Robustness and Regularization in Reinforcement Learning

Esther Derman[*] Mila - Quebec AI Institute
Yevgeniy Men Technion
Matthieu Geist Google Deepmind
Shie Mannor Technion, Nvidia Research

Abstract

Robust Markov decision processes (MDPs) tackle changing or partially known system dynamics. To solve them, one typically resorts to robust optimization, which can significantly increase computational complexity and limit scalability. On the other hand, policy regularization improves learning stability without impairing time complexity. Yet, it does not encompass uncertainty in the model dynamics. In this work, we aim to learn robust MDPs using regularization. We first show that policy regularization methods solve a particular instance of robust MDPs with uncertain rewards. 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 then introduce twice regularized MDPs (R^2 MDPs), *i.e.*, MDPs with value *and* policy regularization. The corresponding Bellman operators lead to planning and learning schemes with convergence and generalization guarantees, thus reducing robustness to regularization. We numerically show this two-fold advantage on tabular and physical domains, and illustrate the persistent efficacy of R^2 regularization.

1 Introduction

MDPs provide a practical framework for solving sequential decision problems under uncertainty [36]. However, the chosen strategy can be overly sensitive to sampling errors or inaccurate model estimates. This can lead to complete failure if the model parameters vary adversarially or are simply unknown [31]. Robust MDPs mitigate such sensitivity by assuming that the transition and/or reward function (P, r) varies arbitrarily inside a given *uncertainty set* $\mathcal{U}$ [24, 35]. In this setting, an optimal solution maximizes return under the worst-case parameters, thus enhancing stability and generalization of the learned policy [48]. Indeed, by construction, any MDP similar to the one the robust agent was trained on would incur stable performance, where 'similar' means belonging to the uncertainty set.

The robust MDP objective can be thought of as a dynamic zero-sum game with an agent choosing the best action while Nature imposes the most adversarial model. As such, solving robust MDPs involves max-min problems, which can be computationally challenging and limit scalability. In recent years, several methods have been developed to alleviate the computational concerns raised by robust reinforcement learning (RL). Apart from [32, 33, 14] which consider specific types of coupled uncertainty sets, all rely on a rectangularity assumption without which the problem can be NP-hard [2, 47]. This assumption is key to deriving tractable solvers of robust MDPs such as robust value iteration [2, 15] or more general robust modified policy iteration (MPI) [27]. Yet, reducing time complexity in robust Bellman updates remains challenging and is still researched today [20, 15, 3, 21, 22].

At the same time, the empirical success of regularization in policy search has motivated a wide range of algorithms for improved exploration [16, 30] or stability [39, 17]. Geist et al. [13] proposed a unified view from which many existing algorithms can be derived. Their regularized MDP formalism allows for error propagation analysis in approximate MPI [38] and leads to the same bounds as for

[*]Contact author: esther.derman@mila.quebec

NeurIPS 2023 Workshop on Generalization in Planning (GenPlan 2023).

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 [41, 8], little attention has been given to understanding *why* it can generate strategies that are robust to external perturbations [17].

To our knowledge, the only works that relate robustness to regularization in RL are [9, 23, 12, 6]. In Derman & Mannor [9], a distributionally robust optimization approach is employed to regularize an empirical value function. Unfortunately, computing this empirical value necessitates several policy evaluation procedures, which is quickly unpractical. The studies [23, 6] 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 [17]. These two works show that regularization ensures robustness, but do not enclose any algorithmic novelty. Similarly, [12] focuses on maximum entropy methods and relates them to either reward or transition robustness.

As opposed to RL theory, the robustness-regularization duality is well established in statistical learning [48, 40, 28]. 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 guarantees. As such, we introduce a regularization function $\Omega_{\mathcal{U}}$ that depends on the uncertainty set $\mathcal{U}$ and is defined over both policy and value spaces, thus inducing a *twice regularized* Bellman operator (see Sec. 5). 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 [36]. 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*. This establishes equivalence as the two problems admit the same optimum for any policy. Moreover, the inequality constraint of (CO) enables us 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.

We make the following contributions: (i) We show that policy regularization leads to a specific instance of robust MDPs with uncertain rewards, and explicitly formulate the uncertainty sets induced by standard policy regularizers. (ii) We extend this duality to MDPs with uncertain transitions and provide the first regularizer that recovers robust MDPs. (iii) We introduce twice regularized MDPs (R^2 MDPs) that apply both policy and value regularization to retrieve robust MDPs with ball constraints. The corresponding Bellman operators are shown to be contracting, which leads to a converging R^2 MPI algorithm of similar time complexity as vanilla MPI. (iv) We introduce R^2 q-learning, a model-free algorithm that provably converges and efficiently solves robust MDPs. (v) We extend R^2 q-learning to a deep variant. In particular, we provide an easy method to estimate the value regularization term when a tabular representation is no longer available. Experiments on tabular and continuous domains prove the efficiency of R^2 for both planning and learning, thus opening new perspectives towards practical and scalable robust RL.

2 Preliminaries

This section describes the background material that we use throughout our work. We first define general notations and recall useful properties in convex analysis. Secondly, we address classical discounted MDPs. Thirdly, we briefly detail regularized MDPs and the associated operators, and lastly, we focus on the robust MDP setting.

2.1 Convex Analysis

We designate the extended reals by $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$. Given a finite set $\mathcal{Z}$, the class 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 $\mathbb{1}_{\mathcal{Z}}$. Similarly, for any set $\mathcal{X}$, $\Delta_{\mathcal{Z}}^{\mathcal{X}}$ denotes the class of functions defined over $\mathcal{X}$ and valued in $\Delta_{\mathcal{Z}}$. 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 $f^*(\mathbf{y}) := \max_{\mathbf{a} \in \mathbb{R}^{\mathcal{Z}}} \{\langle \mathbf{a}, \mathbf{y} \rangle - f(\mathbf{a})\}$. 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 $\sigma_{\mathfrak{Z}}(\mathbf{y}) = \max_{\mathbf{a} \in \mathfrak{Z}} \langle \mathbf{a}, \mathbf{y} \rangle$ [4, Ex. 1.6.1].

Let $C \subset \mathbb{R}^{\mathcal{Z}}$ be a convex set and $\Omega : C \to \mathbb{R}$ a strongly convex function. In our study, the function Ω plays the role of a policy and/or value regularizer. Our work uses the following result [19, 34]:

Proposition 1. *Given $\Omega : C \to \mathbb{R}$ strongly convex, the following properties hold:*
(i) $\nabla \Omega^*$ is Lipschitz and satisfies $\nabla \Omega^*(\mathbf{y}) = \arg\max_{\mathbf{a} \in C} \langle \mathbf{a}, \mathbf{y} \rangle - \Omega(\mathbf{a}), \forall\, \mathbf{y} \in \mathbb{R}^{\mathcal{Z}}$.
(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

Consider an infinite horizon MDP $(\mathcal{S}, \mathcal{A}, \mu_0, \gamma, P, r)$ 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 $r \in \mathbb{R}^{\mathcal{X}}$ is 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:

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

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

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

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 $T_{(P,r)}^{\pi}$ and $T_{(P,r)}$ are γ-contractions with respect to (w.r.t.) the supremum norm, so each admits a unique fixed point $v_{(P,r)}^{\pi}$ and $v_{(P,r)}^{*}$, respectively. The set of greedy policies w.r.t. value v defines $\mathcal{G}_{(P,r)}(v)$, and any policy $\pi \in \mathcal{G}_{(P,r)}(v_{(P,r)}^{*})$ is optimal [36]. For all $v \in \mathbb{R}^{\mathcal{S}}$, the associated function $q \in \mathbb{R}^{\mathcal{X}}$ is given by $q(s, a) = r(s, a) + \gamma \langle P(\cdot|s, a), v \rangle \quad \forall (s, a) \in \mathcal{X}$. In particular, the fixed point $v_{(P,r)}^{\pi}$ satisfies $v_{(P,r)}^{\pi} = \langle \pi_s, q_{(P,r)}^{\pi}(s, \cdot) \rangle$ where $q_{(P,r)}^{\pi}$ is its associated q-function.

2.3 Regularized MDPs

A regularized MDP is a tuple $(\mathcal{S}, \mathcal{A}, \mu_0, \gamma, P, r, \Omega)$ with $(\mathcal{S}, \mathcal{A}, \mu_0, \gamma, P, r)$ an infinite horizon MDP as 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]

[2] In the formalism of Geist et al. [13], Ω_s is initially constant over $\mathcal{S}$. However, later in the paper [13, Sec. 5], it changes according to policy iterates. Here, we alternatively define a family Ω of state-dependent regularizers, which accounts for state-dependent uncertainty sets (see Sec. 5 below).

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 by $T^{*,\Omega}_{(P,r)}v := \max_{\pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}} T^{\pi,\Omega}_{(P,r)}v \quad \forall v \in \mathbb{R}^{\mathcal{S}}$ [13]. 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 defines the *regularized value function* (resp. *regularized optimal value function*). Although the regularized MDP formalism stems from the aforementioned Bellman operators in [13], it turns out that regularized MDPs are MDPs with modified reward. 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 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 Ω_s being strongly convex. Also, this modification does not apply to the reward function r but only to its expectation r^{π}, as we cannot regularize the original reward without making it policy-independent.

2.4 Robust MDPs

In general, the MDP model is not explicitly known but rather estimated from sampled trajectories. Robust MDPs aim to mitigate over-sensitive outcomes this may yield [31]. 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 $P \in \mathcal{P} \subseteq \Delta^{\mathcal{X}}_{\mathcal{S}}$ and reward $r \in \mathcal{R} \subseteq \mathbb{R}^{\mathcal{X}}$ [24, 47]. The uncertainty set $\mathcal{U}$ is given and typically controls the confidence level of a model estimate, which in turn determines the agent's level of robustness. The robust agent seeks to maximize performance under the worst-case model $(P, r) \in \mathcal{U}$. Although intractable in general, this problem can be solved in polynomial time for *rectangular* uncertainty sets, *i.e.*, when $\mathcal{U} = \times_{s \in \mathcal{S}} \mathcal{U}_s = \times_{s \in \mathcal{S}} (\mathcal{P}_s \times \mathcal{R}_s)$ [47, 32]. 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* $v^{*,\mathcal{U}}(s) := \max_{\pi \in \Delta^{\mathcal{A}}_{\mathcal{S}}} v^{\pi,\mathcal{U}}(s)$. Each of them is the unique fixed point of the contracting robust Bellman operators, respectively:

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

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

3 Reward-robust MDPs

This section focuses on reward-robust MDPs, *i.e.*, robust MDPs with uncertain reward but known transition model. We first show that regularized MDPs represent a particular instance of reward-robust MDPs, as both solve the same optimization problem. This equivalence provides a theoretical motivation for the heuristic success of policy regularization. Then, we explicit the uncertainty set underlying some standard regularization functions, which formally explains their empirical robustness.

We first establish the following Prop. 2 that slightly extends [24][Lemma 3.2]. A proof is in Appx. A.1.

Proposition 2. *For any policy $\pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}$, 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, as we shall see in the sequel. As standard in the robust RL literature [37, 20, 1], the remaining of this work focuses on uncertainty sets centered around a known *nominal model*. Formally, given P_0 (resp. r_0) a nominal transition kernel (resp. reward function), we consider uncertainty sets of the form $(P_0 + \mathcal{P}) \times (r_0 + \mathcal{R})$. The size of $\mathcal{P} \times \mathcal{R}$ quantifies our level of uncertainty or alternatively, the desired 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. 1 establishes that reward-robust MDPs are in fact regularized MDPs whose regularizer is given by a support function (see proof 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 [13].

Theorem 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:*

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

where $\sigma_{\mathcal{R}_s}$ is the support function of the reward uncertainty set (see definition in Sec. 2.1).

Thm. 1 highlights a convex regularizer $\Omega_s(\pi_s) := \sigma_{\mathcal{R}_s}(-\pi_s)$, and recovers a regularized MDP by setting $[T^{\pi,\Omega} v](s) = T^{\pi}_{(P_0,r_0)} v(s) - \sigma_{\mathcal{R}_s}(-\pi_s) \quad \forall s \in \mathcal{S}$. In particular, when $\mathcal{R}_s$ is a ball of radius α_s^r, the support function (or regularizer) becomes $\Omega_s(\pi_s) := \alpha_s^r \|\pi_s\|$, which is strongly convex. We formalize this in Appx. A.3.

3.2 Related Algorithms

Thm. 1 shows that regularization induces reward-robustness. At the same time, specific reward-robust MDPs recover well-known policy regularization methods. Consider a reward uncertainty set of the form $\mathcal{R} := \times_{(s,a) \in \mathcal{X}} \mathcal{R}_{s,a}$, *i.e.*, an (s,a)-rectangular $\mathcal{R}$ whose rectangles $\mathcal{R}_{s,a}$ are independent at each state-action pair. For the regularizers below, we derive $\mathcal{R}_{s,a}$-s that produce the same regularized value function. Detailed proofs are in Appx. A.4, along with a table comparing the properties of some RL regularizers with ours (Sec. 5). Note that the reward uncertainty sets here depend on the policy. This is due to the fact that standard regularizers are defined over the policy space and not at each state-action pair. Similarly, the reward transformation induced by policy regularization does not apply to the original function, as already mentioned in Sec. 2.3.

Negative Shannon entropy: Let $\mathcal{R}^{\mathrm{NS}}_{s,a}(\pi) := \left[\ln\left({}^{1}/_{\pi_s(a)}\right), +\infty\right), \quad \forall (s,a) \in \mathcal{X}$. The associated support function gives:

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

which recovers the negative Shannon entropy $\Omega(\pi_s) = \sum_{a \in \mathcal{A}} \pi_s(a) \ln(\pi_s(a))$ [17].

Kullback-Leibler divergence: Given an action distribution $0 < d \in \Delta_{\mathcal{A}}$, let $\mathcal{R}^{\mathrm{KL}}_{s,a}(\pi) := \ln(d(a)) + \mathcal{R}^{\mathrm{NS}}_{s,a}(\pi) \quad \forall (s,a) \in \mathcal{X}$. It amounts to translating the interval $\mathcal{R}^{\mathrm{NS}}_{s,a}$ by the given constant. Writing the support function yields $\Omega(\pi_s) = \sum_{a \in \mathcal{A}} \pi_s(a) \ln\left({}^{\pi_s(a)}/_{d(a)}\right)$, which reveals the KL divergence [39].

