Regularized Policies are Reward Robust

Hisham Husain
The Australian National University
CSIRO Data61

Kamil Ciosek*
Spotify Research

Ryota Tomioka
Microsoft Research Cambridge

Abstract

Entropic regularization of policies in Reinforcement Learning (RL) is a commonly used heuristic to ensure that the learned policy explores the state-space sufficiently before overfitting to a local optimal policy. The primary motivation for using entropy is for exploration and disambiguating optimal policies; however, the theoretical effects are not entirely understood. In this work, we study the more general regularized RL objective and using Fenchel duality; we derive the dual problem which takes the form of an adversarial reward problem. In particular, we find that the optimal policy found by a regularized objective is precisely an optimal policy of a reinforcement learning problem under a worst-case adversarial reward. Our result allows us to reinterpret the popular entropic regularization scheme as a form of robustification. Furthermore, due to the generality of our results, we apply to other existing regularization schemes. Our results thus give insights into the effects of regularization of policies and deepen our understanding of exploration through robust rewards at large.

1 Introduction

Reinforcement Learning (RL) is a paradigm of algorithms which learn policies that maximize the expected discounted reward specified by a Markov Decision Process (MDP) (Sutton and Barto, 2018). The formulation of an MDP is well-posed with links in utility theory (Russell and Norvig, 2002) and specifies a reward function where the solution can be found precisely in a deterministic form. However, in practice, the reward function is typically an idealization, and it turns out that an optimal policy in this model will cope terribly when presented to unseen or uncertain situations. Intuitively, it is anticipated that there exist multiple policies that are near-optimal to this reward yet exhibit more robust and diversified behaviour. In particular, having multiple solutions of this form would be preferred since they can help the practitioner in understanding the environment and problem better.

Finding near-optimal policies in this sense requires balancing between ensuring that the policy is optimal for the given reward and demonstrates some form of robustness or diversity. This is commonly recollected as the *exploration* vs *exploitation* trade-off[1]. One of the most effective ways in ensuring this balance is by altering the objective of the MDP to include a form of penalty so that the resulting policy reflects characteristics of diversified behaviour. Causal entropy (Ziebart, 2010) is a popular example of this, where the policy is penalized for being deterministic in favour of exploration and disambiguating optimal policies. This has lead to the MaxEnt framework (Haarnoja et al., 2018c) and shown compelling relations to probabilistic inference (Dayan and Hinton, 1997; Neumann et al., 2011; Todorov, 2007; Kappen, 2005; Toussaint, 2009; Rawlik et al., 2013; Theodorou et al., 2010; Ziebart, 2010) whilst maintaining empirically superior performance on several tasks (Haarnoja et al., 2018c,b), including robustness in the face of uncertainty (Haarnoja et al., 2018a). In the case where the reward function is not specified, the entropy alone as an objective is also prevalent to ensure exploration (Hazan et al., 2019). Similar forms of regularization have appeared in Wu et al. (2019), which ensure that the policy is stabilized in accordance with a pre-determined behaviour and other forms of diversifying schemes using policy regularization have been developed in (Hong et al., 2018). Furthermore, the benefits of regularizers have also been observed in adversarial imitation learning

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

(*) Work done while at Microsoft Research Cambridge.

[1] Traditionally, this refers to the sequential behavior where one is interested in finding better policies at each timestep.

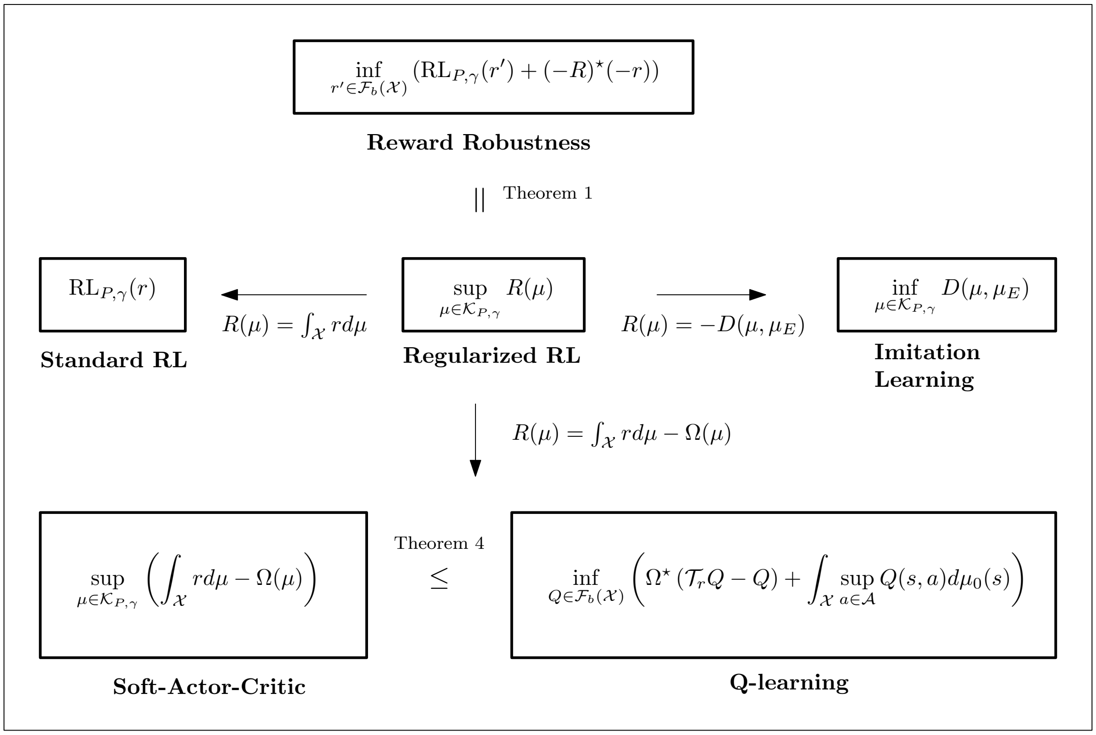
$\inf_{r' \in \mathcal{F}_b(\mathcal{X})} (\mathrm{RL}_{P,\gamma}(r') + (-R)^\star(-r))$
Reward Robustness
|| Theorem 1
$\mathrm{RL}_{P,\gamma}(r)$
$R(\mu) = \int_{\mathcal{X}} r d\mu$
Standard RL
$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu)$
Regularized RL
$R(\mu) = -D(\mu, \mu_E)$
$\inf_{\mu \in \mathcal{K}_{P,\gamma}} D(\mu, \mu_E)$
Imitation Learning
$R(\mu) = \int_{\mathcal{X}} r d\mu - \Omega(\mu)$
$\sup_{\mu \in \mathcal{K}_{P,\gamma}} \left(\int_{\mathcal{X}} r d\mu - \Omega(\mu) \right)$
Soft-Actor-Critic
Theorem 4
$\leq$
$\inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\Omega^\star (\mathcal{T}_r Q - Q) + \int_{\mathcal{X}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) \right)$
Q-learning

