pi_s) = \sum_{a\in\mathcal{A}} \pi_s(a)\ln(\pi_s(a))$ [12].

KL divergence. Let $\mathcal{R}^{\mathrm{KL}}_{s,a}(\pi) := \ln(1/|\mathcal{A}|) + \mathcal{R}^{\mathrm{NS}}_{s,a}(\pi), \forall (s,a) \in \mathcal{X}$. It amounts to translating the interval $\mathcal{R}^{\mathrm{NS}}_{s,a}$ by the given constant. Then, by similarly writing the support function, we recover $\Omega(\pi_s) = \sum_{a\in\mathcal{A}} \pi_s(a)\ln(\pi_s(a)) + \ln(|\mathcal{A}|)$, which is exactly the Kullback-Leibler divergence [34].

Negative Tsallis entropy. Letting $\mathcal{R}^{\mathrm{T}}_{s,a}(\pi) := \left[\frac{1-\pi_s(a)}{2}, +\infty\right), \forall (s,a) \in \mathcal{X}$, we recover the negative Tsallis entropy $\Omega(\pi_s) = \frac{1}{2}(\|\pi_s\|^2 - 1)$ [19].

3.3 Policy-gradient for reward-robust MDPs

The equivalence between reward-robust and regularized MDPs leads us to ask whether we can employ policy-gradient methods [37] on reward-robust MDPs using regularization. The following result establishes that indeed, a policy-gradient theorem can be derived for reward-robust MDPs (see proof in Appx. A.5).

Proposition 3.2. *Assume that $\mathcal{U} = \{P_0\} \times (r_0 + \mathcal{R})$ with $\mathcal{R}_s = \{r_s \in \mathbb{R}^{\mathcal{A}} : \|r_s\| \leq \alpha^r_s\}$. Then, the gradient of the reward-robust objective $J_{\mathcal{U}}(\pi) := \langle v_{\pi,\mathcal{U}}, \mu_0 \rangle$ is given by*

$$\nabla J_{\mathrm{R}^2}(\pi) = \mathbb{E}_{(s,a)\sim\mu_\pi}\left[\nabla \ln \pi_s(a) \left(q_{\pi,\mathcal{U}}(s,a) - \alpha^r_s \frac{\pi_s(a)}{\|\pi_s\|}\right)\right].$$

Although Prop. 3.2 is a particular case of [9, Appx. D.3] for regularized MDPs, its application to reward-robust MDPs is novel and suggests another simplification of robust methods. Indeed, previous works that derive policy-gradient for robust MDPs involve the occupancy measure of the worst-case model, whereas our result sticks to the nominal. In practice, Prop. 3.2 enables to learn a robust policy by sampling transitions from the nominal model instead of all uncertain models. This has a twofold advantage: (i) it avoids an additional computation of the minimum as done in [29, 20, 6]; there, the authors sample next-state transitions and rewards based on all parameters from the uncertainty set, then update the policy based on the worst outcome; (ii) it releases from restricting to finite uncertainty sets. In fact, our regularizer accounts for robustness regardless of the sampling procedure, whereas the parallel simulations in [29, 20, 6] require the uncertainty set to be finite. Technical difficulties are yet to be addressed for generalizing this to transition-uncertain MDPs, because of the interdependence between the regularizer and the value function (see Secs. 4-5). We detail more on this in Appx. A.5.

4 General robust MDPs

Now that we have established policy regularization as a reward-robust problem, we would like to study the opposite question: can any robust MDP, with both uncertain reward and transition, be solved using regularization instead of robust optimization? If so, is the regularization function easy to determine? In this section, we answer positively to both questions for properly defined robust MDPs. This greatly facilitates robust reinforcement learning, as it avoids the increased complexity of robust planning algorithms while still reaching robust performance.

The following theorem establishes that similarly to reward-robust MDPs, any robust MDP can be formulated through regularization (see proof in Appx. B.1). Although the regularizer is also a support function in that case, it depends on both the policy and the value objective, which explains the added difficulty of dealing with robust MDPs.