Negative Tsallis entropy: Letting $\mathcal{R}^{\mathrm{T}}_{s,a}(\pi) := \left[{}^{(1-\pi_s(a))}/_{2}, +\infty\right) \quad \forall (s,a) \in \mathcal{X}$, we recognize the negative Tsallis entropy $\Omega(\pi_s) = \frac{1}{2}(\|\pi_s\|^2 - 1)$ [30].

4 General robust MDPs

Now that we have established policy regularization as a reward-robust problem, we are interested in the opposite question: can any robust MDP with uncertain reward *and* transition be solved using regularization instead of robust optimization? If so, is the regularization function easy to determine? This section answers positively to both questions. It greatly facilitates robust RL, 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, robust MDPs can be formulated through regularization (see proof in Appx. B.1). The regularizer is also a support function in that case, but it depends on the policy *and* the value objective.

Theorem 2 (General robust MDP)**.** *Assume that $\mathcal{U} = (P_0 + \mathcal{P}) \times (r_0 + \mathcal{R})$. 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:*

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

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

The upper bound in the inequality constraint (2) is similar to the regularized Bellman operator except that here, the regularization is a policy *and* value-dependent function. It further simplifies when the uncertainty set is a ball, as shown below.

Corollary 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:*

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

5 R^2 MDPs

In Sec. 4, we showed that for general robust MDPs, the optimization constraint involves a regularizer that depends on the value function itself. This adds difficulty to the reward-robust case where the regularization only depends on the policy. In this section, we focus on general robust MDPs that are ball-constrained and introduce R^2 MDPs, an extension of regularized MDPs that combines policy with value regularization. The core idea is to regularize the Bellman operators twice and recover the support functions derived in Secs. 3-4.

Definition 1 (R^2 Bellman operators)**.** *For all* $v \in \mathbb{R}^{\mathcal{S}}$, *define* $\Omega_{v,\mathrm{R}^2} : \Delta_{\mathcal{A}} \to \mathbb{R}$ *as* $\Omega_{v,\mathrm{R}^2}(\pi_s) := \|\pi_s\|(\alpha_s^r + \alpha_s^P \gamma \|v\|)$. *The R^2 Bellman evaluation and optimality operators are defined as*

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

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 the following assumption, it is contracting and we can apply Banach's fixed point theorem to define the R^2 value function (see proof in Appx. C.1).

Assumption 1 (Bounded radius)**.** *For all* $s \in \mathcal{S}$, *there exists* $\epsilon_s > 0$ *such that*

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

Proposition 3. *Suppose that Asm. 1 holds. Then, the following properties hold:*
(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 $v_1 \in \mathbb{R}^{\mathcal{S}}, c \in \mathbb{R}$, *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* $v_1, v_2 \in \mathbb{R}^{\mathcal{S}}$, *we have* $\|T^{\pi,\mathrm{R}^2} v_1 - T^{\pi,\mathrm{R}^2} v_2\|_\infty \leq (1 - \epsilon_*)\|v_1 - v_2\|_\infty$ *and* $\|T^{*,\mathrm{R}^2} v_1 - T^{*,\mathrm{R}^2} v_2\|_\infty \leq (1 - \epsilon_*)\|v_1 - v_2\|_\infty$.

The contracting coefficient $1 - \epsilon^*$ from Prop. 3 is different from the original discount γ. Yet, as Asm. 1 suggests it, an intrinsic dependence between γ and ϵ^* makes the R^2 Bellman updates similar to the standard ones: when γ tends to 0, the value of ϵ^* required for Asm. 1 to hold increases, which makes the contracting coefficient $1 - \epsilon^*$ tend to 0 as well, *i.e.*, the two contracting coefficients behave similarly. The contracting feature of both R^2 Bellman operators finally leads us to introduce R^2 value functions.

Definition 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 $q^{\pi,\mathrm{R}^2}(s,a) = r_0(s,a) + \gamma \langle P_0(\cdot|s,a), v^{\pi,\mathrm{R}^2} \rangle$. (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 $q^{*,\mathrm{R}^2}(s,a) = r_0(s,a) + \gamma \langle P_0(\cdot|s,a), v^{*,\mathrm{R}^2} \rangle$.*

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

Theorem 3 (R^2 optimal policy)**.** *The greedy 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 $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, $v^{\pi^*,\mathrm{R}^2} = v^{*,\mathrm{R}^2} \geq v^{\pi,\mathrm{R}^2}$.*

Remark 1. *An optimal R^2 policy may be stochastic. This is because our R^2 MDP framework builds upon the general s-rectangularity assumption. Robust MDPs with s-rectangular uncertainty sets may similarly yield an optimal robust policy that is stochastic [47, Table 1]. Nonetheless, the R^2 MDP formulation recovers a deterministic optimal policy in the more specific (s,a)-rectangular case, which is in accordance with the robust MDP setting (see proof in Appx. C.3).* [3]

6 Planning in R^2 MDPs

The results above ensure convergence of MPI in R^2 MDPs, along with the same geometric convergence rate as in standard and robust MDPs. We call that method R^2 MPI and provide its pseudo-code in Alg. 1. R^2 MPI reduces the computational complexity of robust MPI by avoiding solving a max-min problem at each iteration, which can take polynomial time for general convex programs. The only optimization in R^2 MPI appears in the greedy step, which can efficiently be performed in linear time [11]. In the (s,a)-rectangular case, it even suffices to choose a greedy action (see Rmk. 1).

Algorithm 1 R^2 MPI

Initialize: $v_k \in \mathbb{R}^{\mathcal{S}}$
repeat
  $\pi_{k+1} \leftarrow \mathcal{G}_{\Omega_{\mathrm{R}^2}}(v_k)$
  $v_{k+1} \leftarrow (T^{\pi_{k+1},\mathrm{R}^2})^m v_k$
until convergence
Return: π_{k+1}, v_{k+1}

We compare the computing time of R^2 MPI with that of MPI [36] and robust MPI [27]. The code is available at `https://github.com/EstherDerman/r2mdp`. To do so, we run experiments on a 5×5 grid-world domain: The agent starts from a random position and seeks to reach a goal state in order to maximize reward. Thus, the reward function is zero in all states but two: one provides a reward of 1 while the other gives 10. An episode ends when either one of those two states is attained. Parameter values and other implementation details are deferred to Appx. E. Table 1 shows the time spent by each algorithm until convergence. R^2 PE converges in 0.02 seconds, whereas robust PE takes 54.8 seconds to converge, *i.e.*, 2740 times longer. R^2 PE still takes 2.5 times longer than its standard, non-regularized counterpart, because of the additional computation of regularization terms.

We then study the overall MPI process for each approach. We can see in Table 1 that the increased complexity of robust MPI is even more prominent than its PE thread, as robust MPI takes 3953 (resp. 3270) times longer than R^2 MPI when $m = 1$ (resp. $m = 4$). Robust MPI with $m = 4$ is a bit more advantageous than $m = 1$, as it needs less iterations (31 versus 67), *i.e.*, less optimization solvers to converge. Interestingly, for both $m \in \{1, 4\}$, progressing from PE to MPI did not cost much more computing time to either the vanilla or the R^2 version: both take less than one second to run.

	Vanilla	**R^2**	**Robust**
PE	$0.008 \pm 0.$	$0.02 \pm 0.$	54.8 ± 1.2
MPI ($m = 1$)	$0.01 \pm 0.$	$0.03 \pm 0.$	118.6 ± 1.3
MPI ($m = 4$)	$0.01 \pm 0.$	$0.03 \pm 0.$	98.1 ± 4.1

Table 1: Computing time (in sec.) of planning algorithms using vanilla, R^2 and robust approaches. Each cell displays the mean $\pm$ std obtained from 5 running seeds.

[3] The stochasticity of an optimal entropy-regularized policy as in the examples of Sec. 3.1 is not contradicting. Indeed, even though the corresponding uncertainty set is (s,a)-rectangular there, it is policy-dependent.

7 Learning in R^2 MDPs

In general, we do not know the nominal model (P_0, r_0) and can only interact with the underlying system. Thus, we are interested in devising a model-free method that achieves a robust optimal policy with low time complexity. In the remainder, we assume that $\mathcal{U} = \times_{(s,a)\in\mathcal{X}}\mathcal{U}_{s,a}$ so there exists a deterministic policy which is R^2 optimal. We introduce R^2 q-learning, which provably converges to the optimal robust q-value. Then, we extend R^2 q-learning to a deep variant. In particular, we introduce an easy method for estimating the norm of the R^2 value regularizer in non-tabular settings. The source code for R^2 q-learning and its deep extension is available at `https://github.com/yevgm/r2rl`.

7.1 R^2 q-learning

R^2 q-learning is an R^2 variant of vanilla q-learning [46] aiming to learn a robust optimal policy. Its pseudo-code can be found in Alg. 2 and its convergence in Appx. D.1. The difference with standard q-learning is that we update an R^2 temporal difference (TD) to target an R^2 Bellman recursion. Unlike robust q-learning [37], R^2 q-learning does not involve an optimization problem at each R^2 TD update.

Algorithm 2 R^2 q-learning

Input: Uncertainty levels $\alpha^P, \alpha^r \in \mathbb{R}_+^{\mathcal{X}}$; Learning rates $(\beta_t)_{t\in\mathbb{N}}$ with $\beta_t \in [0,1]^{\mathcal{X}}$;
Initialize: $t = 0$; $q = q_0$ - Arbitrary q-function;
repeat
  Act ϵ-greedily according to $a_t \leftarrow \arg\max_{b\in\mathcal{A}} q_t(s_t, b)$, observe s_{t+1} and obtain r_t
  Set $v_t = \max_{b\in\mathcal{A}} q_t(\cdot, b)$
  Set $\delta_t^{\mathrm{R}^2} = r_t + \gamma \max_{b\in\mathcal{A}} q_t(s_{t+1}, b) - \alpha^r_{s_t a_t} - \gamma\alpha^P_{s_t a_t}\|v_t\| - q_t(s_t, a_t)$
  Update $q_{t+1}(s_t, a_t) = q_t(s_t, a_t) + \beta_t(s_t, a_t)\delta_t^{\mathrm{R}^2}$
until convergence
Return: R^2 value q

7.2 Deep R^2 Learning

The expression of R^2 TD (line 6 of Alg. 2) requires to have access to the whole q-table for computing the current value's norm. This is not possible for infinite or even continuous state-space. Instead, we need to estimate the norm based on sampled observations. We thus keep track of a replay buffer that memorizes and updates past information online. At each iteration, we sample a batch $\mathcal{B}_t$ to which we derive an empirical norm estimate. Formally, $\|v_t\|^2_{\mathcal{B}_t} := \sum_{s\in\mathcal{B}_t} v_t(s)^2$, where $\|\cdot\|_{\mathcal{B}_t}$ indicates the empirical nature of the norm. Finally, our approximate setting motivates us to stabilize value norm estimates. Thus, in the same spirit as [26, 45], we use a moving average mixing the previous estimate with the current one, *i.e.*, at iteration $t+1$, the value norm squared is given by $\beta\|v_t\|^2_{\mathcal{B}_t} + (1-\beta)\|v_{t+1}\|^2_{\mathcal{B}_{t+1}}$.

We thus scale tabular R^2 q-learning to a deep variant we name R^2 double DQN (DDQN) and compare it to vanilla and robust baselines. R^2 DDQN (resp. robust DDQN) is similar to DDQN [18], except that it minimizes an R^2 TD (resp. robust TD) when updating the q-network. For the three algorithmic variants, we use a fully connected network with an input size of the state space dimension, 2 hidden layers of size 256, and an output size corresponding to the dimension of the action space (see Appx. F). We select three physical environments from OpenAI Gym [7]. In each environment, the underlying transition model is directly affected by the physical properties assigned to the agent. Therefore, changing these properties implicitly introduces transition uncertainty into the MDP. We train the three agents on one nominal environment and five different seeds. For a fair comparison, each seed set is taken to be the same for vanilla, robust and R^2 DDQN. Robust and R^2 DDQN are trained under the same uncertainty level, namely, $\alpha^P = \alpha^r = 10^{-4}$. Fig. 1 shows that all three agents converge to similar performance while in Mountaincar, R^2 DDQN outperforms vanilla and robust DDQN.

To check the computational advantage of R^2 DDQN over robust DDQN, we calculate the average time each algorithm takes to perform one update of the q-network. As we see in Tab. 2, one learning step of robust DDQN is slower than one R^2 update by an order of magnitude. On the other hand, one R^2 update is approximately four times slower than vanilla because of the additional computations it re-

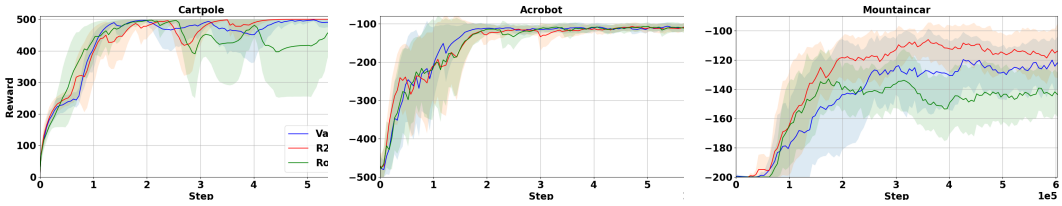


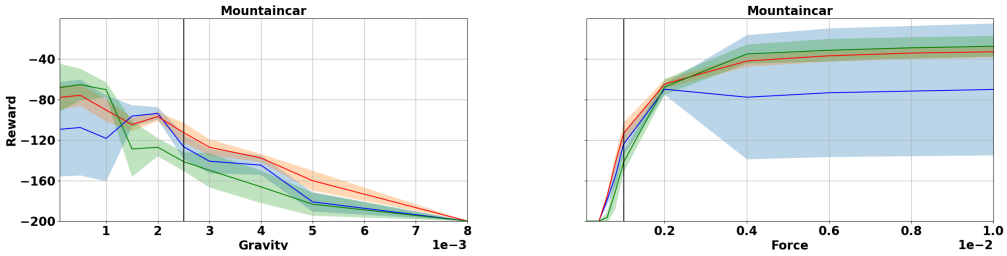
Figure 1: Convergence graphs of vanilla, R^2 and robust DDQN algorithms. Each graph displays the mean $\pm$ standard deviation obtained from 5 running seeds in each environment. The graphs were smoothed with an exponential moving average.

quires. This confirms the results we obtained previously for R^2 MPI and R^2 q-learning: robust updates take much longer than R^2 updates, themselves being slightly slower than standard, non-robust updates.