Figure 1: Our main is to provide a unified view of existing objectives in Reinforcement Learning and relate them to a reward robustness problem as highlighted above through Theorem 1. Additionally, we show another link between regularized policies and Q-learning in Theorem 4.

methods (Ho and Ermon, 2016; Li et al., 2017).

While the empirical success should rejoice, it is somewhat unsettling that changing the objective deviates from the MDP set-up, which was initially motivated through the axioms of utility theory (Russell and Norvig, 2002). In particular, it is not clear what kind of policy these regularized objectives are learning from the perspective of the original reward maximization problems, especially since it is apparent that regularized policies pose successfully in these schemes. On this front, there exists work that shows entropic regularization smoothens the optimization landscape (Ahmed et al., 2019) and induces sparse policies when considering a larger class of policy regularizers (Yang et al., 2019). While these works advocate the effects of policy regularization, the benefits of regularization from an accuracy or robustness perspective and not very well understood. This is especially relevant since in machine learning more generally, regularization has shown strong links to generalization and robustness (Duchi et al., 2016; Sinha et al., 2017; Husain, 2020). The first attempt is (Eysenbach and Levine, 2019), which shows that MaxEnt performs explicitly well on a robust reward problem. This approach however, is limited to only the MaxEnt and cannot apply to other schemes such as regularized imitation learning.

In this work, we tackle this precisely and focus on the problem specified by finding a policy that maximizes an objective R that is concave in the space of state-action visitation distributions. This objective includes the standard reward objective and subsumes other popular objectives such as the MaxEnt framework and imitation learning. Our main insight is that the policy learned using a concave objective R is *robust* against rewards chosen by an adversary, where R determines the nature of the adversary. We find that the policy is precisely a maximizer against the worst-case reward r'. Moreover, we characterize the analytic form of r' (using a technical assumption on R), which delivers more insight onto the nature of robustness. Our results thus allow us to reinterpret entropic regularization and exploration more generally as a robustifying mechanism and add to the advocation for using such methods in practice. In summary, our contributions are

1. A duality result linking generalized RL objectives as adversarial reward problems[2], which allows us to reinterpret the extant MaxEnt framework, among others, as a robustifying mechanism.

[2] We remark that this is not the same as conventional adversarial training, as found in supervised learning.

2. Characterization of the adversarial reward solved by these regularized policy objectives. In doing so, we derive a generalized value function interpretation of entropic regularization.

3. A primal-dual link between the regularized policy objective and Q-learning loss. This allows us to reinterpret the mean-squared error Q-learning as a form regularization of policies and robustification against rewards in light of our main result.

4. Deriving the robust-reward problem for other popular frameworks such as imitation learning and model-free entropic optimization. This allows us to compare and unify these separate problems under reward-robustness. We illustrate this diagrammatically in Figure 1.