Theorem 4.1 (Transition-robust MDP)**.** *Assume that $\mathcal{U} = (P_0 + \mathcal{P}) \times (r_0 + \mathcal{R})$. Then, for all policy $\pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}$, the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of the convex optimization problem:*

$$\max_{v\in\mathbb{R}^{\mathcal{S}}} \langle v, \mu_0 \rangle \text{ s. t. } v(s) \leq T^\pi_{(P_0,r_0)} v(s) - \sigma_{\mathcal{R}_s}(-\pi_s) - \sigma_{\mathcal{P}_s}(-\gamma v \cdot \pi_s) \text{ for all } s \in \mathcal{S}, \tag{2}$$

where $[v \cdot \pi_s](s',a) := v(s')\pi_s(a), \forall (s',a) \in \mathcal{X}$.

The upper-bound in the inequality constraint (2) is of the same spirit as the regularized Bellman operator: the first term is a standard, non-regularized Bellman operator on the nominal model (P_0, r_0) to which we subtract a policy-dependent function playing the role of regularization. That support function reminds that of [5, Thm. 3.1], also coming from conjugacy. This is the only similarity between both regularizers: in [5], the Legendre-Fenchel transform is applied on a different type of function, which results in a regularization term that has no closed form but can only be bounded from above. Moreover, the setup considered in [5] is different from our robust MDP setting since it studies a Wasserstein distributionally robust MDP. Therefore, it involves a general convex optimization problem, whereas we focus on the robust formulation of an LP.

When the uncertainty set is a ball, the support function simplifies further, as we show in the following Cor. 4.1. Yet, the dependence of the regularization term in the value function prevents us from readily applying the tool-set of regularized MDPs. We shall later study the properties of this new regularization function in Sec. 5.

Corollary 4.1. *Assume that $\mathcal{U} = (P_0 + \mathcal{P}) \times (r_0 + \mathcal{R})$ with $\mathcal{P}_s := \{P_s \in \mathbb{R}^{\mathcal{X}} : \|P_s\| \leq \alpha_s^P\}$ and $\mathcal{R}_s := \{r_s \in \mathbb{R}^{\mathcal{A}} : \|r_s\| \leq \alpha_s^r\}$ for all $s \in \mathcal{S}$. Then, the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of the convex optimization problem:*

$$\max_{v \in \mathbb{R}^{\mathcal{S}}} \langle v, \mu_0 \rangle \text{ s. t. } v(s) \leq T^{\pi}_{(P_0,r_0)} v(s) - \alpha_s^r \|\pi_s\| - \alpha_s^P \|\pi_s\| \cdot \gamma \|v\| \text{ for all } s \in \mathcal{S}\,. \tag{3}$$

In fact, Cor. 4.1 can be reformulated for any arbitrary norm by replacing each $\|\cdot\|$ by its dual norm $\|\cdot\|_*$ in Eq. (3) (see proof in Appx. B.2): we state Cor. 4.1 with the ℓ_2-norm for notation convenience only, the dual norm of ℓ_2 being norm-ℓ_2 itself. Thus, our regularization function recovers a robust value function independently of the chosen norm, which extends previous results in [14, 10]. Indeed, Ho et al. [14] lighten complexity of robust planning for ℓ_1-norm only, while Grand-Clément & Kroer [10] focus on KL and ℓ_2 ball-constrained uncertainty sets. Both works rely on the specific structure induced by the divergence they consider to derive more efficient robust Bellman updates. Differently, we circumvent these updates using a generic regularization method that is problem-independent while still applying to s-rectangular uncertainty sets as [14, 10] do.

5 R^2 MDPs

In Sec. 4, we showed that for transition-robust MDPs, the optimization constraint involves a regularization term that depends on the value function itself. This adds a difficulty to the reward-robust case where the regularization only depends on the policy. On the other hand, we provided an explicit regularizer for either reward or transition robust MDPs that are ball-constrained. In this section, we introduce R^2 MDPs, a generalization of regularized MDPs that combines policy and value regularization. The core idea is to further regularize the Bellman operators with a value-dependent term that recovers the support functions we derived from the robust optimization problems of Secs. 3-4.