	Vanilla	**R2**	**Robust**
Cartpole	2.5 ± 0.1	8.3 ± 1.0	76.9 ± 15.3
Acrobot	2.3 ± 0.1	8.1 ± 0.2	73.0 ± 15.3
Mountaincar	2.5 ± 0.8	8.2 ± 0.5	77.6 ± 16.0

Table 2: Average computing time (in $0.1\cdot$ ms) of a learning step for vanilla, R^2 and robust DDQN. Each cell displays the mean $\pm$ std obtained from 1000 iterations.

We aim to check the generalization properties of each algorithm to new dynamics. After training, we select two environment parameters across a range of values and evaluate the average performance over several episodes run under the corresponding dynamics. Fig. 2 displays the performance obtained by each agent undergoing such treatment: R^2 and robust DDQN generalize better than vanilla DDQN, while R^2 is more robust to changing gravity than the other two agents.



(a) Evaluation on Mountaincar: Left - 'gravity', right - 'force' parameter.

Figure 2: Comparison of the average reward over 5 seeds of Vanilla, R^2 and Robust algorithms. The black vertical line represents the nominal parameter value each algorithm was trained on.

8 Discussion

This study settles the theoretical foundations for scalable robust RL. We should note that our results naturally extend to continuous but compact action spaces in the same manner as standard MDPs do [36]. Theoretical extension to infinite state space would be more involved because of the state-dependent regularizer in R^2 MDPs. In fact, it would be interesting to study the R^2 MDP setting under function approximation, as such approximation would have a direct effect on the regularizer. Similarly, one could analyze approximate dynamic programming for R^2 MDPs in light of its robust analog [42, 1]. Apart from its practical effect, we believe our work opens the path to more theoretical contributions in robust RL. For example, extending R^2 MPI to the approximate case [38] would be an interesting problem to solve because of the R^2 evaluation operator being non-linear. So would be a sample complexity analysis for R^2 MDPs with a comparison to robust MDPs [49]. Another line of research is to extend policy-gradient to R^2 MDPs, as this would avoid parallel learning of adversarial models [10, 44] and be very useful for continuous control.

References

[1] Badrinath, K. P. and Kalathil, D. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pp. 511–520. PMLR, 2021.

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

[3] Behzadian, B., Petrik, M., and Ho, C. P. Fast algorithms for l_∞-constrained s-rectangular robust MDPs. *Advances in Neural Information Processing Systems*, 34:25982–25992, 2021.

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

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

[6] Brekelmans, R., Genewein, T., Grau-Moya, J., Delétang, G., Kunesch, M., Legg, S., and Ortega, P. Your policy regularizer is secretly an adversary. *Transactions on Machine Learning Research (TMLR)*, 2022.

[7] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

[8] Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022.

[9] Derman, E. and Mannor, S. Distributional robustness and regularization in reinforcement learning. *International Conference on Machine Learning Workshop*, 2020.

[10] 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.

[11] Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l_1-ball for learning in high dimensions. In *International Conference on Machine learning*, pp. 272–279, 2008.

[12] Eysenbach, B. and Levine, S. Maximum entropy RL (provably) solves some robust RL problems. *International Conference on Learning Representations*, 2022.

[13] 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.

[14] Goyal, V. and Grand-Clement, J. Robust Markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

[15] Grand-Clément, J. and Kroer, C. Scalable first-order methods for robust MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12086–12094, 2021.

[16] 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.

[17] 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.

[18] Hasselt, H. v., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pp. 2094–2100. AAAI Press, 2016.

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

[20] 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.

[21] Ho, C. P., Petrik, M., and Wiesemann, W. Partial policy iteration for l_1-robust Markov decision processes. *J. Mach. Learn. Res.*, 22:275–1, 2021.

[22] Ho, C. P., Petrik, M., and Wiesemann, W. Robust ϕ-divergence MDPs. *arXiv preprint arXiv:2205.14202*, 2022.

[23] 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.

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

[25] Jaakkola, T., Jordan, M., and Singh, S. Convergence of stochastic iterative dynamic programming algorithms. *Advances in Neural Information Processing Systems*, 6, 1993.

[26] Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*. Citeseer, 2002.

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

[28] 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.

[29] Kumar, N., Levy, K., Wang, K., and Mannor, S. Efficient policy iteration for robust Markov decision processes via regularization. *arXiv preprint arXiv:2205.14327*, 2022.

[30] 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.

[31] 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.

[32] Mannor, S., Mebel, O., and Xu, H. Lightning does not strike twice: Robust MDPs with coupled uncertainty. *International Conference on Machine Learning*, 2012.

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

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

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

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

[37] Roy, A., Xu, H., and Pokutta, S. Reinforcement learning under model mismatch. *Advances in Neural Information Processing Systems*, 2017.

[38] 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.

[39] 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.

[40] Shafieezadeh-Abadeh, S., Esfahani, P. M., and Kuhn, D. Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 2015.

[41] 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.

[42] 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.

[43] Tessler, C., Mankowitz, D. J., and Mannor, S. Reward constrained policy optimization. *International Conference on Learning Representations*, 2018.

[44] 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.

[45] Vieillard, N., Pietquin, O., and Geist, M. Deep conservative policy iteration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6070–6077, 2020.

[46] Watkins, C. J. and Dayan, P. Q-learning. *Machine learning*, 8(3):279–292, 1992.

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

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

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

Appendix: Robustness and Regularization in Reinforcement Learning

This appendix provides proofs for all of the results stated in the paper. We first recall the following theorem used in the sequel and referred to as Fenchel-Rochafellar duality [5, Thm 3.3.5].

Theorem (Fenchel-Rockafellar duality)**.** *Let X, Y two Euclidean spaces, $f : X \to \overline{\mathbb{R}}$ and $g : Y \to \overline{\mathbb{R}}$ two proper, convex functions, and $A : X \to Y$ a linear mapping such that $0 \in \operatorname{core}(\operatorname{dom}(g) - A(\operatorname{dom}(f)))$.*[4] *Then, it holds that*

$$\min_{x \in X} f(x) + g(Ax) = \max_{y \in Y} -f^*(-A^*y) - g^*(y). \tag{4}$$

A Reward-Robust MDPs

A.1 Proof of Proposition 2

Proposition. *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:*

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

Proof. Let v^* an optimal point of ($\mathrm{P}_{\mathcal{U}}$). By definition of the robust value function, $v^{\pi,\mathcal{U}} = T^{\pi,\mathcal{U}} v^{\pi,\mathcal{U}} = \min_{(P,r) \in \mathcal{U}} T^{\pi}_{(P,r)} v^{\pi,\mathcal{U}}$. In particular, $v^{\pi,\mathcal{U}} \leq T^{\pi}_{(P,r)} v^{\pi,\mathcal{U}}$ for all $(P,r) \in \mathcal{U}$, so the robust value is feasible and by optimality of v^*, we get $\langle v^*, \mu_0 \rangle \geq \langle v^{\pi,\mathcal{U}}, \mu_0 \rangle$. Now, we aim to show that any feasible $v \in \mathbb{R}^{\mathcal{S}}$ satisfies $v \leq v^{\pi,\mathcal{U}}$. Let an arbitrary $\epsilon > 0$. By definition of $T^{\pi,\mathcal{U}}$, there exists $(P_\epsilon, r_\epsilon) \in \mathcal{U}$ such that

$$T^{\pi,\mathcal{U}} v^{\pi,\mathcal{U}} + \epsilon > T^{\pi}_{(P_\epsilon, r_\epsilon)} v^{\pi,\mathcal{U}}. \tag{5}$$

This yields:

$$\begin{aligned}
v - v^{\pi,\mathcal{U}} &= v - T^{\pi,\mathcal{U}} v^{\pi,\mathcal{U}} && [v^{\pi,\mathcal{U}} = T^{\pi,\mathcal{U}} v^{\pi,\mathcal{U}}] \\
&< v + \epsilon - T^{\pi}_{(P_\epsilon, r_\epsilon)} v^{\pi,\mathcal{U}} && [\text{By Eq. (5)}] \\
&\leq T^{\pi,\mathcal{U}} v + \epsilon - T^{\pi}_{(P_\epsilon, r_\epsilon)} v^{\pi,\mathcal{U}} && [v \text{ is feasible for } (\mathrm{P}_{\mathcal{U}})] \\
&\leq T^{\pi}_{(P_\epsilon, r_\epsilon)} v + \epsilon - T^{\pi}_{(P_\epsilon, r_\epsilon)} v^{\pi,\mathcal{U}} && [T^{\pi,\mathcal{U}} v \leq T^{\pi}_{(P,r)} v \text{ for all } (P,r) \in \mathcal{U}] \\
&= T^{\pi}_{(P_\epsilon, r_\epsilon)} (v - v^{\pi,\mathcal{U}}) + \epsilon. && [\text{By linearity of } T^{\pi}_{(P_\epsilon, r_\epsilon)}]
\end{aligned}$$

Thus, $v - v^{\pi,\mathcal{U}} \leq T^{\pi}_{(P_\epsilon, r_\epsilon)} (v - v^{\pi,\mathcal{U}}) + \epsilon$, which we iteratively apply as follows:

$$\begin{aligned}
v - v^{\pi,\mathcal{U}} &\leq T^{\pi}_{(P_\epsilon, r_\epsilon)} (v - v^{\pi,\mathcal{U}}) + \epsilon \\
&\leq T^{\pi}_{(P_\epsilon, r_\epsilon)} (T^{\pi}_{(P_\epsilon, r_\epsilon)} (v - v^{\pi,\mathcal{U}}) + \epsilon) + \epsilon && [u \leq w \implies T^{\pi}_{(P_\epsilon, r_\epsilon)} u \leq T^{\pi}_{(P_\epsilon, r_\epsilon)} w] \\
&= (T^{\pi}_{(P_\epsilon, r_\epsilon)})^2 (v - v^{\pi,\mathcal{U}}) + \gamma\epsilon + \epsilon \\
&\leq (T^{\pi}_{(P_\epsilon, r_\epsilon)})^2 (T^{\pi}_{(P_\epsilon, r_\epsilon)} (v - v^{\pi,\mathcal{U}}) + \epsilon) + \gamma\epsilon + \epsilon \\
&\vdots \\
&\leq (T^{\pi}_{(P_\epsilon, r_\epsilon)})^{n+1} (v - v^{\pi,\mathcal{U}}) + \sum_{k=0}^{n} \gamma^k \epsilon \\
&= (T^{\pi}_{(P_\epsilon, r_\epsilon)})^{n+1} (v - v^{\pi,\mathcal{U}}) + \frac{1 - \gamma^{n+1}}{1 - \gamma} \epsilon.
\end{aligned}$$

[4] Given $C \subseteq \mathbb{R}^{\mathcal{S}}$, we say that $x \in \operatorname{core}(C)$ if for all $d \in \mathbb{R}^{\mathcal{S}}$ there exists a small enough $t \in \mathbb{R}$ such that $x + td \in C$ [5].

By definition of the sup-norm and applying the triangular inequality we obtain:

$$\begin{aligned} v - v^{\pi,\mathcal{U}} &\leq \left\| (T^{\pi}_{(P_\epsilon, r_\epsilon)})^{n+1}(v - v^{\pi,\mathcal{U}}) \right\|_\infty + \frac{1-\gamma^{n+1}}{1-\gamma}\epsilon \\ &\leq \gamma^{n+1}\|v - v^{\pi,\mathcal{U}}\|_\infty + \frac{1-\gamma^{n+1}}{1-\gamma}\epsilon \qquad [T^{\pi}_{(P_\epsilon, r_\epsilon)} \text{ is } \gamma\text{-contracting}] \end{aligned}$$

Setting $n \to \infty$ yields $v - v^{\pi,\mathcal{U}} \leq \frac{\epsilon}{1-\gamma}$. Since both $\epsilon > 0$ and v were taken arbitrarily, $v^* - v^{\pi,\mathcal{U}} \leq 0$, while we have already shown that $\langle v^*, \mu_0 \rangle \geq \langle v^{\pi,\mathcal{U}}, \mu_0 \rangle$. By positivity of the probability distribution μ_0, it results that $\langle v^*, \mu_0 \rangle = \langle v^{\pi,\mathcal{U}}, \mu_0 \rangle$, and since $\mu_0 > 0$, $v^{\pi,\mathcal{U}} = v^*$. □

A.2 Proof of Theorem 1

Theorem (Reward-robust MDP)**.** *Assume that $\mathcal{U} = \{P_0\} \times (r_0 + \mathcal{R})$. Then, for any 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:*

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

Proof. For all $s \in \mathcal{S}$, define: $F(s) := \max_{(P,r)\in\mathcal{U}} \{v(s) - r^\pi(s) - \gamma P^\pi v(s)\}$. It corresponds to the robust counterpart of ($\mathrm{P}_{\mathcal{U}}$) at $s \in \mathcal{S}$. Thus, the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of:

$$\max_{v \in \mathbb{R}^{\mathcal{S}}} \langle v, \mu_0 \rangle \text{ s. t. } F(s) \leq 0 \text{ for all } s \in \mathcal{S}\,. \tag{6}$$

Based on the structure of the uncertainty set $\mathcal{U} = \{P_0\} \times (r_0 + \mathcal{R})$, we compute the robust counterpart:

$$\begin{aligned} F(s) &= \max_{r' \in r_0 + \mathcal{R}} \{v(s) - r'^\pi(s) - \gamma P_0^\pi v(s)\} \\ &= \max_{r':r' = r_0 + r, r \in \mathcal{R}} \{v(s) - r'^\pi(s) - \gamma P_0^\pi v(s)\} \\ &= \max_{r \in \mathcal{R}} \{v(s) - (r_0^\pi(s) + r^\pi(s)) - \gamma P_0^\pi v(s)\} \qquad [(r_0 + r)^\pi = r_0^\pi + r^\pi \quad \forall \pi \in \Delta^{\mathcal{S}}_{\mathcal{A}}] \\ &= \max_{r \in \mathcal{R}} \{v(s) - r^\pi(s) - r_0^\pi(s) - \gamma P_0^\pi v(s)\} \\ &= \max_{r \in \mathcal{R}} \left\{ v(s) - r^\pi(s) - T^{\pi}_{(P_0,r_0)} v(s) \right\} \qquad [T^{\pi}_{(P_0,r_0)} v(s) = r_0^\pi(s) + \gamma P_0^\pi v(s)] \\ &= \max_{r \in \mathcal{R}} \{-r^\pi(s)\} + v(s) - T^{\pi}_{(P_0,r_0)} v(s) \\ &= \max_{r \in \mathbb{R}^{\mathcal{X}}} \{-r^\pi(s) - \delta_{\mathcal{R}}(r')\} + v(s) - T^{\pi}_{(P_0,r_0)} v(s) \\ &= -\min_{r \in \mathbb{R}^{\mathcal{X}}} \{r^\pi(s) + \delta_{\mathcal{R}}(r)\} + v(s) - T^{\pi}_{(P_0,r_0)} v(s) \\ &= -\min_{r \in \mathbb{R}^{\mathcal{X}}} \{\langle r_s, \pi_s \rangle + \delta_{\mathcal{R}}(r)\} + v(s) - T^{\pi}_{(P_0,r_0)} v(s). \qquad [r^\pi(s) = \langle r_s, \pi_s \rangle] \end{aligned}$$

By the rectangularity assumption, $\mathcal{R} = \times_{s \in \mathcal{S}} \mathcal{R}_s$ and for all $r := (r_s)_{s \in \mathcal{S}} \in \mathbb{R}^{\mathcal{X}}$, we have $\delta_{\mathcal{R}}(r) = \sum_{s' \in \mathcal{S}} \delta_{\mathcal{R}_{s'}}(r_{s'})$. As such,

$$\begin{aligned} F(s) &= -\min_{r \in \mathbb{R}^{\mathcal{X}}} \{\langle r_s, \pi_s \rangle + \sum_{s' \in \mathcal{S}} \delta_{\mathcal{R}_{s'}}(r_{s'})\} + v(s) - T^{\pi}_{(P_0,r_0)} v(s) \\ &= -\min_{r \in \mathbb{R}^{\mathcal{X}}} \{\langle r_s, \pi_s \rangle + \delta_{\mathcal{R}_s}(r_s)\} + v(s) - T^{\pi}_{(P_0,r_0)} v(s), \end{aligned}$$

where the last equality holds since the objective function is minimal if and only if $r_s \in \mathcal{R}_s$.