2 Preliminaries

Reinforcement Learning We use a compact set $\mathcal{S}$ to denote the state space, $\mathcal{A}$ the action space and set $\mathcal{X} = \mathcal{S} \times \mathcal{A}$. We assume these spaces are Polish and furthermore use $\mathscr{P}(\mathcal{S})$, $\mathscr{P}(\mathcal{A})$ and $\mathscr{P}(\mathcal{X})$ to denote the set of Borel probability measures. Similarly, we use $\mathcal{F}_b(\mathcal{S})$, $\mathcal{F}_b(\mathcal{A})$ and $\mathcal{F}_b(\mathcal{X})$ to denote the set of bounded and measurable functions on the sets $\mathcal{S}, \mathcal{A}$ and $\mathcal{X}$ respectively. A reward function is a mapping $r : \mathcal{X} \to \mathbb{R}$, a transition kernel is specified as $P : \mathcal{X} \to \mathscr{P}(\mathcal{S})$ and a policy is a mapping $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$. Let $\gamma > 0$ be an implicit fixed discount parameter. It can be shown that each $\mathcal{S}$, $\mathcal{A}$, P, initial distribution μ_0 and policy π uniquely define a Markov chain $\{(S_t, A_t)\}_{t=1}^{\infty} \subseteq \mathcal{X}$. We denote the underlying probability space as $(\mathcal{X}, \mathscr{T}, P_{\mu_0,\pi})$ where $P_{\mu_0,\pi} \in \mathscr{P}(\mathcal{X})$ is referred to as the state-action visitation distribution. We refer the reader to (Meyn and Tweedie, 2012, Chapter 3) and (Revuz, 2008, Chapter 2) for more detailed constructions. The goal in RL is to find a policy that maximizes expected return over the state-action pairs visited, which can be concretely summarized in the optimization problem:

$$\sup_{\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})} \mathbb{E}_{P_{\mu_0,\pi}(s,a)} \left[r(s,a) \right]. \tag{1}$$

This objective is linear in the space of state-action visitation distributions and thus is equivalent to the linear program $\max_{\mu \in \mathcal{K}_{P,\gamma}} \int_{\mathcal{X}} r(s,a) d\mu(s,a)$ where

$$\mathcal{K}_{P,\gamma} = \Bigg\{ \mu \in \mathscr{P}(\mathcal{X}) : \int_{\mathcal{A}} \mu(s,a) da = (1-\gamma)\mu_0(s) + \gamma \int_{\mathcal{X}} P(s \mid s', a') d\mu(s', a') \Bigg\}.$$

In particular, for any policy π, we have that $P_{\mu_0,\pi} \in \mathcal{K}_{P,\gamma}$ and that for any element $\mu \in \mathcal{K}_{P,\gamma}$, we can construct the corresponding policy $\pi_\mu(s) = \mu(s,a) / \int_{\mathcal{A}} \mu(s,a) da$. We will now introduce notation to formally write the reinforcement learning problem described in 1 since it will serve useful for the remainder of the paper.

Definition 1 *For a reward function $r : \mathcal{X} \to \mathbb{R}$, we define*

$$\mathrm{RL}_{P,\gamma}(r) := \sup_{\mu \in \mathcal{K}_{P,\gamma}} \int_{\mathcal{X}} r(s,a) d\mu(s,a)$$
$$M_{P,\gamma}(r) := \arg\sup_{\mu \in \mathcal{K}_{P,\gamma}} \int_{\mathcal{X}} r(s,a) d\mu(s,a)$$

In the above, $\mathrm{RL}_{P,\gamma}(r)$ is the same as (1) and represents the maximum expected reward possible under an environment P, discount factor γ and reward function r. The set $M_{P,\gamma}(r) \subseteq \mathscr{P}(\mathcal{X})$ represent the solutions that achieve maximal expected reward.

Convex Analysis and Legendre-Fenchel Duality We use $\mathscr{B}(\mathcal{X})$ to denote the set of finitely-additive measures and denote its topological dual to be $\mathcal{F}_b(\mathcal{X})$, the set of measurable and bounded functions mapping from $\mathcal{X}$ to $\mathbb{R}$. For any functional $F : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$, we define the Legendre-Fenchel dual, for any $h \in \mathcal{F}_b(\mathcal{X})$ as

$$F^\star(h) = \sup_{\mu \in \mathscr{B}(\mathcal{X})} \left(\int_{\mathcal{X}} h(x) d\mu(x) - F(\mu) \right).$$

For a set of functions $\mathcal{F} \subseteq \mathcal{F}_b(\mathcal{X})$, we use $\iota_{\mathcal{F}}(h)$ to denote the convex indicator function defined which is 0 if $h \in \mathcal{F}$ and $+\infty$ otherwise. For any two measures $\mu, \nu \in \mathscr{B}(\mathcal{X})$, we define the f-divergence between μ and ν to be $D_f(\mu,\nu) = \int_{\mathcal{X}} f(d\mu/d\nu) d\nu - \int_{\mathcal{X}} d\nu + 1$ where $f : \mathbb{R} \to (-\infty, \infty]$ is a lower semicontinuous convex function with $f(1) = 0$. In particular, the setting of $f(t) = t \log t$ is the popular Kullback-Leiber divergence, which we denote by $\mathrm{KL}(\mu,\nu) = D_f(\mu,\nu)$.