Definition 5.1. *For all $v \in \mathbb{R}^{\mathcal{S}}$, let $\Omega_{v,\mathrm{R}^2} : \Delta_{\mathcal{A}} \to \mathbb{R}$ defined as $\Omega_{v,\mathrm{R}^2}(\pi_s) := \|\pi_s\|(\alpha_s^r + \alpha_s^P \gamma \|v\|)$. The R^2 Bellman evaluation operator is defined as*

$$[T^{\pi,\mathrm{R}^2} v](s) := T^{\pi}_{(P_0,r_0)} v(s) - \Omega_{v,\mathrm{R}^2}(\pi_s), \quad \forall s \in \mathcal{S}\,.$$

The R^2 Bellman optimal operator is defined as

$$[T^{*,\mathrm{R}^2} v](s) := \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} [T^{\pi,\mathrm{R}^2} v](s) = \Omega^*_{v,\mathrm{R}^2}(q_s), \quad \forall s \in \mathcal{S}\,.$$

For any function $v \in \mathbb{R}^{\mathcal{S}}$, the associated unique greedy policy is defined as

$$\pi_s = \arg\max_{\pi_s \in \Delta_{\mathcal{A}}} T^{\pi,\mathrm{R}^2} v(s) = \nabla \Omega^*_{v,\mathrm{R}^2}(q_s), \quad \forall s \in \mathcal{S},$$

that is, in vector form, $\pi = \nabla \Omega^*_{v,\mathrm{R}^2}(q) =: \mathcal{G}_{\Omega_{\mathrm{R}^2}}(v) \iff T^{\pi,\mathrm{R}^2} v = T^{*,\mathrm{R}^2} v$.

The R^2 Bellman evaluation operator is not linear because of the functional norm appearing in the regularization function. Yet, under mild conditions, it is contracting and we can apply Banach's fixed point theorem to define the R^2 value function.

Assumption 5.1 (Bounded radius). *For all $s \in \mathcal{S}$, there exists $\epsilon_s > 0$ such that*

$$\alpha_s^P \leq \min \left(\frac{1-\gamma-\epsilon_s}{\gamma}; \min_{\substack{\mathbf{u}_\mathcal{A} \in \mathbb{R}_+^\mathcal{A}, \|\mathbf{u}_\mathcal{A}\|=1 \\ \mathbf{v}_\mathcal{S} \in \mathbb{R}_+^\mathcal{S}, \|\mathbf{v}_\mathcal{S}\|=1}} \mathbf{u}_\mathcal{A}^\top P_0(\cdot|s,\cdot)\mathbf{v}_\mathcal{S} \right).$$

Algorithm 1: R^2 MPI

Result: π_{k+1}, v_{k+1}
Initialize $v_k \in \mathbb{R}^\mathcal{S}$;
while *not converged* **do**
  $\pi_{k+1} \leftarrow \mathcal{G}_{\Omega_{\mathrm{R}^2}}(v_k)$;
  $v_{k+1} \leftarrow (T^{\pi_{k+1},\mathrm{R}^2})^m v_k$;
end

Asm. 5.1 requires an upper-bound on ball radius for transition uncertainty sets. The first term in the minimum is needed for establishing the contraction property of R^2 Bellman operators, while the second one is used for monotonicity. We remark that the former depends on the original discount factor γ: radius α_s^P must be smaller as γ tends to 1, but as γ decreases to 0, it can grow arbitrarily without altering contraction. Indeed, larger γ implies longer time horizon and higher stochasticity. At the same time, Asm. 5.1 requires tighter level of uncertainty for solving robust MDPs via regularization. Otherwise, both value and policy regularization seem unable to handle the mixed effects of parameter and stochastic uncertainties. Although we recognize a generalized Rayleigh quotient-type problem in the second minimum [28], its interpretation in our context remains unclear. Asm. 5.1 enables the R^2 Bellman operators to admit a unique fixed point, among other nice properties. We formalize this in the following proposition (see proof in Appx. C.1).