We now aim to apply Fenchel-Rockafellar duality to the minimization problem. Let the function $f : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}$ defined as $r_s \mapsto \langle r_s, \pi_s \rangle$, and consider the support function $\delta_{\mathcal{R}_s} : \mathbb{R}^{\mathcal{A}} \to \overline{\mathbb{R}}$ together with the identity mapping $\mathbf{Id}_{\mathcal{A}} : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}^{\mathcal{A}}$. Clearly, $\mathrm{dom}(f) = \mathbb{R}^{\mathcal{A}}$, $\mathrm{dom}(\delta_{\mathcal{R}_s}) = \mathcal{R}_s$, and $\mathrm{dom}(\delta_{\mathcal{R}_s}) - \mathbf{Id}_{\mathcal{A}}(\mathrm{dom}(f)) = \mathcal{R}_s - \mathbb{R}^{\mathcal{A}} = \mathbb{R}^{\mathcal{A}}$. Therefore, $\mathrm{core}(\mathrm{dom}(\delta_{\mathcal{R}_s}) - A(\mathrm{dom}(f))) = \mathrm{core}(\mathbb{R}^{\mathcal{A}}) = \mathbb{R}^{\mathcal{A}}$ and $0 \in \mathbb{R}^{\mathcal{A}}$. We can thus apply Fenchel-Rockafellar duality: noting that $\mathbf{Id}_{\mathcal{A}} = (\mathbf{Id}_{\mathcal{A}})^*$ and $(\delta_{\mathcal{R}_s})^*(y) = \sigma_{\mathcal{R}_s}(y)$, we get

$$\min_{r_s \in \mathbb{R}^{\mathcal{A}}} \{f(r_s) + \delta_{\mathcal{R}_s}(r_s)\} = -\min_{y \in \mathbb{R}^{\mathcal{A}}} \{f^*(-y) + (\delta_{\mathcal{R}_s})^*(y)\} = -\min_{y \in \mathbb{R}^{\mathcal{A}}} \{f^*(-y) + \sigma_{\mathcal{R}_s}(y)\}.$$

It remains to compute

$$f^*(-y) = \max_{r_s \in \mathbb{R}^{\mathcal{A}}} -\langle r_s, y\rangle - \langle r_s, \pi_s\rangle = \max_{r_s \in \mathbb{R}^{\mathcal{A}}} \langle r_s, -y - \pi_s\rangle = \begin{cases} 0 \text{ if } -y - \pi_s = 0 \\ +\infty \text{ otherwise} \end{cases},$$

and obtain

$$F(s) = \min_{y \in \mathbb{R}^{\mathcal{A}}} \{f^*(-y) + \sigma_{\mathcal{R}_s}(y)\} + v(s) - T^{\pi}_{(P_0,r_0)} v(s) = \sigma_{\mathcal{R}_s}(-\pi_s) + v(s) - T^{\pi}_{(P_0,r_0)} v(s).$$

We can thus rewrite the optimization problem (6) as:

$$\max_{v \in \mathbb{R}^{\mathcal{S}}} \langle v, \mu_0\rangle \text{ s. t. } \sigma_{\mathcal{R}_s}(-\pi_s) + v(s) - T^{\pi}_{(P_0,r_0)} v(s) \leq 0 \text{ for all } s \in \mathcal{S},$$

which concludes the proof. □

A.3 Reward uncertainty: the ball constraint case

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

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

Proof. We evaluate the support function:

$$\sigma_{\mathcal{R}_s}(-\pi_s) = \max_{r_s \in \mathbb{R}^{\mathcal{A}} : \|r_s\| \leq \alpha_s^r} \langle r_s, -\pi_s\rangle \overset{(1)}{=} \alpha_s^r \|-\pi_s\| = \alpha_s^r \|\pi_s\|,$$

where equality (1) holds by definition of the dual norm. Applying Thm. 1, the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of: $\max_{v \in \mathbb{R}^{\mathcal{S}}} \langle v, \mu_0\rangle$ s. t. $\alpha_s^r \|\pi_s\| + v(s) - T^{\pi}_{(P_0,r_0)} v(s) \leq 0$ for all $s \in \mathcal{S}$, which concludes the proof.

Ball-constraint with arbitrary norm. In the case where reward ball-constraints are defined according to an arbitrary norm $\|\cdot\|_a$ with dual norm $\|\cdot\|_{a^*}$, the support function becomes:

$$\sigma_{\mathcal{R}_s}(-\pi_s) = \max_{r_s \in \mathbb{R}^{\mathcal{A}} : \|r_s\|_a \leq \alpha_s^r} \langle r_s, -\pi_s\rangle = \alpha_s^r \|-\pi_s\|_{a^*} = \alpha_s^r \|\pi_s\|_{a^*}.$$

□

A.4 Related Algorithms: Uncertainty sets from regularizers

Negative Shannon entropy. Each (s,a)-reward uncertainty set is $\mathcal{R}^{\text{NS}}_{s,a}(\pi) := [\ln(1/\pi_s(a)), +\infty)$. We compute the associated support function:

$$\begin{aligned}
\sigma_{\mathcal{R}^{\text{NS}}_s(\pi)}(-\pi_s) &= \max_{r_s \in \mathcal{R}^{\text{NS}}_s(\pi)} \langle r_s, -\pi_s\rangle \\
&= \max_{r(s,a') : r(s,a') \in \mathcal{R}^{\text{NS}}_{s,a'}(\pi), a' \in \mathcal{A}} \sum_{a \in \mathcal{A}} -r(s,a)\pi_s(a) \\
&= \max_{r(s,a') : r(s,a') \geq \ln(1/\pi_s(a)), a' \in \mathcal{A}} -\sum_{a \in \mathcal{A}} \pi_s(a) r(s,a) \\
&= \sum_{a \in \mathcal{A}} \pi_s(a) \ln(\pi_s(a)), \qquad (7)
\end{aligned}$$

where the last equality results from the fact that $\pi_s \geq 0$, and $-r(s,a)\pi_s(a)$ is maximal when $r(s,a)$ is minimal. We thus obtain the negative Shannon entropy.

	Negative Shannon	**KL divergence**	**Negative Tsallis**	**R^2 function**
Regularizer Ω	$\sum_{a\in\mathcal{A}} \pi_s(a)\ln(\pi_s(a))$	$\sum_{a\in\mathcal{A}} \pi_s(a)\ln\left(\frac{\pi_s(a)}{d(a)}\right)$	$\frac{1}{2}(\|\pi_s\|^2-1)$	$\|\pi_s\|(\alpha_s^r+\alpha_s^P\gamma\|v\|)$
Conjugate Ω^*	$\ln\left(\sum_{a\in\mathcal{A}} e^{q_s(a)}\right)$	$\ln\left(\sum_{a\in\mathcal{A}} d(a)e^{q_s(a)}\right)$	$\frac{1}{2}+\frac{1}{2}\sum_{a\in\mathfrak{A}}(q_s(a)^2-\tau(q_s)^2)$	Not in closed-form
Gradient $\nabla\Omega^*$	$\pi_s(a)=\frac{e^{q_s(a)}}{\sum_{b\in\mathcal{A}} e^{q_s(b)}}$	$\pi_s(a)=\frac{e^{q_s(a)}}{\sum_{b\in\mathcal{A}} e^{q_s(b)}}$	$\pi_s(a)=(q_s(a)-\tau(q_s))_+$	Not in closed-form
Reward Uncertainty	(s,a)-rectangular	(s,a)-rectangular	(s,a)-rectangular	s-rectangular
	$\mathcal{R}^{\text{NS}}_{s,a}(\pi) = \left[\ln\left(\frac{1}{\pi_s(a)}\right),+\infty\right)$	$\ln(d(a))+\mathcal{R}^{\text{NS}}_{s,a}(\pi)$	$\left[\frac{1-\pi_s(a)}{2},+\infty\right)$	$\mathbf{B}_{\|\cdot\|}(r_{0s},\alpha_s^r)$
Transition Uncertainty	(s,a)-rectangular	(s,a)-rectangular	(s,a)-rectangular	s-rectangular
	$\{P_0(\cdot\|s,a)\}$	$\{P_0(\cdot\|s,a)\}$	$\{P_0(\cdot\|s,a)\}$	$\mathbf{B}_{\|\cdot\|}(P_{0s},\alpha_s^P)$

Table 3: Summary table of existing policy regularizers and generalization to our R^2 function.

KL divergence. Similarly, given $d\in\Delta_{\mathcal{A}}$, let $\mathcal{R}^{\text{KL}}_{s,a}(\pi) := \ln(d(a))+\mathcal{R}^{\text{NS}}_{s,a}(\pi) \quad \forall (s,a)\in\mathcal{X}$. Then

$$
\begin{aligned}
\sigma_{\mathcal{R}^{\text{KL}}_s(\pi)}(-\pi_s) &= \max_{r(s,a'):r(s,a')\in\mathcal{R}^{\text{KL}}_{s,a'}(\pi),a'\in\mathcal{A}} \sum_{a\in\mathcal{A}} -r(s,a)\pi_s(a)\\
&= \max_{\substack{r(s,a')+\ln(d(a)):\\ r(s,a')\in\mathcal{R}^{\text{NS}}_{s,a'}(\pi),a'\in\mathcal{A}}} \sum_{a\in\mathcal{A}} -r(s,a)\pi_s(a)\\
&= \max_{\substack{r(s,a'):\\ r(s,a')\in\mathcal{R}^{\text{NS}}_{s,a'}(\pi),a'\in\mathcal{A}}} \sum_{a\in\mathcal{A}} -(r(s,a)+\ln(d(a))\pi_s(a)\\
&= \max_{\substack{r(s,a'):\\ r(s,a')\in\mathcal{R}^{\text{NS}}_{s,a'}(\pi),a'\in\mathcal{A}}} \{-\sum_{a\in\mathcal{A}} \pi_s(a)r(s,a)\} - \sum_{a\in\mathcal{A}} \pi_s(a)\ln(d(a))\\
&= \sum_{a\in\mathcal{A}} \pi_s(a)\ln(\pi_s(a)) - \sum_{a\in\mathcal{A}} \pi_s(a)\ln(d(a)),
\end{aligned}
$$

where the last equality uses Eq. (7). We thus recover the KL divergence $\Omega(\pi_s) = \sum_{a\in\mathcal{A}} \pi_s(a)\ln\left({}^{\pi_s(a)}/_{d(a)}\right)$.

Negative Tsallis entropy. Given $\mathcal{R}^{\mathrm{T}}_{s,a}(\pi) := \left[\frac{1-\pi_s(a)}{2}, +\infty\right) \quad \forall (s,a) \in \mathcal{X}$, we compute:

$$\begin{aligned}
\sigma_{\mathcal{R}^{\mathrm{T}}_s(\pi)}(-\pi_s) &= \max_{r(s,a'):r(s,a')\in\mathcal{R}^{\mathrm{T}}_{s,a'}(\pi),a'\in\mathcal{A}} \sum_{a\in\mathcal{A}} -r(s,a)\pi_s(a) \\
&= \max_{r(s,a'):r(s,a')\in\left[\frac{1-\pi_s(a')}{2},+\infty\right),a'\in\mathcal{A}} \sum_{a\in\mathcal{A}} -r(s,a)\pi_s(a) \\
&= \sum_{a\in\mathcal{A}} -\frac{1-\pi_s(a)}{2}\pi_s(a) \qquad (8) \\
&= -\frac{1}{2}\sum_{a\in\mathcal{A}} \pi_s(a) + \frac{1}{2}\sum_{a\in\mathcal{A}} \pi_s(a)^2 = -\frac{1}{2} + \frac{1}{2}\|\pi_s\|^2,
\end{aligned}$$

where Eq. (8) also comes from the fact that $\pi_s \geq 0$, and $-r(s,a)\pi_s(a)$ is maximal when $r(s,a)$ is minimal. We thus obtain the negative Tsallis entropy $\Omega(\pi_s) = \frac{1}{2}(\|\pi_s\|^2 - 1)$.

The reward uncertainty sets associated to both KL and Shannon entropy are similar, as the former amounts to translating the latter by a negative constant (translation to the left). As such, both yield reward values that can be either positive or negative. This is not the case of the negative Tsallis, as its minimal reward is 0, attained for a deterministic action policy, *i.e.*, when $\pi_s(a) = 1$.

