Viability of Future Actions: Robust Safety in Reinforcement Learning via Entropy Regularization

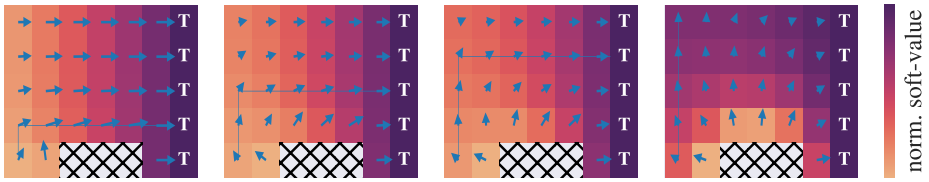
Pierre-François Massiani*,1, Alexander von Rohr*,1,2, Lukas Haverbeck[1], and Sebastian Trimpe[1]

[1] Institute for Data Science in Mechanical Engineering, RWTH Aachen University, Germany {massiani,lukas.haverbeck,trimpe}@dsme.rwth-aachen.de
[2] Learning Systems and Robotics Lab, Technical University of Munich, Germany alex.von.rohr@tum.de
* Equal contribution.

Abstract. Despite the many recent advances in reinforcement learning (RL), the question of learning policies that robustly satisfy state constraints under unknown disturbances remains open. In this paper, we offer a new perspective on achieving robust safety by analyzing the interplay between two well-established techniques in model-free RL: entropy regularization, and constraints penalization. We reveal empirically that entropy regularization in constrained RL inherently biases learning toward maximizing the number of future viable actions, thereby promoting constraints satisfaction robust to action noise. Furthermore, we show that by relaxing strict safety constraints through penalties, the constrained RL problem can be approximated arbitrarily closely by an unconstrained one and thus solved using standard model-free RL. This reformulation preserves both safety and optimality while empirically improving resilience to disturbances. Our results indicate that the connection between entropy regularization and robustness is a promising avenue for further empirical and theoretical investigation, as it enables robust safety in RL through simple reward shaping.

1 Introduction

Safety is the ability of a policy to keep the system away from a failure set of undesirable states. Robustness extends the notion to adversarial or noisy settings; robust policies remain outside of the failure set in spite of the noise or adversary. While robust reinforcement learning (RL) may be formulated as a constrained optimization problem [1], there is a strong appeal in achieving robustly safe policies through reward shaping alone, given the numerous algorithms available for unconstrained RL. The purpose of this work is to reveal how robustly safe policies arise naturally from two common practices in RL; namely, maximum-entropy RL [2] and failure penalization [3]. Our results support that the maximum-entropy RL objective together with failure penalties enable safe operation at testing under action noise stronger than that seen at training; a property we call *robustness*.



(a) $\alpha = 1$ (b) $\alpha = 4$ (c) $\alpha = 8$ (d) Max. ent.

Fig. 1. Fenced cliff — Robustness as a function of α: An entropy-regularized policy avoids states with fewer actions available (d). The degree is controlled by the temperature parameter α. As it increases (a–c), the policy moves away from the constraints, getting more robust but taking longer to reach the target. The mode of the policy is shown as a thin blue line.

Since "robustness" is a highly overloaded term in RL, we emphasize the notion of robustness in this paper differs from those of previous studies [4,5,6]. Indeed, they guarantee that entropy regularization preserves a high return under changes in the reward or dynamics. In other words, the *return* is robust to such changes. In contrast, we want that *safety constraints are still satisfied* under changes in the dynamics (namely, the level of action noise). These two types of objectives are complementary since safety and optimality are generally separate concerns in optimal control, where the goal is to act optimally while abiding by safety constraints. Similarly, the term "entropy-regularized RL" is used in the literature to refer to various formulations of regularized MDPs [7]. In this paper, we use it specifically to refer to methods that optimize the soft RL objective (3) as in [8], which are often referred to as maximum-entropy RL. From here on, we reserve the term "maximum-entropy" for the special case where the reward is identically zero and the agent solely maximizes entropy.

Our approach builds on two key contributions. First, we empirically show that entropy regularization in a *constrained* environment induces robustness to action noise. We do this by first showing that agents optimizing the entropy-regularized objective may sacrifice reward to avoid constraints boundaries, with the degree of avoidance modulated by the temperature parameter. This is illustrated in Figure 1, where higher temperatures yield policies whose mode stays farther from constraints. Then, we provide empirical evidence that this constraints avoidance translates to robustness to action noise, i.e., policies preserving the long-term number of viable actions are generally more robust.