Proposition 5.1. *Suppose that Asm. 5.1 holds. Then, we have the following properties:*
(i) Monotonicity: For all $v_1, v_2 \in \mathbb{R}^\mathcal{S}$ such that $v_1 \leq v_2$, we have $T^{\pi,\mathrm{R}^2} v_1 \leq T^{\pi,\mathrm{R}^2} v_2$ and $T^{*,\mathrm{R}^2} v_1 \leq T^{*,\mathrm{R}^2} v_2$.
(ii) Sub-distributivity: For all $c \in \mathbb{R}$, we have $T^{\pi,\mathrm{R}^2}(v_1 + c\mathbb{1}_\mathcal{S}) \leq T^{\pi,\mathrm{R}^2} v_1 + \gamma c \mathbb{1}_\mathcal{S}$ and $T^{*,\mathrm{R}^2}(v_1 + c\mathbb{1}_\mathcal{S}) \leq T^{*,\mathrm{R}^2} v_1 + \gamma c \mathbb{1}_\mathcal{S}$, $\forall c \in \mathbb{R}$.
(iii) Contraction: Let $\epsilon_* := \min_{s\in\mathcal{S}} \epsilon_s > 0$. Then, for all $v_1, v_2 \in \mathbb{R}^\mathcal{S}$, $\|T^{\pi,\mathrm{R}^2} v_1 - T^{\pi,\mathrm{R}^2} v_2\|_\infty \leq (1-\epsilon_*)\|v_1 - v_2\|_\infty$ and $\|T^{*,\mathrm{R}^2} v_1 - T^{*,\mathrm{R}^2} v_2\|_\infty \leq (1-\epsilon_*)\|v_1 - v_2\|_\infty$.

The fixed point property of R^2 Bellman operators leads us to define R^2 value functions as follows.

Definition 5.2 (R^2 value functions). *(i) The R^2 value function v^{π,R^2} is defined as the unique fixed point of the R^2 Bellman evaluation operator: $v^{\pi,\mathrm{R}^2} = T^{\pi,\mathrm{R}^2} v^{\pi,\mathrm{R}^2}$. The associated q-function is $q^{\pi,\mathrm{R}^2}(s,a) = r(s,a) + \gamma \langle P_0(\cdot|s,a), v^{\pi,\mathrm{R}^2} \rangle$. (ii) The R^2 optimal value function v^{*,R^2} is defined as the unique fixed point of the R^2 Bellman optimal operator: $v^{*,\mathrm{R}^2} = T^{*,\mathrm{R}^2} v^{*,\mathrm{R}^2}$. The associated q-function is $q^{*,\mathrm{R}^2}(s,a) = r(s,a) + \gamma \langle P_0(\cdot|s,a), v^{*,\mathrm{R}^2} \rangle$.*

Monotonicity of R^2 Bellman operators plays a key role for reaching an optimal R^2 policy, as we show in the following. A proof can be found in Appx. C.2.

Theorem 5.1 (R^2 optimal policy). *The policy $\pi^{*,\mathrm{R}^2} = \mathcal{G}_{\Omega_{\mathrm{R}^2}}(v^{*,\mathrm{R}^2})$ is the unique optimal R^2 policy, i.e., for all $\pi \in \Delta_\mathcal{A}^\mathcal{S}$, $v^{\pi^*,\mathrm{R}^2} = v^{*,\mathrm{R}^2} \geq v^{\pi,\mathrm{R}^2}$.*

All of the results above ensure convergence of MPI in R^2 MDPs. We call that method R^2 MPI and provide its pseudo-code in Alg. 1. The convergence proof follows the same lines as in [30]. Regarding time complexity of the greedy step, computing the optimal R^2 operator amounts to projecting onto the simplex, which can be efficiently performed in linear time [7].

6 Related Work

Connections between regularization and robustness have already been established in standard statistical learning settings such as support vector machines [41], logistic regression [35] or maximum likelihood estimation [18]. As stated in Sec. 1, these single-stage decision-making problems are a particular instance of RL problems. In that regard, the generalization of robustness-regularization duality to sequential decision-making is seldom studied in the RL literature.

Two works that interpret policy regularization from a robustness perspective are [15] and [8]. In [15], regularization is applied on the dual objective instead of the primal, which has two shortcomings:

(i) It prevents from deriving regularized Bellman operators and dynamic programming methods; (ii) The feasible set is that of occupancy measures, so the connection with standard policy regularization remains unclear. Furthermore, their work focus on reward robustness only. Differently, Eysenbach & Levine [8] address both reward and transition uncertainty by showing that regularized policies with maximum entropy solve a particular type of robust MDPs. Yet, their analysis treats each uncertainty separately, which questions the robustness of the resulting policy when the whole model (P, r) is adversarial. Moreover, their dual relation between entropy regularization and transition-robust MDPs is weak besides applying to specific uncertainty sets. Both of these works treat robustness as an side-effect of regularization more than an objective on its own, whereas we aim to do the opposite, namely, use regularization to solve robust RL problems.