Table 3 summarizes the properties of each regularizer. For the Tsallis entropy, we denote by $\tau : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}$ the function $q_s \mapsto \frac{\sum_{a\in\mathfrak{A}(q_s)} q_s(a) - 1}{|\mathfrak{A}(q_s)|}$, where $\mathfrak{A}(q_s) \subseteq \mathcal{A}$ is a subset of actions: $\mathfrak{A}(q_s) = \{a \in \mathcal{A} : 1 + iq_s(a_{(i)}) > \sum_{j=0}^{i} q_s(a_{(j)}), i \in \{1, \cdots, |\mathcal{A}|\}\}$, and $a_{(i)}$ is the action with the i-th maximal value [30].

B General robust MDPs

B.1 Proof of Theorem 2

Theorem (General robust MDP)**.** *Assume that $\mathcal{U} = (P_0 + \mathcal{P}) \times (r_0 + \mathcal{R})$. 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:*

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

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

Proof. The robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of:

$$\max_{v\in\mathbb{R}^{\mathcal{S}}} \langle v, \mu_0\rangle \text{ s. t. } F(s) \leq 0 \text{ for all } s \in \mathcal{S}, \qquad (9)$$

where $F(s) := \max_{(P,r)\in\mathcal{U}} \{v(s) - r^\pi(s) - \gamma P^\pi v(s)\}$ is the robust counterpart of ($\mathrm{P}_\mathcal{U}$) at $s \in \mathcal{S}$. Let's compute it based on the structure of the uncertainty set $\mathcal{U} = (P_0 + \mathcal{P}) \times (r_0 + \mathcal{R})$:

$$
\begin{aligned}
F(s) &= \max_{(P',r')\in(P_0+\mathcal{P})\times(r_0+\mathcal{R})} \{v(s) - r'^\pi(s) - \gamma P'^\pi v(s)\} \\
&= \max_{\substack{P':P'=P_0+P,P\in\mathcal{P} \\ r':r'=r_0+r,r\in\mathcal{R}}} \{v(s) - r'^\pi(s) - \gamma P'^\pi v(s)\} \\
&= \max_{P\in\mathcal{P},r\in\mathcal{R}} \{v(s) - (r_0^\pi(s) + r^\pi(s)) - \gamma(P_0^\pi + P^\pi)v(s)\} \quad [(P_0+P)^\pi = P_0^\pi + P^\pi, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (r_0+r)^\pi = r_0^\pi + r^\pi] \\
&= \max_{P\in\mathcal{P},r\in\mathcal{R}} \{v(s) - r_0^\pi(s) - r^\pi(s) - \gamma P_0^\pi v(s) - \gamma P^\pi v(s)\} \\
&= \max_{P\in\mathcal{P},r\in\mathcal{R}} \left\{v(s) - T^\pi_{(P_0,r_0)}v(s) - r^\pi(s) - \gamma P^\pi v(s)\right\} \quad [T^\pi_{(P_0,r_0)}v(s) = r_0^\pi(s) + \gamma P_0^\pi v(s)] \\
&= \max_{P\in\mathcal{P}}\{-\gamma P^\pi v(s)\} + \max_{r\in\mathcal{R}}\{-r^\pi(s)\} + v(s) - T^\pi_{(P_0,r_0)}v(s) \\
&= -\min_{P\in\mathcal{P}}\{\gamma P^\pi v(s)\} - \min_{r\in\mathcal{R}}\{r^\pi(s)\} + v(s) - T^\pi_{(P_0,r_0)}v(s) \\
&= -\min_{P\in\mathbb{R}^{\mathcal{X}\times\mathcal{S}}}\{\gamma P^\pi v(s) + \delta_\mathcal{P}(P)\} - \min_{r\in\mathbb{R}^\mathcal{X}}\{r^\pi(s) + \delta_\mathcal{R}(r)\} \\
&\quad + v(s) - T^\pi_{(P_0,r_0)}v(s) \\
&= -\min_{P\in\mathbb{R}^{\mathcal{X}\times\mathcal{S}}}\{\gamma\langle P_s^\pi, v\rangle + \delta_\mathcal{P}(P)\} - \min_{r\in\mathbb{R}^\mathcal{X}}\{\langle r_s, \pi_s\rangle + \delta_\mathcal{R}(r)\} \\
&\quad + v(s) - T^\pi_{(P_0,r_0)}v(s). \qquad\qquad [P^\pi v(s) = \langle P_s^\pi, v\rangle, r^\pi(s) = \langle r_s, \pi_s\rangle]
\end{aligned}
$$

As shown in the proof of Thm. 1, $\min_{r\in\mathbb{R}^\mathcal{X}}\{\langle r_s, \pi_s\rangle + \delta_\mathcal{R}(r)\} = \min_{r_s\in\mathbb{R}^\mathcal{A}}\{\langle r_s, \pi_s\rangle + \delta_{\mathcal{R}_s}(r_s)\}$ thanks to the rectangularity assumption. Similarly, by rectangularity of the transition uncertainty set, for all $P := (P_s)_{s\in\mathcal{S}} \in \mathbb{R}^\mathcal{X}$, we have $\delta_\mathcal{P}(P) = \sum_{s'\in\mathcal{S}} \delta_{\mathcal{P}_{s'}}(P_{s'})$. As such,

$$
\begin{aligned}
\min_{P\in\mathbb{R}^{\mathcal{X}\times\mathcal{S}}}\{\gamma\langle P_s^\pi, v\rangle + \delta_\mathcal{P}(P)\} &= \min_{P\in\mathbb{R}^{\mathcal{X}\times\mathcal{S}}}\{\gamma\langle P_s^\pi, v\rangle + \sum_{s'\in\mathcal{S}} \delta_{\mathcal{P}_{s'}}(P_{s'})\} \\
&= \min_{P_s\in\mathbb{R}^\mathcal{X}}\{\gamma\langle P_s^\pi, v\rangle + \delta_{\mathcal{P}_s}(P_s)\},
\end{aligned}
$$

where the last equality holds since the objective function is minimal if and only if $P_s \in \mathcal{P}_s$. Finally,

$$
F(s) = -\min_{P_s\in\mathbb{R}^\mathcal{X}}\{\gamma\langle P_s^\pi, v\rangle + \delta_{\mathcal{P}_s}(P_s)\} - \min_{r_s\in\mathbb{R}^\mathcal{A}}\{\langle r_s, \pi_s\rangle + \delta_{\mathcal{R}_s}(r_s)\} + v(s) - T^\pi_{(P_0,r_0)}v(s).
$$

Referring to the proof of Thm. 1, we know that $-\min_{r\in\mathbb{R}^\mathcal{X}}\{\langle r_s, \pi_s\rangle + \delta_\mathcal{R}(r)\} = \sigma_{\mathcal{R}_s}(-\pi_s)$, so

$$
F(s) = -\min_{P_s\in\mathbb{R}^\mathcal{X}}\{\gamma\langle P_s^\pi, v\rangle + \delta_{\mathcal{P}_s}(P_s)\} + \sigma_{\mathcal{R}_s}(-\pi_s) + v(s) - T^\pi_{(P_0,r_0)}v(s).
$$

Let the matrix $v \cdot \pi_s \in \mathbb{R}^\mathcal{X}$ defined as $[v \cdot \pi_s](s', a) := v(s')\pi_s(a)$ for all $(s', a) \in \mathcal{X}$. Further define $\varphi(P_s) := \gamma\langle P_s^\pi, v\rangle$, which we can rewrite as $\varphi(P_s) = \gamma\langle P_s, v \cdot \pi_s\rangle$. Then, we have that:

$$
\min_{P_s\in\mathbb{R}^\mathcal{X}}\{\gamma\langle P_s^\pi, v\rangle + \delta_{\mathcal{P}_s}(P_s)\} = \min_{P_s\in\mathbb{R}^\mathcal{X}}\{\varphi(P_s) + \delta_{\mathcal{P}_s}(P_s)\} = -\min_{\mathbf{B}\in\mathbb{R}^\mathcal{X}}\{\varphi^*(-\mathbf{B}) + \sigma_{\mathcal{P}_s}(\mathbf{B})\},
$$

where the last equality results from Fenchel-Rockafellar duality and the fact that $(\delta_{\mathcal{P}_s})^* = \sigma_{\mathcal{P}_s}$. It thus remains to compute the convex conjugate of φ:

$$
\begin{aligned}
\varphi^*(-\mathbf{B}) &= \max_{P_s\in\mathbb{R}^\mathcal{X}}\{\langle P_s, -\mathbf{B}\rangle - \varphi(P_s)\} \\
&= \max_{P_s\in\mathbb{R}^\mathcal{X}}\{\langle P_s, -\mathbf{B}\rangle - \gamma\langle P_s, v \cdot \pi_s\rangle\} \\
&= \max_{P_s\in\mathbb{R}^\mathcal{X}}\langle P_s, -\mathbf{B} - \gamma v \cdot \pi_s\rangle \\
&= \begin{cases} 0 \text{ if } -\mathbf{B} - \gamma v \cdot \pi_s = 0 \\ +\infty \text{ otherwise,} \end{cases}
\end{aligned}
$$

which yields $\min_{\mathbf{B}\in\mathbb{R}^\mathcal{X}}\{\varphi^*(-\mathbf{B}) + \sigma_{\mathcal{P}_s}(\mathbf{B})\} = \sigma_{\mathcal{P}_s}(-\gamma v \cdot \pi_s)$. Finally, the robust counterpart rewrites as: $F(s) = \sigma_{\mathcal{P}_s}(-\gamma v \cdot \pi_s) + \sigma_{\mathcal{R}_s}(-\pi_s) + v(s) - T^\pi_{(P_0,r_0)}v(s)$, and plugging it into the optimization problem (9) yields the desired result. □

B.2 Proof of Corollary 1

Corollary. *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:*

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

Proof. As we already showed in Appx. A.3, the support function of the reward uncertainty set is $\sigma_{\mathcal{R}_s}(-\pi_s) = \alpha_s^r \|\pi_s\|$. For the transition uncertainty set, we similarly have:

$$\begin{aligned}
\sigma_{\mathcal{P}_s}(-\gamma v \cdot \pi_s) &= \max_{\substack{P_s \in \mathbb{R}^{\mathcal{X}}: \\ \|P_s\| \leq \alpha_s^P}} \langle P_s, -\gamma v \cdot \pi_s \rangle \\
&= \alpha_s^P \|-\gamma v \cdot \pi_s\| \\
&= \alpha_s^P \gamma \|v \cdot \pi_s\| \\
&= \alpha_s^P \gamma \|v\| \|\pi_s\|. \qquad [\|v \cdot \pi_s\|^2 = \sum_{(s',a) \in \mathcal{X}} (v(s')\pi_s(a))^2 \\
&\qquad = \sum_{s' \in \mathcal{S}} v(s')^2 \sum_{a \in \mathcal{A}} \pi_s(a)^2 = \|v\|^2 \|\pi_s\|^2]
\end{aligned}$$

Now we apply Thm. 1 and replace each support function by their explicit form to get that the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of:

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

Ball-constraints with arbitrary norms. As seen in the proof of Thm. 1 and in Appx. A.3, for ball-constrained rewards defined with an arbitrary norm $\|\cdot\|_a$ of dual $\|\cdot\|_{a^*}$, the corresponding support function is $\sigma_{\mathcal{R}_s}(-\pi_s) = \alpha_s^r \|\pi_s\|_{a^*}$. Similarly, for ball-constrained transitions based on a norm $\|\cdot\|_b$ of dual $\|\cdot\|_{b^*}$, we have:

$$\sigma_{\mathcal{P}_s}(-\gamma v \cdot \pi_s) = \max_{\substack{P_s \in \mathbb{R}^{\mathcal{X}}: \\ \|P_s\|_b \leq \alpha_s^P}} \langle P_s, -\gamma v \cdot \pi_s \rangle = \alpha_s^P \|-\gamma v \cdot \pi_s\|_{b^*},$$

in which case the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of:

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

□

C R^2 MDPs

C.1 Proof of Proposition 3

Proposition. *Suppose that Asm. 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 $v_1 \in \mathbb{R}^{\mathcal{S}}, c \in \mathbb{R}$, 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 $\|T^{*,\mathrm{R}^2} v_1 - T^{*,\mathrm{R}^2} v_2\|_\infty \leq (1 - \epsilon_*)\|v_1 - v_2\|_\infty$.