This general behavior aligns with the viability-based notion of robustness (called "safety" therein) introduced by [8,9] (cf. [10]), where the robustness of a state is quantified by the number of viable actions — those that allow indefinite constraint satisfaction. We interpret the cumulative discounted entropy of a policy as a proxy for the long-term number of safe actions it considers, and thus entropy regularization naturally encourages avoidance of states with few viable options.

Our second contribution shows that this constrained setting can be approximated arbitrarily well using failure penalties. For penalties above a finite threshold,

the *mode* of the resulting policy matches that of the constrained problem, offering a practical reward-shaping strategy for learning robustly safe policies.

Making this observation relevant to state-of-the-art RL requires relating the unconstrained and state-constrained optimization problems, as most applications focus on the former [11]. This is the role of *constraints penalization* or, more concisely, of penalties. In the absence of entropy regularization, they are known to make constraints violations suboptimal, and large enough penalties guarantee that policies optimal for the penalized problem are also optimal for the constrained problem [3]. We add entropy regularization to this analysis. Furthermore, penalties above a *finite* threshold recover the mode of the constrained policy.

Our observations emphasize a benefit of entropy regularization that differs from what is commonly mentioned in the literature. Indeed, algorithms such as soft actor-critic (SAC) [2] are often praised for their excellent exploration and their robustness to the choice of hyperparameters [2]. Although crucial in practice, these strengths are relevant *during* learning. In contrast, we focus on the optimal policy, that is, on what occurs *after* successful learning.

Contributions We reveal how robust optimal controllers arise from the combination of entropy regularization together with sufficient constraints penalties. Specifically:

1. We identify empirically that constraints repel trajectories of optimal controllers in the presence of entropy regularization, by favoring controllers maximizing the number of future viable actions.
2. We prove that failure penalties approximate this constrained problem arbitrarily closely.
3. Finally, we show that we can extract a safe policy from the optimal solution to the penalized problem and demonstrate that this policy is robustly safe.

The first contribution strongly supports that the mode of entropy-regularized policies is robust to action noise, as the most-likely trajectory is "repelled" by the constraints. We confirm robustness to action noise empirically, and further theoretical investigation is a promising avenue for future work. Together, our results enable achieving reward-shaping-based robustness, and a novel interpretation of the temperature coefficient in the presence of constraints as a tunable robustness parameter.

The article is organized as follows. We discuss other approaches to robustness in RL in Section 2. We then expose necessary preliminaries in Section 3, and formalize the problem we consider in Section 4. Section 5 contains our theoretical results, with first a high-level interpretation of the constrained, entropy-regularized problem, and our main theorem guaranteeing approximation with penalties. The empirical evidence on robustness follows in Section 6, together with further empirical validation of our theoretical results.

A complete version of this paper together with its appendix is available at this address: `www.doi.org/10.48550/arXiv.2506.10871`

2 Related work

Viability and safety in RL There is a variety of definitions of safety in RL [11]. We consider the case of avoiding state constraints with certainty (level 3 in [11]). Such a definition of safety falls into the general problem of viability [10]. Many specialized algorithms were developed to solve this safe RL problem, both model-free and model-based [12]. It has been shown in [3] that sufficient failure penalties enforce equivalence between the unconstrained and safety-constrained problems, making safe RL amenable to unconstrained algorithms. This idea falls in the class of *penalty methods*, a general idea in optimization which has been studied in the context of optimal control [13,14]; a reformulation of the results of [3] is that the discounted risk is an exact penalty function. Our results show that it is no longer the case for entropy-regularized RL, as no finite penalty exactly recovers the constrained problem. Yet, we show it can be approximated arbitrarily closely. Regardless, the above works only guarantee safety and neglect robustness. We extend the analysis and proof methods of [3] to entropy-regularized RL, which naturally yields robustness in addition to safety.

Robustness in optimal control Robustness is a well-studied topic in optimal control [15] and consists of preserving viability despite *model uncertainties*. Classical approaches consist of robust model predictive control [16,17] and Hamilton-Jacobi reachability analysis [18]. They provide *worst-case* guarantees, mainly through constraints tightening. The robustness of entropy-regularized controllers does not fit directly in this category, as their full support makes them explore the whole viability kernel. Instead, they seem to exhibit a form of "expected" constraints tightening, which translates into robustness to action noise of the mode, as we illustrate empirically. Finally, alternative methods such as scenario optimization [19] address quantitative uncertainty instead of worst-case, but the connection to the robustness discussed in this article is still open.