Derman & Mannor [5] similarly view regularization as a tool for achieving robust policies. Through their distributionally robust MDP framework, they show upper and lower bounds between transition-robustness and regularization. There again, duality is weak while reward uncertainty is not considered. Moreover, since the exact regularization term has no explicit form, it is usable through its upper bound only. Finally, regularization is applied on the mean of several value functions $v^{\pi}_{(\hat{P}_i, r)}$, where each $\hat{P}_i$ is a transition model estimated from an episode run. Computing this quantity would require as many policy evaluations as available model estimates available, yielding at least a linear complexity blowup.

Previous studies analyze robust planning algorithms and provide convergence guarantees: in [1, 26] a value iteration algorithm is proposed, the works [16, 40] introduce robust policy iteration and Kaufman & Schaefer [17] generalize both schemes by studying the conditions under which robust MPI converges. All of them guarantee a robust solution in polynomial time which is often insufficient, as the complexity of each Bellman update grows at least cubically in the number of states [14].

In order to reduce time complexity of robust planning algorithms, Ho et al. [14] propose two algorithms that compute robust Bellman updates in $\mathcal{O}(|\mathcal{S}||\mathcal{A}| \log(|\mathcal{S}||\mathcal{A}|))$ operations for ℓ_1-constrained uncertainty sets. Advantageously, our regularization approach reduces each robust Bellman update to its standard, non-robust version of $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$ computations. Moreover, although (s, a) and s-rectangular uncertainty sets are considered in [14], they concern transition-uncertainty only, while our model includes both uncertainties in the general s-rectangular case. Also, their contribution relies on LP formulations that necessitate restricting to ℓ_1-norm. Differently, our formulation applies to any norm, which may come from the fact that our main Theorems (Thms. 3.1-4.1) use Fenchel-Rockafellar duality [31, 3], a generalization of LP duality (see Appx. A.2-B.1). More recently, in [10], Grand-Clément & Kroer propose a first-order method to accelerate robust value iteration under s-rectangular uncertainty sets that are ellipsoidal or KL balls. To our knowledge, our study is the first that reduces time complexity for uncertainty sets as general as s-rectangular sets with arbitrary norm.

7 Conclusion and future work

In this work, we established strong duality between robust and regularized MDPs with a properly chosen regularization function, thus making regularized MDPs a particular case of robust MDPs with uncertain reward. This enabled us to derive a policy-gradient theorem for reward-robust MDPs. When extending this robustness-regularization duality to robust MDPs with uncertain reward and transition, we found that the regularizer depends on the value function besides the policy. We thus introduced R^2 MDPs, a generalization of regularized MDPs with both policy and value regularization. Their related R^2 Bellman operators enable to derive a converging MPI algorithm that achieves the optimal robust value function.

This study provides the theoretical foundations of scalable robust RL. Although our theory focused on planning, the R^2 Bellman operators and their contracting properties open the path to learning algorithms that can (i) use existing RL algorithms and robustify them by simply changing the regularizer; (ii) scale to deep learning settings. Apart from its practical effect, we believe our work opens the path to more theoretical contributions in robust RL. For example, it would be interesting to extend R^2 MPI to the approximate case [33], as non-linearity of the R^2 evaluation operator complicates the analysis. Other possible directions are sample complexity analysis for R^2 MDPs and its comparison with that of robust MDPs, as done recently in [42]; approximate dynamic programming for R^2 MDPs in light of their robust analog [38, 27]. Another line of research is to extend policy-gradient to twice regularized MDPs, as it would avoid parallel learning of adversarial models [6, 39] and can be very useful for continuous control.

References

[1] Bagnell, J. A., Ng, A. Y., and Schneider, J. G. Solving uncertain Markov decision processes. 2001.

[2] Bertsekas, D. P. *Convex optimization theory*. Athena Scientific Belmont, 2009.

[3] Borwein, J. and Lewis, A. S. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

[4] Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

[5] Derman, E. and Mannor, S. Distributional robustness and regularization in reinforcement learning. *ICML Workshop*, 2020.