3 Related Work

Our main contribution is a reinterpretation of regularized policy maximization as robustifying mechanisms and so we discuss developments at understanding these methods along with similar results existing in machine learning at large. The idea of using causal entropy (Ziebart, 2010) is guided by the intuition of encouraging curious and diversified behavior. Further developed in (Haarnoja et al., 2018c), empirical success of using this penalty has been apparent. In particular, regularized policies unlike standard policies have illustrated robust behavior in the face of uncertainty and diversified behavior in finite sample schemes. Despite the empirical success, there is not much work studying these benefits from a formal perspective. The main existing results show that regularized objectives include smoothen the optimization landscape (Ahmed et al., 2019) and yield sparse policies (Yang et al., 2019). (Eysenbach and Levine, 2019) focuses on the MaxEnt framework and relates the optimal policy to solving a variable reward problem, which is line with our findings. Their results in contrast to ours, cannot be applied to other policy regularizers or other schemes that use causal entropy in the absence of reward functions such as adversarial imitation learning (Li et al., 2017).

In the realm of machine learning more generally, regularization has been principally established as a robustifying strategy. In supervised learning, various forms of robustness have shown connections to a number of regularization penalties such as Lipschitzness (Blanchet and Murthy, 2019; Sinha et al., 2017; Cranko et al., 2020; Husain, 2020), variance (Duchi et al., 2016) and Hilbert space norms (Staib and Jegelka, 2019). In Optimal Transport (OT), it has also been shown that entropic regularization is linked to ground cost robustness (Paty and Cuturi, 2020). Our result thus extends and develops these narratives for RL. (Zhang et al., 2020) also uses technical tools similar to our work such as Fenchel duality however for their purposes and findings are for quite different purposes.

4 Reward Robust Reinforcement Learning

We will be focusing on the problem specified by

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu),$$

where $R : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$ is a concave upper semicontinuous function. Note that when a reward function $r : \mathcal{X} \to \mathbb{R}$ is given, setting $R(\mu) = \int_{\mathcal{X}} r(x) d\mu(x)$ recovers the standard maximum expected reward problem. Furthermore, the above subsumes other developments of RL in the case where the reward is unknown and R is chosen to be the entropy (Hazan et al., 2019) or imitation learning when $R(\mu) = -D(\mu, \mu_E)$ where μ_E is some expert demonstration and D is a divergence between probability measures (Ghasemipour et al., 2019). We present the main result which shows the above as a reward robust RL problem.

Theorem 1 *For any concave upper semicontinuous function $R : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$, we have*

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) = \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\mathrm{RL}_{P,\gamma}(r') + (-R)^\star(-r') \right)$$

Proof (Sketch) The key part of the proof is to rewrite R in terms of the convex conjugate of $-R$, which is well-defined since $-R$ is lower semicontinuous and convex, by assumptions on R. The proof then concludes by moving the supremum over μ inside by an application of a generalized minimax theorem. ■

The key point from the above is that the value of the maximal policy over R is exactly equal to the problem of finding an adversarial reward. In particular, the adversarial reward problem seeks to find a reward r' that makes the maximally achievable reward $\mathrm{RL}_{P,\gamma}$ as small as possible while paying the penalty $(-R)^\star(-r')$, where $(-R)^\star$ is a convex function. We remark that this is a one-party problem involving only an adversary. The conventional notion of robustness would relate this to the optimal model μ. We do this precisely by presenting a result that links the optimal μ and adversarial reward r':

Theorem 2 *Let μ^* and r^* be the optimal solution to the problems specified in Theorem 1, then we have that $\mu^* \in M_{P,\gamma}(r^*)$.*

This result tell us that an optimal policy found by solving the regularized objective is precisely an optimal policy of the Reinforcement Learning problem specified by the adversarial reward r^*. This is particularly striking since it tells us that though we are maximizing some concave R, which may be motivated for separate purposes, we can always guarantee that the policy learned is optimal for some reward r' in the axiomatic utility theory sense. In particular, this reward r^* is chosen to be the worst-case for this environment. The strength of robustness and nature of the adversarial reward clearly depends on the choice of R, as this is what budgets the adversarial reward r'. We will show that under a technical assumption on R, we can characterize the form r^* takes, which happens to depend on a single state-dependent mapping $V \in \mathcal{F}_b(\mathcal{S})$. The particular technical assumption on $(-R)^\star$ is that it is *increasing* by which we mean $r(x) \geq r'(x)$ for every $x \in \mathcal{X}$ implies $(-R)^\star(r) \geq (-R)^\star(r')$. We first introduce a result.

Theorem 3 *Suppose R is concave upper semicontinuous and let $\mathscr{I}$ be the value of the optimization problem*

$$\inf_{V \in \mathcal{F}_b(\mathcal{S}), r \in \mathcal{F}_b(\mathcal{X})} \left((1-\gamma) \int_{\mathcal{S}} V(s) d\mu_0(s) + (-R)^\star(-r) \right), \tag{2}$$

$$\text{s.t.}\, V(s) \geq r(s,a) + \gamma \int_{\mathcal{S}} V(s') dP(s' \mid s, a).$$

It then holds that $\mathscr{I} = \sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu)$.