Robustness in RL Achieving robustness for RL policies is an active area of research [1]. A common formalization is that of a two-player game between the agent and an adversary [20,21]. This setup is akin to that of Hamilton-Jacobi reachability analysis, only with a discounted cost. These approaches achieve robustness through an adversary controlling, for instance, disturbances [21] or action noise [22], yielding worst-case robustness. However, such adversarially-robust RL requires specialized algorithms and training the adversary. In contrast, entropy-regularized RL is a popular framework with many standard implementations, which, as we show, also yields robustness solely through reward shaping.

The work of [9] introduces a state-dependent safety measure based on the number of viable actions available in each state. Our work extends this notion to robust safety of policies. A detailed discussion on the connection with the safety measure therein is in Appendix C.

We are not the first to report that entropy-regularization leads to robustness. Some empirical [4,5] and theoretical works [6] highlight the inherent robustness of entropy-regularized RL. As mentioned above, however, their definition of

robustness differs: [6] consider robustness of the return to changes in the dynamics, whereas we are interested in preserving constraints satisfaction.

The observation that action noise during training can lead to more robust behavior was already made in [23, Example 6.6] on the famous cliff walking grid-world. There, ε-greedy action selection resulted in more robust behavior for the case of on-policy learning (SARSA), whereas Q-learning (an off-policy method) learns to the optimal, non-robust, policy. We take the same example in Fig. 1 and observe that entropy regularization leads to robust behavior in off-policy RL. Similarly, the G-learning algorithm exhibits the same robust behavior on the cliff environment [24]. Our results and interpretation provide a general explanation for this observation.

3 Preliminaries

We introduce concepts to frame the optimization problems and their constraints. In particular, we address entropy-regularized RL and viability.

3.1 Entropy-regularized RL

We consider finite sets $\mathcal{X}$ and $\mathcal{A}$ called the state and action spaces, respectively, and deterministic dynamics $f : \mathcal{Q} \to \mathcal{X}$, where $\mathcal{Q} = \mathcal{X} \times \mathcal{A}$ is the state-action space. A policy $\pi : \mathcal{Q} \to [0, 1]$ is a map whose partial evaluation in any $x \in \mathcal{X}$ is a probability mass function on $\mathcal{A}$; we write $\pi(\cdot \mid x)$, and $\mathit{\Pi}$ is the set of all policies. The state at time $t \in \mathbb{N}$ from initial state $x \in \mathcal{X}$ and following $\pi \in \mathit{\Pi}$ is $X(t; x, \pi)$, and the action taken by π at that time is $A(t; x, \pi)$. If the policy and initial state are unambiguous, we simply write X_t and A_t.

We also consider $r : \mathcal{Q} \to \mathbb{R}$ a bounded reward function. The return of $\pi \in \mathit{\Pi}$ from initial state $x \in \mathcal{X}$ is then

$$G(x, \pi) = \sum_{t=0}^{\infty} \gamma^t r(X_t, A_t), \tag{1}$$

where $\gamma \in (0, 1)$ is the discount factor. A smaller γ disregards delayed rewards, but can be overcome if the said rewards have large magnitude. The expected return is $\bar{G}(x, \pi) = \mathbb{E}[G(x, \pi)]$. With $\mathcal{H}$ as the entropy, we introduce the discounted cumulative entropy of $\pi \in \mathit{\Pi}$ from $x \in \mathcal{X}$ as

$$S(x, \pi) = \sum_{t=0}^{\infty} \gamma^t \mathcal{H}(\pi(\cdot \mid X_t)), \tag{2}$$

and its expectation $\bar{S}(x, \pi) = \mathbb{E}[S(x, \pi)]$. The objective of entropy-regularized RL is then to find an optimal policy, that is, a policy $\pi_{\text{opt}} \in \mathit{\Pi}$ such that

$$\pi_{\text{opt}} \in \arg\max_{\pi \in \mathit{\Pi}} \bar{G}(x, \pi) + \alpha \bar{S}(x, \pi), \quad \forall x \in \mathcal{X}, \tag{3}$$

where $\alpha \in \mathbb{R}_{\geq 0}$ is a design parameter called the *temperature*. It is known that there exists an optimal policy [2]. Specifically, one can be computed by leveraging the optimal soft-Q-value function $q : \mathcal{Q} \to \mathbb{R}$, which satisfies for all $(x, a) \in \mathcal{Q}$ [25]:

$$q(x,a) = r(x,a) + \gamma\alpha \ln \left[\sum_{b \in \mathcal{A}} \exp\left(\frac{1}{\alpha} q(x', b) \right) \right], \tag{4}$$

where we defined the shorthand $x' = f(x, a)$. An equivalent definition is [25, Theorem 16]

$$q(x,a) = \max_{\pi \in \mathit{\Pi}} r(x,a) + \gamma \bar{G}(x', \pi) + \alpha\gamma \bar{S}(x', \pi). \tag{5}$$