Proof. Proof of (i). Consider the evaluation operator and let $v_1, v_2 \in \mathbb{R}^{\mathcal{S}}$ such that $v_1 \leq v_2$. For all $s \in \mathcal{S}$,

$$
\begin{aligned}
&[T^{\pi,\mathrm{R}^2} v_1 - T^{\pi,\mathrm{R}^2} v_2](s) \\
&= T^{\pi}_{(P_0,r_0)} v_1(s) - \alpha^r_s \|\pi_s\| - \alpha^P_s \gamma \|v_1\| \|\pi_s\| \\
&\quad - (T^{\pi}_{(P_0,r_0)} v_2(s) - \alpha^r_s \|\pi_s\| - \alpha^P_s \gamma \|v_2\| \|\pi_s\|) \\
&= T^{\pi}_{(P_0,r_0)} v_1(s) - T^{\pi}_{(P_0,r_0)} v_2(s) + \alpha^P_s \gamma \|\pi_s\| (\|v_2\| - \|v_1\|) \\
&= \gamma P^{\pi}_0 (v_1 - v_2)(s) + \alpha^P_s \gamma \|\pi_s\| (\|v_2\| - \|v_1\|) \\
&= \gamma \langle \pi_s, P_{0s}(v_1 - v_2) \rangle + \alpha^P_s \gamma \|\pi_s\| (\|v_2\| - \|v_1\|) \quad [\forall v \in \mathbb{R}^{\mathcal{S}}, P^{\pi}_0 v(s) = \sum_{(s',a) \in \mathcal{X}} \pi_s(a) P_0(s'|s,a) v(s') \\
&\qquad\qquad = \sum_{a \in \mathcal{A}} \pi_s(a) [P_{0s} v](a) = \langle \pi_s, P_{0s} v \rangle] \\
&= \gamma \|\pi_s\| \left(\left\langle \frac{\pi_s}{\|\pi_s\|}, P_{0s}(v_1 - v_2) \right\rangle + \alpha^P_s (\|v_2\| - \|v_1\|) \right) \\
&\leq \gamma \|\pi_s\| \left(\left\langle \frac{\pi_s}{\|\pi_s\|}, P_{0s}(v_1 - v_2) \right\rangle + \alpha^P_s (\|v_2 - v_1\|) \right) \quad [\forall v, w \in \mathbb{R}^{\mathcal{S}}, \|v\| - \|w\| \leq |\|v\| - \|w\|| \leq \|v - w\|].
\end{aligned}
$$

By Asm. 1, we also have

$$
\alpha^P_s \leq \min_{\substack{\mathbf{u}_{\mathcal{A}} \in \mathbb{R}^{\mathcal{A}}_+, \|\mathbf{u}_{\mathcal{A}}\| = 1 \\ \mathbf{v}_{\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}_+, \|\mathbf{v}_{\mathcal{S}}\| = 1}} \mathbf{u}^{\top}_{\mathcal{A}} P_0(\cdot|s,\cdot) \mathbf{v}_{\mathcal{S}} \quad = \min_{\substack{\mathbf{u}_{\mathcal{A}} \in \mathbb{R}^{\mathcal{A}}_+, \|\mathbf{u}_{\mathcal{A}}\| = 1 \\ \mathbf{v}_{\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}_+, \|\mathbf{v}_{\mathcal{S}}\| = 1}} \langle \mathbf{u}_{\mathcal{A}}, P_0(\cdot|s,\cdot) \mathbf{v}_{\mathcal{S}} \rangle \leq \left\langle \frac{\pi_s}{\|\pi_s\|}, P_0(\cdot|s,\cdot) \frac{(v_2 - v_1)}{\|v_2 - v_1\|} \right\rangle,
$$

so that

$$
\begin{aligned}
[T^{\pi,\mathrm{R}^2} v_1 - T^{\pi,\mathrm{R}^2} v_2](s) &\leq \gamma \|\pi_s\| \left(\left\langle \frac{\pi_s}{\|\pi_s\|}, P_{0s}(v_1 - v_2) \right\rangle + \left\langle \frac{\pi_s}{\|\pi_s\|}, P_0(\cdot|s,\cdot) \frac{(v_2 - v_1)}{\|v_2 - v_1\|} \right\rangle \|v_2 - v_1\| \right) \\
&= \gamma \|\pi_s\| \left(\left\langle \frac{\pi_s}{\|\pi_s\|}, P_{0s}(v_1 - v_2) \right\rangle + \left\langle \frac{\pi_s}{\|\pi_s\|}, P_0(\cdot|s,\cdot)(v_2 - v_1) \right\rangle \right) = 0,
\end{aligned}
$$

where we switch notations to designate $P_0(\cdot|s,\cdot) = P_{0s} \in \mathbb{R}^{\mathcal{A} \times \mathcal{S}}$. This proves monotonicity.

Proof of (ii). We now prove the sub-distributivity of the evaluation operator. Let $v \in \mathbb{R}^{\mathcal{S}}, c \in \mathbb{R}$. For all $s \in \mathcal{S}$,

$$
\begin{aligned}
&[T^{\pi,\mathrm{R}^2}(v + c\mathbb{1}_{\mathcal{S}})](s) \\
=&[T^{\pi}_{(P_0,r_0)}(v + c\mathbb{1}_{\mathcal{S}})](s) - \alpha^r_s \|\pi_s\| - \alpha^P_s \gamma \|v + c\mathbb{1}_{\mathcal{S}}\| \|\pi_s\| \\
=&T^{\pi}_{(P_0,r_0)} v(s) + \gamma c - \alpha^r_s \|\pi_s\| - \alpha^P_s \gamma \|v + c\mathbb{1}_{\mathcal{S}}\| \|\pi_s\| \qquad [T^{\pi}_{(P_0,r_0)}(v + c\mathbb{1}_{\mathcal{S}}) = T^{\pi}_{(P_0,r_0)} v + \gamma c \mathbb{1}_{\mathcal{S}}] \\
\leq&T^{\pi}_{(P_0,r_0)} v(s) + \gamma c - \alpha^r_s \|\pi_s\| - \alpha^P_s \gamma \|\pi_s\| (\|v\| + \|c\mathbb{1}_{\mathcal{S}}\|) \\
=&T^{\pi}_{(P_0,r_0)} v(s) + \gamma c - \alpha^r_s \|\pi_s\| - \alpha^P_s \gamma \|\pi_s\| \|v\| \\
&\quad - \alpha^P_s \gamma \|\pi_s\| \|c\mathbb{1}_{\mathcal{S}}\| \\
=&[T^{\pi,\mathrm{R}^2} v](s) + \gamma c - \alpha^P_s \gamma \|\pi_s\| \|c\mathbb{1}_{\mathcal{S}}\| \\
\leq&[T^{\pi,\mathrm{R}^2} v](s) + \gamma c. \qquad [\gamma > 0, \alpha^P_s > 0, \|\cdot\| \geq 0]
\end{aligned}
$$

Proof of (iii). We prove the contraction of a more general evaluation operator with ℓ_p regularization, $p \geq 1$. This will establish contraction of the R^2 operator T^{π,R^2} by simply setting $p = 2$. Define as q the conjugate value of p, *i.e.*, such that $\frac{1}{p} + \frac{1}{q} = 1$. As seen in the proof of Thm. 2, for balls that are constrained according to the ℓ_p-norm $\|\cdot\|_p$, the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of:

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

because $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$, and we can define the R^2 operator accordingly:

$$
[T^{\pi,\mathrm{R}^2}_q v](s) := T^{\pi}_{(P_0,r_0)} v(s) - \alpha^r_s \|\pi_s\|_q - \alpha^P_s \gamma \|v \cdot \pi_s\|_q \quad \forall v \in \mathbb{R}^{\mathcal{S}}, s \in \mathcal{S}\,.
$$

We make the following assumption:

Assumption (A_q). *For all $s \in \mathcal{S}$, there exists $\epsilon_s > 0$ such that $\alpha_s^P \leq \frac{1-\gamma-\epsilon_s}{\gamma|\mathcal{S}|^{\frac{1}{q}}}$.*

Let $v_1, v_2 \in \mathbb{R}^{\mathcal{S}}$. For all $s \in \mathcal{S}$,

$$
\begin{aligned}
&\left|[T_q^{\pi,\mathrm{R}^2} v_1](s) - [T_q^{\pi,\mathrm{R}^2} v_2](s)\right| \\
=&\mid T^{\pi}_{(P_0,r_0)} v_1(s) - \alpha_s^r \|\pi_s\|_q - \alpha_s^P \gamma \|v_1 \cdot \pi_s\|_q \\
&\qquad - (T^{\pi}_{(P_0,r_0)} v_2(s) - \alpha_s^r \|\pi_s\|_q - \alpha_s^P \gamma \|v_2 \cdot \pi_s\|_q) \mid \\
=&\left|T^{\pi}_{(P_0,r_0)} v_1(s) - T^{\pi}_{(P_0,r_0)} v_2(s)\right| + \left|\alpha_s^P \gamma (\|v_2 \cdot \pi_s\|_q - \|v_1 \cdot \pi_s\|_q)\right| \\
=&\left|T^{\pi}_{(P_0,r_0)} v_1(s) - T^{\pi}_{(P_0,r_0)} v_2(s)\right| + \alpha_s^P \gamma \left|\|v_2 \cdot \pi_s\|_q - \|v_1 \cdot \pi_s\|_q\right| \\
\leq&\left|T^{\pi}_{(P_0,r_0)} v_1(s) - T^{\pi}_{(P_0,r_0)} v_2(s)\right| + \alpha_s^P \gamma \|v_2 \cdot \pi_s - v_1 \cdot \pi_s\|_q \\
&\qquad\qquad [\forall\, \mathbf{A}, \mathbf{B} \in \mathbb{R}^{\mathcal{X}}, \left|\|\mathbf{A}\|_q - \|\mathbf{B}\|_q\right| \leq \|\mathbf{A} - \mathbf{B}\|_q] \\
\leq&\gamma \|v_1 - v_2\|_\infty + \alpha_s^P \gamma \|v_2 \cdot \pi_s - v_1 \cdot \pi_s\|_q \\
&\qquad\qquad [\|T^{\pi}_{(P_0,r_0)} v_1 - T^{\pi}_{(P_0,r_0)} v_2\|_\infty \leq \gamma \|v_1 - v_2\|_\infty] \\
=&\gamma \|v_1 - v_2\|_\infty + \alpha_s^P \gamma \|(v_2 - v_1) \cdot \pi_s\|_q \\
&\qquad\qquad [\forall v, w \in \mathbb{R}^{\mathcal{S}}, v \cdot \pi_s - w \cdot \pi_s = (v - w) \cdot \pi_s] \\
\leq&\gamma \|v_1 - v_2\|_\infty + \alpha_s^P \gamma \|v_2 - v_1\|_q \qquad [\forall v \in \mathbb{R}^{\mathcal{S}}, \|v \cdot \pi_s\|_q \leq \|v\|_q] \\
\leq&\gamma \|v_1 - v_2\|_\infty + \alpha_s^P \gamma |\mathcal{S}|^{\frac{1}{q}} \|v_1 - v_2\|_\infty \quad [\forall v, w \in \mathbb{R}^{\mathcal{S}}, \|v - w\|_q \leq |\mathcal{S}|^{\frac{1}{q}} \|v - w\|_\infty] \\
=&\gamma (1 + \alpha_s^P |\mathcal{S}|^{\frac{1}{q}}) \|v_1 - v_2\|_\infty \\
\leq&\gamma \left(1 + \frac{1 - \gamma - \epsilon_s}{\gamma}\right) \|v_1 - v_2\|_\infty \qquad [\alpha_s^P \leq \frac{1 - \gamma - \epsilon_s}{\gamma |\mathcal{S}|^{\frac{1}{q}}} \text{ by Asm. } (\mathrm{A}_q)] \\
=&(1 - \epsilon_s) \|v_1 - v_2\|_\infty \\
\leq&(1 - \epsilon_*) \|v_1 - v_2\|_\infty,
\end{aligned}
$$

where $\epsilon_* := \min_{s \in \mathcal{S}} \epsilon_s$. Setting $q = 2$ and remarking that: (i) the first bound in Asm. 1 recovers Asm. (A_q); (ii) $T_2^{\pi,\mathrm{R}^2} = T^{\pi,\mathrm{R}^2}$, establishes contraction of the R^2 evaluation operator. For the optimality operator, the proof is exactly the same as that of [13, Prop. 3], using Prop. 1. □

C.2 Proof of Theorem 3

Theorem (R^2 optimal policy). *The greedy 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 $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, $v^{\pi^{*,\mathrm{R}^2},\mathrm{R}^2} = v^{*,\mathrm{R}^2} \geq v^{\pi,\mathrm{R}^2}$.*

Proof. By strong convexity of the norm, the R^2 function $\Omega_{v,\mathrm{R}^2} : \pi_s \mapsto \|\pi_s\|(\alpha_s^r + \alpha_s^P \gamma \|v\|)$ is strongly convex in π_s. As such, we can invoke Prop. 1 to state that the greedy policy π^{*,R^2} is the unique maximizing argument for v^{*,R^2}. Moreover, by construction,

$$T^{\pi^{*,\mathrm{R}^2},\mathrm{R}^2} v^{*,\mathrm{R}^2} = T^{*,\mathrm{R}^2} v^{*,\mathrm{R}^2} = v^{*,\mathrm{R}^2}.$$

Supposing that Asm. 1 holds, the evaluation operator $T^{\pi^{*,\mathrm{R}^2},\mathrm{R}^2}$ is contracting and has a unique fixed point $v^{\pi^{*,\mathrm{R}^2},\mathrm{R}^2}$. Therefore, v^{*,R^2} being also a fixed point, we have $v^{\pi^{*,\mathrm{R}^2},\mathrm{R}^2} = v^{*,\mathrm{R}^2}$. It remains to show the last inequality: the proof is exactly the same as that of [13, Thm. 1], and relies on the monotonicity of the R^2 operators. □

C.3 Proof of Remark 1

Remark 2. *An optimal R^2 policy may be stochastic. This is due to the fact that our R^2 MDP framework builds upon the general s-rectangularity assumption. Robust MDPs with s-rectangular uncertainty sets similarly yield an optimal robust policy that is stochastic [47, Table 1]. Nonetheless, the*

R^2 MDP formulation recovers a deterministic optimal policy in the more specific (s,a)-rectangular case, which is in accordance with the robust MDP setting.