It should be first noted that the above is a strong duality Theorem and indeed is a generalized version of the standard linear programming duality between policy maximization and value function minimization as described in (Agarwal et al., 2019), which is recovered when $R(\mu) = \int_{\mathcal{X}} r(x) d\mu(x)$ for some reward r. We will now show that the optimal value function of this objective gives the optimal reward. In particular, note that by solving the above constraint for the reward yields

$$r_V(s,a) := V(s) - \gamma \cdot \int_{\mathcal{S}} V(s') dP(s' \mid s,a). \qquad (3)$$

We then have the following result

Lemma 1 *Suppose $(-R)^\star$ is increasing and V^* is the optimal solution of (2) then r_{V^*} is the optimal adversarial reward.*

The main consequence of the above Lemma is that it characterizes the shape of the adversarial reward chosen. In particular, it tells us that as long as as R satisfies the technical assumption ($(-R)^\star$ is increasing), the adversarial reward will be of the form r_V for some V. This is insightful since it tells us that the adversarial reward relates rewards between states through the dynamics of P. For example, note that if a particular state-action pair (s,a) yields the same state s then $r_V(s,a) = (1-\gamma)V(s)$. This technical condition on R can be satisfied for any R with a simple reparametrization, which we lay out in Lemma 1 in the supplementary material, and exploit when deriving $(-R)^\star$ for Soft-Actor-Critic. Moreover, we will show that the common choices of R which are motivated for smoothing or other empirical benefits naturally satisfy this technical assumption.

Generalized Soft-Actor-Critic Regularization Consider the case of having an available reward and using a convex penalty $\Omega : \mathscr{B}(\mathcal{X}) \times \mathscr{B}(\mathcal{X}) \to \mathbb{R}$ for the policy so we select $R = R_\Omega$ of the form

$$R_\Omega(\mu) = \int_{\mathcal{X}} r(s,a) d\mu(s,a) - \varepsilon \cdot \Omega(\mu),$$

for some $\varepsilon > 0$. It can easily be shown (see Appendix) that $(-R)^\star(-r') = \varepsilon \Omega^\star \left(\frac{r-r'}{\varepsilon}\right)$, so that we have the following.

Corollary 1 *Let $\Omega : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$ be a convex penalty then for any $\varepsilon > 0$ we have*

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R_\Omega(\mu) = \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\mathrm{RL}_{P,\gamma}(r') + \varepsilon \Omega^\star \left(\frac{r - r'}{\varepsilon} \right) \right).$$

The above tells us that the adversarial reward problem pays a price for deviating from the given reward r due to the second term $\varepsilon \Omega^\star \left(\frac{r-r'}{\varepsilon}\right)$. In the Soft-Actor-Critic (SAC) method, this corresponds to selecting (upto some constant) $\Omega_{\mathrm{SAC}}(\mu) = \mathbb{E}_{\mu(s,a)}\left[\mathrm{KL}(\pi_\mu(\cdot \mid s), U)\right]$, where π_μ is the policy induced by μ and U is the uniform distribution over $\mathcal{A}$. We presented Corollary 1 with a general Ω, which we believe will be useful for future developments. In this work, we consider the causal policy entropy along with 2-Tsallis entropy in the next next section. For the SAC case, we have the following result

Lemma 2 (Soft-Actor-Critic) *For any $\varepsilon > 0$ and $r, r' \in \mathcal{F}(\mathcal{X})$, we have*

$$\varepsilon \Omega^\star_{\mathrm{SAC}} \left(\frac{r - r'}{\varepsilon} \right) = \varepsilon \cdot \sup_{s \in \mathcal{S}} \left(\int_{\mathcal{X}} \exp\left(\frac{r(s,a) - r'(s,a)}{\varepsilon} \right) dU(a) - 1 \right)$$

If one reasons about how the adversary behaves, the first incentive is to make $\mathrm{RL}_{P,\gamma}(r')$ small by selecting very small rewards across the environment. However, we can see that for the case of entropic regularization, the adversary pays a big price for selecting r' to be far from the original reward r for any given state. Note that in this case, we have $(-R)^\star$ is increasing and so in light of the concrete insight found in Lemma 1, we are able to reason about the SAC policy maximizing a reward of the worst-case reward of the form (3). This is striking since it tells us that the adversarial reward r' will respect the environment dynamics across the action space even if the ground reward r does not.

Derivation of Q-learning through robust learning In this subsection, we derive Q-learning through the reward-robust RL framework. In this context, learning a policy that is robust to a small variation in the reward corresponds to allowing a small violation of the Bellman equation with respect to the original reward function. For any Q-function $Q \in \mathcal{F}_b(\mathcal{X})$, we define the bellman operator $\mathcal{T}_r : \mathcal{F}_b(\mathcal{X}) \to \mathcal{F}_b(\mathcal{X})$ as

$$\mathcal{T}_r Q(s,a) = r(s,a) + \gamma \int_{\mathcal{X}} \sup_{a' \in \mathcal{A}} Q(s',a') dP(s' \mid s,a)$$

The maximum reward problem can be restated as

$$\mathrm{RL}_{P,\gamma}(r) = \inf_{Q \geq \mathcal{T}_r Q} \int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s), \qquad (4)$$

where the optimal $Q^* \in \mathcal{F}_b(\mathcal{X})$ from the above is a contraction of $\mathcal{T}_r$ meaning that $\mathcal{T}_r Q^* = Q^*$. As it is difficult to find this contraction, one method known as *deep Q-learning* tackles this by parametrizing Q with a deep neural network and uses regression in the supervised learning sense to match $\mathcal{T}_r Q$ to Q (Sutton and Barto, 2018). This will deviate from the original objective since it relaxes this constraint $Q = \mathcal{T}_r Q$ into the term appearing in the objective, which will naturally introduce bias. We now show quite a remarkable connection that doing so is related to policy regularization and by virtue of Corollary 1, linked to reward robustness.