[6] Derman, E., Mankowitz, D., Mann, T., and Mannor, S. Soft-robust actor-critic policy-gradient. *AUAI press for Association for Uncertainty in Artificial Intelligence*, pp. 208–218, 2018.

[7] Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pp. 272–279, 2008.

[8] Eysenbach, B. and Levine, S. Maximum entropy RL (provably) solves some robust RL problems. *arXiv preprint arXiv:2103.06257*, 2021.

[9] Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

[10] Grand-Clément, J. and Kroer, C. Scalable first-order methods for robust MDPs. *arXiv preprint arXiv:2005.05434*, 2020.

[11] Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361. PMLR, 2017.

[12] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018.

[13] Hiriart-Urruty, J.-B. and Lemaréchal, C. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

[14] Ho, C. P., Petrik, M., and Wiesemann, W. Fast bellman updates for robust MDPs. In *International Conference on Machine Learning*, pp. 1979–1988. PMLR, 2018.

[15] Husain, H., Ciosek, K., and Tomioka, R. Regularized policies are reward robust. In *International Conference on Artificial Intelligence and Statistics*, pp. 64–72. PMLR, 2021.

[16] Iyengar, G. N. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

[17] Kaufman, D. L. and Schaefer, A. J. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.

[18] Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pp. 130–166. INFORMS, 2019.

[19] Lee, K., Choi, S., and Oh, S. Sparse Markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3): 1466–1473, 2018.

[20] Mankowitz, D., Mann, T., Bacon, P.-L., Precup, D., and Mannor, S. Learning robust options. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[21] Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

[22] Mannor, S., Mebel, O., and Xu, H. Lightning does not strike twice: Robust MDPs with coupled uncertainty. *ICML*, 2012.

[23] Mannor, S., Mebel, O., and Xu, H. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

[24] Mensch, A. and Blondel, M. Differentiable dynamic programming for structured prediction and attention. In *International Conference on Machine Learning*, pp. 3462–3471. PMLR, 2018.

[25] Nachum, O. and Dai, B. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

[26] Nilim, A. and El Ghaoui, L. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[27] Panaganti, K. and Kalathil, D. Robust reinforcement learning using least squares policy iteration. *arXiv preprint arXiv:2006.11608*, 2020.

[28] Parlett, B. N. The rayleigh quotient iteration and some generalizations for nonnormal matrices. *Mathematics of Computation*, 28(127):679–693, 1974.

[29] Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826. PMLR, 2017.

[30] Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[31] Rockafellar, R. T. *Convex analysis*, volume 36. Princeton university press, 1970.

[32] Roy, A., Xu, H., and Pokutta, S. Reinforcement learning under model mismatch. *NIPS*, 2017.

[33] Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B., and Geist, M. Approximate modified policy iteration and its application to the game of Tetris. *J. Mach. Learn. Res.*, 16:1629–1676, 2015.

[34] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

[35] Shafieezadeh-Abadeh, S., Esfahani, P. M., and Kuhn, D. Distributionally robust logistic regression. *NIPS*, 2015.

[36] Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.

[37] Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pp. 1057–1063. Citeseer, 1999.

[38] Tamar, A., Mannor, S., and Xu, H. Scaling up robust MDPs using function approximation. In *International Conference on Machine Learning*, pp. 181–189. PMLR, 2014.

[39] Tessler, C., Efroni, Y., and Mannor, S. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pp. 6215–6224. PMLR, 2019.

[40] Wiesemann, W., Kuhn, D., and Rustem, B. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[41] Xu, H., Caramanis, C., and Mannor, S. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009.

[42] Yang, W. and Zhang, Z. Non-asymptotic performances of Robust Markov Decision Processes. *arXiv preprint arXiv:2105.03863*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The main contributions are consistently listed in the abstract and introduction. These are: Showing that regularized MDPs are a particular instance of robust MDPs with uncertain reward (Sec. 3); Extending this relationship to MDPs with uncertain transitions (Sec. 4); Generalizing regularized MDPs to R^2 MDPs (Sec. 5).
   (b) Did you describe the limitations of your work? [Yes] These are described when discussing Prop. 3.2 and Asm. 5.1, near the corresponding statements.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Assumptions are introduced by "Assume that..." or "If..." in each theoretical result.
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs can be found in the Appendix. At each theoretical statement in the text body, we precisely refer to the corresponding proof.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]