Proof. In the (s,a)-rectangular case, the uncertainty set is structured as $\mathcal{U} = \times_{(s,a)\in\mathcal{X}}\mathcal{U}(s,a)$, where $\mathcal{U}(s,a) := P_0(\cdot|s,a) \times r_0(s,a) + \mathcal{P}(s,a) \times \mathcal{R}(s,a)$. The robust counterpart of problem ($\mathrm{P}_\mathcal{U}$) is:

$$\begin{aligned}
F(s) &= \max_{(P,r)\in\mathcal{U}} \{v(s) - r^\pi(s) - \gamma P^\pi v(s)\} \\
&= \max_{(P(\cdot|s,a),r(s,a))\in\mathcal{P}(s,a)\times\mathcal{R}(s,a)} \{v(s) - r_0^\pi(s) - r^\pi(s) - \gamma P_0^\pi v(s) - \gamma P^\pi v(s)\} \\
&= \max_{(P(\cdot|s,a),r(s,a))\in\mathcal{P}(s,a)\times\mathcal{R}(s,a)} \{-r^\pi(s) - \gamma P^\pi v(s)\} + v(s) - r_0^\pi(s) - \gamma P_0^\pi v(s) \\
&= \max_{r(s,a)\in\mathcal{R}(s,a)} \{-r^\pi(s)\} + \gamma \max_{P(\cdot|s,a)\in\mathcal{P}(s,a)} \{-P^\pi v(s)\} + v(s) - T^\pi_{(P_0,r_0)}v(s) \\
&= \max_{r(s,a)\in\mathcal{R}(s,a)} \left\{ -\sum_{a\in\mathcal{A}} \pi_s(a) r(s,a) \right\} + \gamma \max_{P(\cdot|s,a)\in\mathcal{P}(s,a)} \left\{ -\sum_{a\in\mathcal{A}} \pi_s(a) \langle P(\cdot|s,a), v\rangle \right\} \\
&\quad + v(s) - T^\pi_{(P_0,r_0)}v(s) \\
&= \sum_{a\in\mathcal{A}} \pi_s(a) \left(\max_{r(s,a)\in\mathcal{R}(s,a)} \{-r(s,a)\} + \gamma \max_{P(\cdot|s,a)\in\mathcal{P}(s,a)} \{\langle P(\cdot|s,a), -v\rangle\} \right) \\
&\quad + v(s) - T^\pi_{(P_0,r_0)}v(s).
\end{aligned}$$

In particular, if we have ball uncertainty sets $\mathcal{P}(s,a) := \{P(\cdot|s,a) \in \mathbb{R}^\mathcal{S} : \|P(\cdot|s,a)\| \le \alpha^P_{s,a}\}$ and $\mathcal{R}(s,a) := \{r(s,a) \in \mathbb{R} : |r(s,a)| \le \alpha^r_{s,a}\}$ for all $(s,a) \in \mathcal{X}$, then we can explicitly compute the support functions:

$$\max_{r(s,a):|r(s,a)|\le\alpha^r_{s,a}} -r(s,a) = \alpha^r_{s,a} \text{ and } \max_{P(\cdot|s,a):\|P(\cdot|s,a)\|\le\alpha^P_{s,a}} \langle P(\cdot|s,a), -v\rangle = \alpha^P_{s,a}\|v\|.$$

Therefore, the robust counterpart rewrites as:

$$F(s) = \sum_{a\in\mathcal{A}} \pi_s(a)(\alpha^r_{s,a} + \gamma\alpha^P_{s,a}\|v\|) + v(s) - T^\pi_{(P_0,r_0)}v(s),$$

and the robust value function $v^{\pi,\mathcal{U}}$ is the optimal solution of the convex optimization problem:

$$\max_{v\in\mathbb{R}^\mathcal{S}} \langle v, \mu_0\rangle \text{ s. t. } v(s) \le T^\pi_{(P_0,r_0)}v(s) - \sum_{a\in\mathcal{A}} \pi_s(a)(\alpha^r_{s,a} + \gamma\alpha^P_{s,a}\|v\|) \text{ for all } s \in \mathcal{S}\,.$$

This derivation enables us to derive an R^2 Bellman evaluation operator for the (s,a)-rectangular case. Indeed, the R^2 regularization function now becomes

$$\Omega_{v,\mathrm{R}^2}(\pi_s) := \sum_{a\in\mathcal{A}} \pi_s(a)(\alpha^r_{s,a} + \gamma\alpha^P_{s,a}\|v\|),$$

which yields the following R^2 operator:

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

We aim to show that we can find a deterministic policy $\pi^d \in \Delta^\mathcal{S}_\mathcal{A}$ such that $[T^{\pi^d,\mathrm{R}^2}v](s) = [T^{*,\mathrm{R}^2}v](s)$ for all $s \in \mathcal{S}$. Given an arbitrary policy $\pi \in \Delta^\mathcal{S}_\mathcal{A}$, we first rewrite:

$$\begin{aligned}
[T^{\pi,\mathrm{R}^2}v](s) &= r_0^\pi(s) + \gamma P_0^\pi v(s) - \Omega_{v,\mathrm{R}^2}(\pi_s) \\
&= \sum_{a\in\mathcal{A}} \pi_s(a) r_0(s,a) + \gamma \sum_{a\in\mathcal{A}} \pi_s(a) \langle P_0(\cdot|s,a), v\rangle - \left(\sum_{a\in\mathcal{A}} \pi_s(a)(\alpha^r_{s,a} + \gamma\alpha^P_{s,a}\|v\|) \right) \\
&= \sum_{a\in\mathcal{A}} \pi_s(a) \Big(r_0(s,a) - \alpha^r_{s,a} + \gamma(\langle P_0(\cdot|s,a), v\rangle - \alpha^P_{s,a}\|v\|) \Big)
\end{aligned}$$

By [36, Lemma 4.3.1], we have that:

$$\sum_{a\in\mathcal{A}} \pi_s(a)\Big(r_0(s,a) - \alpha^r_{s,a} + \gamma(\langle P_0(\cdot|s,a), v\rangle - \alpha^P_{s,a}\|v\|)\Big)$$
$$\leq \max_{a\in\mathcal{A}}\Big\{r_0(s,a) - \alpha^r_{s,a} + \gamma(\langle P_0(\cdot|s,a), v\rangle - \alpha^P_{s,a}\|v\|)\Big\},$$

and since the action set is finite, there exists an action $a^* \in \mathcal{A}$ reaching the maximum. Setting $\pi^d(a^*) = 1$ thus gives the desired result. We just derived a regularized formulation of robust MDPs with (s,a)-rectangular uncertainty set and ensured that the corresponding R^2 Bellman operators yield a deterministic optimal policy. In that case, the optimal R^2 Bellman operator becomes:

$$[T^{*,\mathrm{R}^2}v](s) = \max_{a\in\mathcal{A}}\Big\{r_0(s,a) - \alpha^r_{s,a} + \gamma(\langle P_0(\cdot|s,a), v\rangle - \alpha^P_{s,a}\|v\|)\Big\}.$$

□

D R^2 q-learning

D.1 The R^2 q-function

Theorem 4. *Assume that $\mathcal{U} = (\{P_0\} + \mathcal{P}) \times (\{r_0\} + \mathcal{R})$ and $\mathcal{U}$ is (s,a)-rectangular. Then, its corresponding robust action-value function is an optimal solution of:*

$$\max_{q\in\mathbb{R}^{\mathcal{X}}} \langle q, \mu_0\cdot\pi\rangle \text{ s.t. } q(s,a) \leq T^{\pi}_{(P_0,r_0)}q(s,a) - \sigma_{\mathcal{R}(s,a)}(-1) - \sigma_{\mathcal{P}(s,a)}(-\gamma q\cdot\pi) \text{ for all } (s,a)\in\mathcal{X}, \tag{10}$$

where $[q\cdot\pi](s') := \sum_{a'\in\mathcal{A}}\pi_{s'}(a')q(s',a'), \forall s'\in\mathcal{S}$.

Proof. It is known from [24] that the robust action-value function is an optimal solution of:

$$\max_{q\in\mathbb{R}^{\mathcal{X}}} \langle q, \mu_0\cdot\pi\rangle \text{ s.t. } q(s,a) \leq T^{\pi}_{(P,r)}q(s,a) \text{ for all } (s,a)\in\mathcal{X}, (P(\cdot|s,a), r(s,a))\in\mathcal{U}(s,a),$$

which can be rewritten as:

$$\max_{q\in\mathbb{R}^{\mathcal{X}}} \langle q, \mu_0\cdot\pi\rangle \text{ s.t. } q(s,a) \leq T^{\pi}_{(P_0,r_0)}q(s,a) + r(s,a) + \gamma\langle P(\cdot|s,a)\cdot\pi, q\rangle$$
$$\text{for all } (s,a)\in\mathcal{X}, (P(\cdot|s,a), r(s,a))\in\mathcal{U}(s,a),$$

with $[P(\cdot|s,a)\cdot\pi](s',a') := \pi_{s'}(a')P(s'|s,a), \forall (s',a'\in\mathcal{X})$. More synthetically, the robust action-value function is an optimal solution of:

$$\begin{aligned}&\max_{q\in\mathbb{R}^{\mathcal{X}}} \langle q, \mu_0\cdot\pi\rangle \\ &\text{s.t.} \max_{(P(\cdot|s,a),r(s,a))\in\mathcal{U}(s,a)} \Big\{q(s,a) - T^{\pi}_{(P_0,r_0)}q(s,a) - r(s,a) - \gamma\langle P(\cdot|s,a)\cdot\pi, q\rangle\Big\} \leq 0 \\ &\text{for all } (s,a)\in\mathcal{X}.\end{aligned} \tag{11}$$

We now compute the robust counterpart. For any $(s,a)\in\mathcal{X}$ and policy $\pi\in\Delta^{\mathcal{S}}_{\mathcal{A}}$, denote by:

$$F^{\pi}(s,a) := \max_{(P(\cdot|s,a),r(s,a))\in\mathcal{U}(s,a)} \Big\{q(s,a) - T^{\pi}_{(P_0,r_0)}q(s,a) - r(s,a) - \gamma\langle P(\cdot|s,a)\cdot\pi, q\rangle\Big\}.$$

Removing the constant terms from the maximization and using the indicator function yields:

$$\begin{aligned}F^{\pi}(s,a) &= q(s,a) - T^{\pi}_{(P_0,r_0)}q(s,a) + \max_{(P(\cdot|s,a),r(s,a))\in\mathcal{U}(s,a)} \{-r(s,a) - \gamma\langle P(\cdot|s,a)\cdot\pi, q\rangle\} \\ &= q(s,a) - T^{\pi}_{(P_0,r_0)}q(s,a) - \min_{(P(\cdot|s,a),r(s,a))\in\mathcal{U}(s,a)} \{r(s,a) + \gamma\langle P(\cdot|s,a)\cdot\pi, q\rangle\} \\ &= q(s,a) - T^{\pi}_{(P_0,r_0)}q(s,a) - \min_{r(s,a)\in\mathcal{R}(s,a)} r(s,a) - \min_{P(\cdot|s,a)\in\mathcal{P}(s,a)} \gamma\langle P(\cdot|s,a)\cdot\pi, q\rangle \\ &= q(s,a) - T^{\pi}_{(P_0,r_0)}q(s,a) - \min_{r(s,a)\in\mathbb{R}} \{r(s,a) + \delta_{\mathcal{R}(s,a)}(r(s,a))\} \\ &\quad - \min_{P(\cdot|s,a)\in\mathbb{R}^{\mathcal{S}}} \{\gamma\langle P(\cdot|s,a)\cdot\pi, q\rangle + \delta_{\mathcal{P}(s,a)}(P(\cdot|s,a))\}.\end{aligned}$$

Applying Fenchel-Rockafellar duality to both minimization problems yields the desired result. □

Corollary 2. *If, additionally, $\mathcal{P}_{sa}$ is a ball of radius α^P_{sa} w.r.t. some norm $\|\cdot\|$ and $\mathcal{R}_{sa}$ an interval of radius α^r_{sa}, then the robust action-value function is an optimal solution of:*

$$\max_{q \in \mathbb{R}^{\mathcal{X}}} \langle q, \mu_0 \cdot \pi \rangle \text{ s.t. } q(s,a) \leq T^{\pi}_{(P_0,r_0)} q(s,a) - \alpha^r_{sa} - \gamma \alpha^P_{sa} \|q \cdot \pi\|_* \text{ for all } (s,a) \in \mathcal{X} \,. \tag{12}$$

The upper-bound in the optimization problem enables to define the R^2 Bellman operator on q-functions as:

$$[T^{\pi,\mathrm{R}^2} q](s,a) := T^{\pi}_{(P_0,r_0)} q(s,a) - \alpha^r_{sa} - \gamma \alpha^P_{sa} \|q \cdot \pi\|_*$$

D.2 Distinguishing between R^2 and robust q-functions

We aim to show that although we can interchangeably optimize an R^2 q-function or a robust q-value under (s,a)-rectangularity, the R^2 q-function obtained from the R^2 value v is *not* the same as the q-function obtained from the original robust optimization problem. This nuance is reminiscent of the regularized MDP setting, where defining the regularized q-function w.r.t. the regularized value v is not equivalent to taking v as the expected q-function over a policy.

Let thus assume that the uncertainty set is (s,a)-rectangular. Then, by Sec. C.3, the R^2 value function v^{π,R^2} is the unique fixed point of the R^2 Bellman operator as below:

$$[T^{\pi,\mathrm{R}^2} v](s) := T^{\pi}_{(P_0,r_0)} v(s) - \sum_{a \in \mathcal{A}} \pi_s(a)(\alpha^r_{s,a} + \gamma \alpha^P_{s,a} \|v\|), \quad \forall s \in \mathcal{S} \,.$$

This rewrites as:

$$\begin{aligned} v^{\pi,\mathrm{R}^2}(s) &= \sum_{a \in \mathcal{A}} \pi_s(a) \left(r_0(s,a) + \gamma \langle P_0(\cdot|s,a), v^{\pi,\mathrm{R}^2} \rangle - \alpha^r_{s,a} - \gamma \alpha^P_{s,a} \|v^{\pi,\mathrm{R}^2}\| \right) \\ &= \sum_{a \in \mathcal{A}} \pi_s(a) \left(q^{\pi,\mathrm{R}^2}(s,a) - \alpha^r_{s,a} - \gamma \alpha^P_{s,a} \|v^{\pi,\mathrm{R}^2}\| \right), \end{aligned}$$

where the last equality holds by definition of the q-function associated with v^{π,R^2} (Def. 2). As a result,

$$q^{\pi,\mathrm{R}^2} \cdot \pi(s) = v^{\pi,\mathrm{R}^2}(s) + [\alpha^r + \gamma \alpha^P \|v^{\pi,\mathrm{R}^2}\|] \cdot \pi(s)$$

Alternatively, by optimizing w.r.t. $q \in \mathbb{R}^{\mathcal{X}}$ instead of $v \in \mathbb{R}^{\mathcal{S}}$ and applying Cor. 2, the robust action-value function $q^{\pi,\mathcal{U}}$ satisfies:

$$q^{\pi,\mathcal{U}}(s,a) = T^{\pi}_{(P_0,r_0)} q^{\pi,\mathcal{U}}(s,a) - \alpha^r_{sa} - \gamma \alpha^P_{sa} \|q^{\pi,\mathcal{U}} \cdot \pi\|_* \text{ for all } (s,a) \in \mathcal{X} \,.$$

Taking the expectation over policy π yields:

$$q^{\pi,\mathcal{U}} \cdot \pi = r^{\pi}_0 + P^{\pi}_0 (q^{\pi,\mathcal{U}} \cdot \pi) - \sum_{a \in \mathcal{A}} \pi_s(a)(\alpha^r_{sa} - \gamma \alpha^P_{sa} \|q^{\pi,\mathcal{U}} \cdot \pi\|_*) \text{ for all } (s,a) \in \mathcal{X},$$

so that $q^{\pi,\mathcal{U}} \cdot \pi$ is a fixed point of the R^2 Bellman operator. By unicity of its fixed point, we obtain that $q^{\pi,\mathcal{U}} \cdot \pi = v^{\pi,\mathrm{R}^2}$. As a result:

$$\begin{aligned} q^{\pi,\mathcal{U}} \cdot \pi &= v^{\pi,\mathrm{R}^2} \\ &= (q^{\pi,\mathrm{R}^2} - \alpha^r - \gamma \|v^{\pi,\mathrm{R}^2}\| \alpha^P) \cdot \pi \\ &= q^{\pi,\mathrm{R}^2} \cdot \pi - [\alpha^r + \gamma \alpha^P \|v^{\pi,\mathrm{R}^2}\|] \cdot \pi \end{aligned}$$