Theorem 4 *For any $\varepsilon > 0$ and convex Ω such that $\Omega^\star$ is increasing, we have*

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R_\Omega(\mu) = \inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\varepsilon \Omega^\star \left(\frac{\mathcal{T}_r Q - Q}{\varepsilon} \right) + \int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) \right).$$

We remark that the above is an inequality if $\Omega^\star$ is not increasing which results in *weak duality.* First note that the Theorem is precisely a relaxed *unconstrained* version of *constraint* objective appearing in (4). The most notable aspect of this result is that it links the regularized objective to finding a Q-function that minimizes the difference in the Bellman update $\varepsilon\Omega^\star\left(\frac{\mathcal{T}_r Q - Q}{\varepsilon}\right)$, depending on the choice of Ω. There exists work that show a relationship between gradients in entropy regularization and Q-learning (Schulman et al., 2017), however we state a more generalized result and bridge it to reward robustness. To see how this relates to the existing losses used in Q-learning, let us consider both the finite and continuous case. In the finite case, we can pick $\Omega(\mu) = \sum_{x \in \mathcal{X}} \mu(x)^2$, which is the 2-Tsallis entropy. One can easily derive the dual $\Omega^\star(r) = \frac{1}{4}\sum_{x\in\mathcal{X}} r(x)^2$ and thus the right side of Theorem 4 becomes (setting $\varepsilon = 1$)

$$\inf_{Q \in \mathcal{F}_b(\mathcal{X})} \left(\frac{1}{4} \sum_{(s,a)\in\mathcal{X}} \left(\mathcal{T}_r Q(s,a) - Q(s,a)\right)^2 + \int_{\mathcal{S}} \sup_{a \in \mathcal{A}} Q(s,a) d\mu_0(s) \right).$$

The variational problem above is a regression problem between Q and $\mathcal{T}_r Q$ using the squared loss, which is the typical objective in deep Q-learning. The consequence of our result is that using this particular choice of loss to learn the Q function is related to learning a policy with the 2-Tsallis entropy, which is rather striking. Furthermore, the 2-Tsallis entropy behaves similar to the Shannon entropy in the sense that it is maximized when μ is uniform and minimized when μ is degenerate. In the continuous case, a buffer distribution $\nu \in \mathscr{P}(\mathcal{X})$ is used for the loss by defining the mean-squared error as L^2 norm with respect to ν between $\mathcal{T}_r Q$ and Q: given by $\|\mathcal{T}_r Q - Q\|^2_{L^2(\nu)}$. In this case, it can be shown that if $\Omega(\mu) = \frac{1}{4}\int_{\mathcal{X}} \left(\frac{d\mu}{d\nu}\right)^2 d\nu$ when $\mu \ll \nu$ and $+\infty$ otherwise then $\Omega^\star(h) = \|h\|^2_{L^2(\nu)}$.

Imitation Learning One method of learning a policy is to imitate expert data which comes in the form of a given distribution $\mu_E \in \mathscr{P}(\mathcal{X})$. Unlike the regularized schemes above, there is no specified reward function. Using the unified perspective provided in (Ghasemipour et al., 2019), where imitation learning is cast as divergence minimization, we can write these methods into our framework by selecting $R(\mu) = -D(\mu, \mu_E)$ (for each corresponding divergence). In particular, our goal is to not only derive the corresponding robust-reward problem but also show that $(-R)^\star$ will be increasing for these cases. We delegate the technical derivations to the Supplementary Section 1.8 and only present the results here. First, we focus on Adversarial Inverse Reinforcement Learning (AIRL) (Fu et al., 2017) selecting $R(\mu) = -\operatorname{KL}(\mu, \mu_E)$ in which case we have

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) = \inf_{r' \in \mathcal{F}_b(\mathcal{X})} \left(\mathrm{RL}_{P,\gamma}(r') + \int_{\mathcal{X}} \exp\left(-r'(x)\right) d\mu_E(x) - 1 \right),$$

noting that $(-R)^\star$ is increasing. We show the more general result that when $R(\mu) = -D_f(\mu, \mu_E)$ where D_f is an f-divergence then $(-R)^\star$ will be increasing. Using this choice of R corresponds to f-MAX (Ghasemipour et al., 2019). Another method for imitation learning is to use a discriminator based divergence as employed in InfoGAIL (Li et al., 2017). In this setting we assume we have a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and denoting the Lipschitz constant of a function $h \in \mathcal{F}_b(\mathcal{X})$ as $\mathrm{Lip}_d(h) := \sup_{x,x' \in \mathcal{X}} |h(x) - h(x')| / d(x,x')$, we set

$$R(\mu) = - \sup_{h : \mathrm{Lip}_d(h) \leq L} \left(\int_{\mathcal{X}} h(x) d\mu(x) - \int_{\mathcal{X}} h(x) d\mu_E(x) \right),$$

where $L > 0$ is chosen as a hyperparameter. In this case, we have

$$\sup_{\mu \in \mathcal{K}_{P,\gamma}} R(\mu) = \inf_{r' : \mathrm{Lip}_d(r') \leq L} \left(\mathrm{RL}_{P,\gamma}(r') - \int_{\mathcal{X}} r' d\mu_E \right).$$

It is clear from the above that the adversarial reward seeks to ensure $\mathrm{RL}_{P,\gamma}$ is as low as possible while maintaining that r' is large around the expert trajectory due to the second term. It should also be noted that the choice of L reflects as the budget of the adversary. We do not have $(-R)^\star$ increasing for this choice of R. On the other hand, it is typical in practice that an

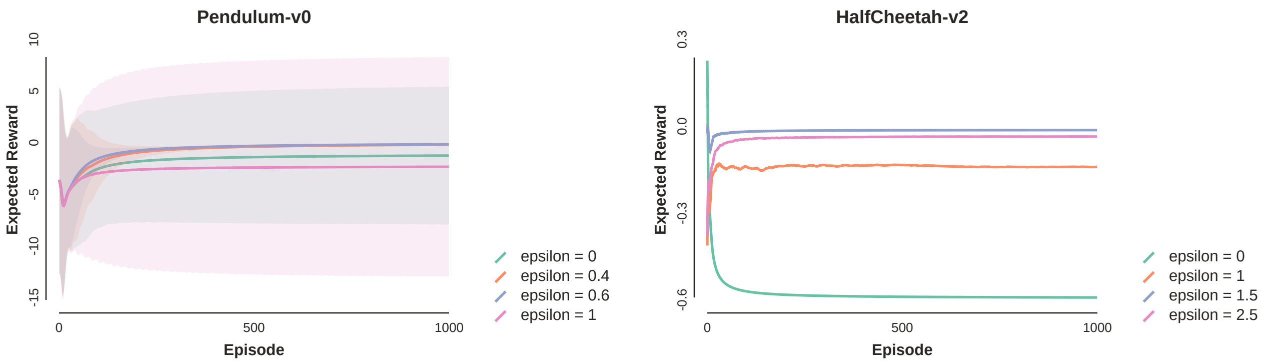

Figure 2: Expected reward over 1000 episodes of policies returned by SAC trained on an adversarial reward r_{adv} and tested on the true reward using different weighting ε for entropy.

entropy term is included in this term:

$$R(\mu) = -\sup_{h:\text{Lip}_d(h)\le L}\left(\int_{\mathcal{X}} h(x)d\mu(x) - \int_{\mathcal{X}} h(x)d\mu_E(x)\right) - \varepsilon\mathbb{E}_{\mu(s,a)}\left[\text{KL}(\pi_\mu(\cdot \mid s), U_A)\right],$$

for some $\varepsilon > 0$ where U_A is the uniform distribution over $\mathcal{A}$. Under this setting, it turns out that $(-R)^\star$ is now increasing, in which case Lemma 1 applies. It is rather intriguing that the role of entropy here ensures that the reward that the InfoGAIL policy maximizes is worst-case, of high value around trajectories from the expert, and attains the familiar shape in Equation (3). This further advocates for the use of entropy regularization.

Entropic Exploration We now consider the case where there is no reward function or expert distribution specified and the only objective to maximize is entropy. For such a scheme, there exists efficient algorithms (Hazan et al., 2019). More specifically, we have $R(\mu) = -\text{KL}(\mu, U_{\mathcal{X}})$ where $U_{\mathcal{X}}$ is the uniform distribution over $\mathcal{X}$. We then have that

$$\sup_{\mu\in\mathcal{K}_{P,\gamma}} R(\mu) = \inf_{r'\in\mathcal{F}_b(\mathcal{X})}\left(\text{RL}_{P,\gamma}(r') + \int_{\mathcal{X}} \exp\left(-r'(x)\right) dU_{\mathcal{X}}(x) - 1\right),$$

and similar to the other choices of R, we have that $(-R)^\star$ is increasing. We would like to remark that if one defines KL to be $+\infty$ when μ is not a probability measure then $(-R)^\star(r) = \log\int_{\mathcal{X}} \exp(r(x))dU_{\mathcal{X}}(x)$ (Ruderman et al., 2012).

5 Experiments

The main practical ramification of our work is to advocate the use of regularized policies by highlighting the robustification aspect, for which we derived a strong theoretical link. There exists extensive empirical evidence for which our work provides foundation for. However, we will show some brief yet illustrative examples which focus on the reward adversarial aspect of regularized policies, as illustrated by our main result Theorem 1. Our goal is thus to see the performance of regularized policies on rewards they are not trained on and analyze their behavior based on the robustness parameter ε. First we consider the Pendulum-v0 environment and train the Soft-Actor-Critic (SAC) method on a reward that has been altered with. We do so by constructing an adversarial reward r_{adv} using

$$r_{\text{adv}} = \begin{cases} r(s,a) + \delta & \text{if } r(s,a) \le -5 \\ r(s,a) & \text{otherwise} \end{cases}$$

where δ is drawn from a normal distribution centered at 5 with variance 0.1. In doing so, initial states of the pendulum will be favored and easier to reach however the maximal reward will still be attained at the inverted position. We train SAC for various values of ε and test their performance on the true reward in Figure 2 (left). We find that the effect of increasing ε yields better performance than no entropy however adding too much entropy (in the case of $\varepsilon = 1$) damages performance. We repeat a similar experiment for HalfCheetah-v2 however using an adversarial reward specified by

$$r_{\text{adv}} = \begin{cases} r(s,a) + \delta & \text{if } r(s,a) \le 0 \\ r(s,a) & \text{otherwise} \end{cases}$$

where δ is drawn from a normal distribution centered at 3 with variance 0.1. We plot the performance under the expected reward in Figure 2 (right). It can also be seen that adding entropy surpasses the non-regularized policy $\varepsilon = 0$ and that increasing ε higher will worsen performance (as seen by $\varepsilon = 2.5$).