Taking deterministic policies on each possible action, we end up with an element-wise identity:

$$q^{\pi,\mathcal{U}}(s,a) = q^{\pi,\mathrm{R}^2}(s,a) - \alpha^r_{s,a} - \gamma \|v^{\pi,\mathrm{R}^2}\| \alpha^P_{s,a}$$

D.3 Convergence of R^2 q-learning

Theorem 5 (Convergence of R^2 q-learning)**.** *For any $(s,a) \in \mathcal{X}$, let a sequence of step-sizes $(\beta_t(s,a))_{t\in\mathbb{N}}$ satisfying $0 \leq \beta_t(s,a) \leq 1$, $\sum_{t\in\mathbb{N}} \beta_t(s,a) = \infty$ and $\sum_{t\in\mathbb{N}} \beta_t^2(s,a) < \infty$. Then, the R^2 q-learning algorithm as given in Alg. 2 converges almost surely to the optimal R^2 q-function.*

Proof. We will use the convergence result from [25]. The update rule is given by:

$$\begin{aligned} q_{t+1}(s_t,a_t) = q_t(s_t,a_t) \\ + \beta_t(s_t,a_t)\left(r_{t+1} + \gamma \max_{b\in\mathcal{A}} q_t(s_{t+1},b) - \alpha^r_{s_t a_t} - \gamma\alpha^P_{s_t a_t}\|\max_{b\in\mathcal{A}} q_t(\cdot,b)\|_* - q_t(s_t,a_t)\right) \end{aligned}$$

which we rewrite as:

$$\begin{aligned} q_{t+1}(s_t,a_t) = (1-\beta_t(s_t,a_t))q_t(s_t,a_t) \\ + \beta_t(s_t,a_t)\left(r_{t+1} + \gamma \max_{b\in\mathcal{A}} q_t(s_{t+1},b) - \alpha^r_{s_t a_t} - \gamma\alpha^P_{s_t a_t}\|\max_{b\in\mathcal{A}} q_t(\cdot,b)\|_*\right). \end{aligned} \tag{13}$$

Further let $\Delta_t(s,a) := q_t(s,a) - q^{*,\mathrm{R}^2}(s,a), \forall (s,a) \in \mathcal{X}$. Then Eq. (13) rewrites as:

$$\begin{aligned} \Delta_{t+1}(s,a) = (1-\beta_t(s_t,a_t))\Delta_t(s_t,a_t) \\ + \beta_t(s_t,a_t)\left(r_{t+1} + \gamma \max_{b\in\mathcal{A}} q_t(s_{t+1},b) - \alpha^r_{s_t a_t} - \gamma\alpha^P_{s_t a_t}\|\max_{b\in\mathcal{A}} q_t(\cdot,b)\|_* - q^{*,\mathrm{R}^2}(s_t,a_t)\right). \end{aligned}$$

We introduce the following random variable:

$$G_t(s,a) := r(s,a) + \gamma \max_{b\in\mathcal{A}} q_t(X(s,a),b) - \alpha^r_{sa} - \gamma\alpha^P_{sa}\|\max_{b\in\mathcal{A}} q_t(\cdot,b)\|_* - q^{*,\mathrm{R}^2}(s,a),$$

so that

$$\begin{aligned} \mathbb{E}\left[G_t(s,a)|\mathcal{F}_t\right] &= \mathbb{E}\left[r(s,a) + \gamma \max_{b\in\mathcal{A}} q_t(X(s,a),b) - \alpha^r_{sa} - \gamma\alpha^P_{sa}\|\max_{b\in\mathcal{A}} q_t(\cdot,b)\|_* - q^{*,\mathrm{R}^2}(s,a)|\mathcal{F}_t\right] \\ &= r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \max_{b\in\mathcal{A}} P(s'|s,a) q_t(s',b) - \alpha^r_{sa} - \gamma\alpha^P_{sa}\|\max_{b\in\mathcal{A}} q_t(\cdot,b)\|_* - q^{*,\mathrm{R}^2}(s,a) \\ &= [T^{*,\mathrm{R}^2} q_t](s,a) - q^{*,\mathrm{R}^2}(s,a) \\ &= [T^{*,\mathrm{R}^2} q_t](s,a) - [T^{*,\mathrm{R}^2} q^{*,\mathrm{R}^2}](s,a). \end{aligned}$$

By contraction property of the R^2 Bellman operator, we thus obtain:

$$\begin{aligned} \|\mathbb{E}\left[G_t(s,a)|\mathcal{F}_t\right]\|_\infty &= \left\|[T^{*,\mathrm{R}^2} q_t](s,a) - [T^{*,\mathrm{R}^2} q^{*,\mathrm{R}^2}](s,a)\right\|_\infty \\ &\leq (1-\epsilon^*)\left\|q_t(s,a) - q^{*,\mathrm{R}^2}(s,a)\right\|_\infty = (1-\epsilon^*)\|\Delta_t(s,a)\|_\infty \end{aligned}$$

□

E Planning on a Maze

Number of seeds per experiment	5
Discount factor γ	0.9
Convergence Threshold θ	1e-3
Reward Radius α^r	1e-3
Transition Radius α^P	1e-5

Table 4: Hyperparameter set to obtain the results from Table 1

In the following experiment, we play with the radius of the uncertainty set and analyze the distance of the robust/R^2 value function to the vanilla one obtained after convergence of MPI. Except for the radius parameters of Table 4, all other parameters remain unchanged. In both figures 3 and 4, we see that the distance norm converges to 0 as the size of the uncertainty set gets closer to 0: this sanity check ensures an increasing relationship between the level of robustness and the radius value. As shown in Fig. 3, the plots for robust MPI and R^2 MPI coincide in the reward-robust case, but they diverge from each other as the transition model gets more uncertain. This does not contradict our theoretical findings from Thms. 1-2. In fact, each iteration of robust MPI involves an optimization problem which is solved using a black-box solver and yields an approximate solution. As such, errors propagate across iterations and according to Fig. 4, they are more sensitive to transition than reward uncertainty. This is easy to understand: as opposed to the reward function, the transition kernel interacts with the value function at each Bellman update, so errors on the value function also affect those on the optimum and vice versa. Moreover, the gap grows with the radius level because of the simplex constraint we ignored when computing the support function of the transition uncertainty set. The work [29] accounts for this additional constraint to derive a regularization function that recovers the robust value under transition uncertainty.

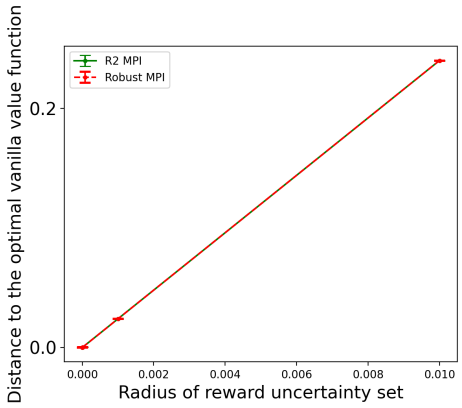


Figure 3: Distance norm between the optimal robust/R^2 value and the vanilla one as a function of α^r ($\alpha^P = 0$) after 5 runs of robust/R^2 MPI

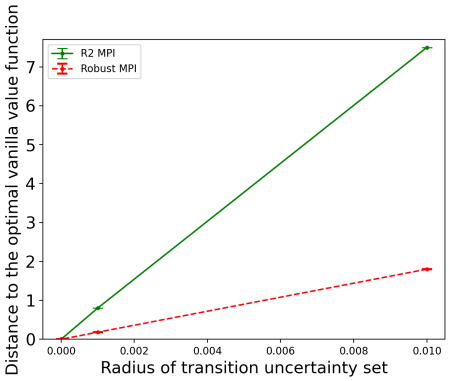


Figure 4: Distance norm between the optimal robust/R^2 value and the vanilla one as a function of α^P ($\alpha^r = 0$) after 5 runs of robust/R^2 MPI

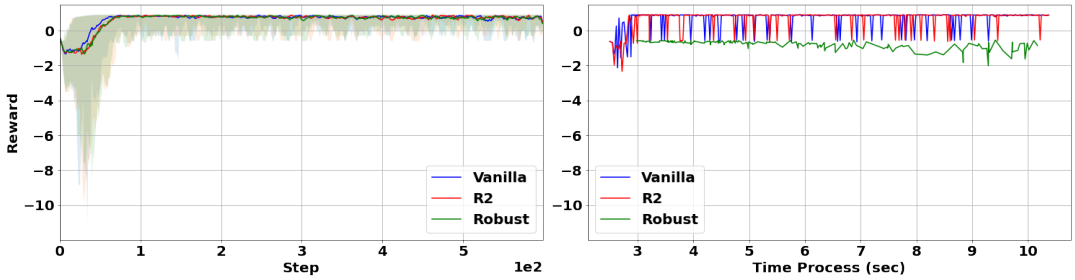
F R^2 Learning Experiments

In this section, we provide additional details and experiments regarding our R^2 q-learning algorithm.

F.1 Tabular case

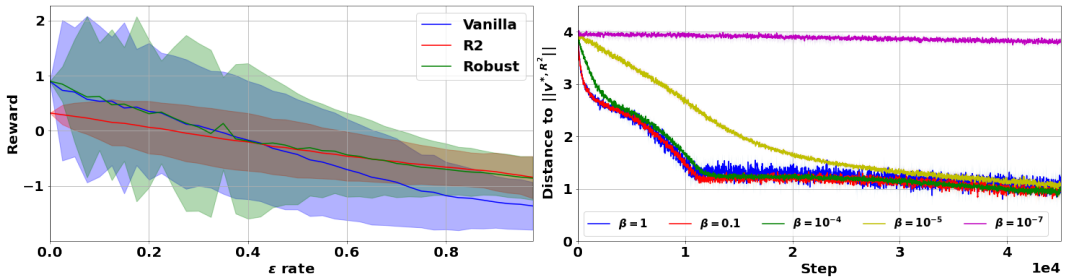
As proof of concept, we perform experiments in a tabular environment. Here, our goals are the following: (i) numerically illustrate the convergence of R^2 q-learning; (ii) highlight its computational advantage over robust q-learning concurrently with its robustness properties.

We consider a Mars Rover domain as in [43]. The objective is to find the shortest path to a goal state in a 10×10 grid. However, taking a shorter path implies higher risk: the rover has a greater chance to hit a rocket and get a negative reward. The transition function is stochastic: the agent moves to the chosen direction with probability $1 - \epsilon$, and randomly otherwise. At each step, it receives a small penalty r_{step}. An episode terminates whenever the rover reaches the goal state or hits a rock. The two scenarios yield a reward of r_{success} and r_{fail} respectively. We thus have $r_{\text{success}} > 0 > r_{\text{step}} > r_{\text{fail}}$. We compare our R^2 q-learning algorithm with two baselines: *vanilla* and *robust* q-learning. Vanilla is the standard method that ignores model uncertainty and assumes the reward and dynamics are fixed. Robust q-learning trains a robust optimal policy using robust Bellman updates, thus requiring solving an optimization problem at each iteration [37].



(a) (b)

Figure 5: Convergence plots for Mars Rover. (a) Cumulative reward w.r.t. the number of iteration steps, averaged over 10 seeds. For R^2 and robust q-learning, $\alpha_p = \alpha_r = 0.01$. (b) Cumulative reward w.r.t. time process in seconds. Performance peaks appear because data are sometimes logged in the middle of an episode, so the agent has accumulated negative rewards.



(a) (b)

Figure 6: Mars Rover: (a) Evaluation of q-learning over new transition models. Each algorithm was trained over 10 seeds on nominal $\epsilon = 0$. (b) Comparison of different β-values for moving average. Each β-value is run over 5 seeds (these are the same for the exact and the estimated case).

Fig. 5a shows the convergence plot across iteration steps for the three agents: vanilla, robust and R^2. All of them have similar sample complexity and fulfill the task within 100 iteration steps. The difference between them arises when we look at the time complexity of each algorithm. As we can see in Fig. 5b, robust q-learning takes more than 2 minutes to converge, whereas vanilla and R^2 q-learning achieve the highest reward within 4 seconds (see also Fig. 7). Similarly, we calculated the average time necessary to perform one learning step in each algorithm: one R^2 update takes $7.7 \pm 5.9 \times 10^{-6}$ seconds to run, which is slightly slower than vanilla with $1.24 \pm 0.89 \times 10^{-6}$ seconds. On the other hand, a robust q-update takes $3 \pm 0.9 \times 10^{-2}$ seconds, thus representing 10^4 higher cost than the other two approaches. This highlights the clear advantage of R^2 over robust q-learning in terms of computational cost. To check robustness, after training, we evaluate each policy under varying dynamics. In particular, we increase the value of ϵ to make the environment more adversarial. Fig. 6a shows that the R^2 policy performs similarly to the robust one under more adversarial transitions *i.e.*, when ϵ tends to 1, both being less sensitive than vanilla.

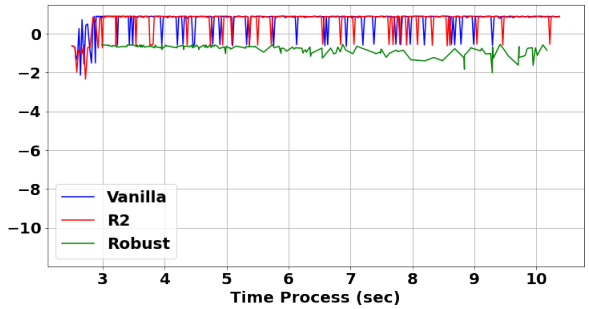


Figure 7: Mars Rover: Cumulative reward w.r.t. time process in seconds (zoom in)

Parameter	Value
random rate ϵ	0

(a) Mars Rover parameters

Parameter	Value
gravity	9.8
masscart	1.0
masspole	0.1
length	0.5
force_mag	10.0

(b) Cartpole

Parameter	Value
link_length_2	1.0
link_mass_1	1.0
link_mass_2	1.0
link_com_pos_1	1.0
link_com_pos_2	1.0
link_moi	1.0
link_length_1	1.0

(c) Acrobot

Parameter	Value
force	0.001
gravity	0.0025

(d) Mountaincar

Table 5: Nominal environment parameters on which all algorithms have been trained