6 Conclusion

Our results allow us to reason about regularization of policies and the regression Q-learning objective from the perspective of robustness. This is not surprising given the advancements in machine learning more generally pointing at the link between regularization and robustness along with the impressive empirical evidence of these schemes. Regularized objectives, however, offer other benefits that are inherently sample based phenomenon such as smoothened objectives or stable training. While our results do not directly target this, we have built a connection between two objectives which will pose modular for future developments.

Acknowledgements

We would like to thank Zakaria Mhammedi for useful feedback regarding the technical analysis. HH is funded by the Australian Government Research Training Program and Data61.

References

Agarwal, A., Jiang, N., and Kakade, S. M. (2019). Reinforcement learning: Theory and algorithms. Technical report, Technical Report, CS Department, UW Seattle.

Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. (2019). Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pages 151–160.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Cranko, Z., Shi, Z., Zhang, X., Nock, R., and Kornblith, S. (2020). Generalised lipschitz regularisation equals distributional robustness. *arXiv preprint arXiv:2002.04197.*

Dayan, P. and Hinton, G. E. (1997). Using expectation-maximization for reinforcement learning. *Neural Computation*, 9(2):271–278.

Duchi, J., Glynn, P., and Namkoong, H. (2016). Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425.*

Eysenbach, B. and Levine, S. (2019). If maxent rl is the answer, what is the question? *arXiv preprint arXiv:1910.01913.*

Fu, J., Luo, K., and Levine, S. (2017). Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248.*

Ghasemipour, S. K. S., Zemel, R., and Gu, S. (2019). A divergence minimization perspective on imitation learning methods. *arXiv preprint arXiv:1911.02256.*

Haarnoja, T., Ha, S., Zhou, A., Tan, J., Tucker, G., and Levine, S. (2018a). Learning to walk via deep reinforcement learning. *arXiv preprint arXiv:1812.11103.*

Haarnoja, T., Pong, V., Zhou, A., Dalal, M., Abbeel, P., and Levine, S. (2018b). Composable deep reinforcement learning for robotic manipulation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6244–6251. IEEE.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018c). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290.*

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. (2019). Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573.

Hong, Z.-W., Shann, T.-Y., Su, S.-Y., Chang, Y.-H., Fu, T.-J., and Lee, C.-Y. (2018). Diversity-driven exploration strategy for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 10489–10500.

Husain, H. (2020). Distributional robustness with ipms and links to regularization and gans. *Advances in Neural Information Processing Systems*, 33.

Kappen, H. J. (2005). Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11):P11011.

Li, Y., Song, J., and Ermon, S. (2017). Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*, pages 3812–3822.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability.* Springer Science & Business Media.

Neumann, G. et al. (2011). Variational inference for policy search in changing situations. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 817–824.

Paty, F.-P. and Cuturi, M. (2020). Regularized optimal transport is ground cost adversarial. *arXiv preprint arXiv:2002.03967.*

Rawlik, K., Toussaint, M., and Vijayakumar, S. (2013). On stochastic optimal control and reinforcement learning by approximate inference. In *Twenty-third international joint conference on artificial intelligence.*

Revuz, D. (2008). *Markov chains.* Elsevier.

Ruderman, A., Reid, M., García-García, D., and Petterson, J. (2012). Tighter variational representations of f-divergences via restriction to probability measures. *arXiv preprint arXiv:1206.4664.*

Russell, S. and Norvig, P. (2002). Artificial intelligence: a modern approach.

Schulman, J., Chen, X., and Abbeel, P. (2017). Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440.*

Sinha, A., Namkoong, H., and Duchi, J. (2017). Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2.

Staib, M. and Jegelka, S. (2019). Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, pages 9131–9141.

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

Theodorou, E., Buchli, J., and Schaal, S. (2010). A generalized path integral control approach to reinforcement learning. *The Journal of Machine Learning Research*, 11:3137–3181.

Todorov, E. (2007). Linearly-solvable markov decision problems. In *Advances in neural information processing systems*, pages 1369–1376.

Toussaint, M. (2009). Robot trajectory optimization using approximate inference. In *Proceedings of the 26th annual international conference on machine learning*, pages 1049–1056.

Wu, Y., Tucker, G., and Nachum, O. (2019). Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361.*

Yang, W., Li, X., and Zhang, Z. (2019). A regularized approach to sparse optimal policy in reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5940–5950.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. *arXiv preprint arXiv:2007.02151.*

Ziebart, B. D. (2010). Modeling purposeful adaptive behavior with the principle of maximum causal entropy.