Once q is known, the softmax policy solves (3):

$$\pi_{\text{opt}}(a \mid x) = \text{softmax}\left[\frac{1}{\alpha} q(x, \cdot) \right](a) \quad \forall (x,a) \in \mathcal{Q}. \tag{6}$$

Finally, for any policy $\pi \in \mathit{\Pi}$, its *mode* is the policy

$$\hat{\pi}(a \mid x) = \frac{1}{|\arg\max \pi(\cdot \mid x)|} \delta_{\arg\max \pi(\cdot|x)}(a), \tag{7}$$

where $|A|$ is the cardinality and $\delta_A(a)$ is the indicator function of a set $A \subset \mathcal{A}$.

3.2 Viability

We consider a set of failure states $\mathcal{X}_{\text{C}} \subset \mathcal{X}$ that the system should never visit. Avoiding $\mathcal{X}_{\text{C}}$ is a dynamic concern, and some states that are not in $\mathcal{X}_{\text{C}}$ themselves may still lead there inevitably. We address this through viability theory [10, Chapter 2].

Definition 1 (Viability kernel). *The viability kernel $\mathcal{X}_{\text{V}}$ is the set of states from where $\mathcal{X}_{\text{C}}$ can be avoided at all times almost surely:*

$$\mathcal{X}_{\text{V}} = \{x \in \mathcal{X} \mid \exists \pi \in \mathit{\Pi}, \forall t \in \mathbb{N}_{>0}, \mathbb{P}[X_t \notin \mathcal{X}_{\text{C}}] = 1\}.$$

By definition, any state that is not in the viability kernel leads to $\mathcal{X}_{\text{C}}$ in finite time. Such states are called *unviable*. The viability kernel is therefore the largest set that enables recursive feasibility of the problem of avoiding transitions into $\mathcal{X}_{\text{C}}$. A closely related concept is the *viable set*, which is the set of state-action pairs that preserve viability [8]:

$$\mathcal{Q}_{\text{V}} = \{(x,a) \in \mathcal{Q} \mid x \in \mathcal{X}_{\text{V}} \wedge f(x,a) \in \mathcal{X}_{\text{V}}\}.$$

We also define the unviable set $\mathcal{Q}_{\text{U}} = \mathcal{Q} \setminus \mathcal{Q}_{\text{V}}$, and the critical set $\mathcal{Q}_{\text{crit}} = \mathcal{Q}_{\text{U}} \cap (\mathcal{X}_{\text{V}} \times \mathcal{A})$ [26].

Definition 2. *Let $\pi \in \Pi$. We say that π is* safe from the state $x \in \mathcal{X}$ *if* $\mathbb{P}[X_t \notin \mathcal{X}_\mathrm{C}] = 1$ *for all* $t \in \mathbb{N}_{>0}$. *We say that π is* safe *if it is safe from any* $x \in \mathcal{X}_\mathrm{V}$. *For $\delta > 0$, we say that π is* δ-safe *if* $\max_{\mathcal{Q}_\mathrm{crit}} \pi \leq \delta$. *We denote the set of policies safe from the state x by $\Pi_\mathrm{V}(x)$ and that of safe policies by* $\Pi_\mathrm{V} = \bigcap_{x \in \mathcal{X}_\mathrm{V}} \Pi_\mathrm{V}(x)$.

By definition of the viability kernel, the condition for a safe policy can be replaced with $\mathbb{P}[X_t \in \mathcal{X}_\mathrm{V}] = 1$ for all $t \in \mathbb{N}$.

Remark 1. Another meaningful definition of δ-safety could be that the policy assigns at most δ of probability mass to unviable actions, that is, $\sum_{a \in \mathcal{Q}_\mathrm{crit}[x]} \pi(a \mid x) \leq \delta$ for all $x \in \mathcal{X}_\mathrm{V}$, where $\mathcal{Q}_\mathrm{crit}[x]$ is the $\mathcal{X}$-slice of $\mathcal{Q}_\mathrm{crit}$ in x. This is equivalent to Definition 2 up to the choice of δ, since a δ-safe policy satisfies $\sum_{a \in \mathcal{Q}_\mathrm{crit}[x]} \pi(a \mid x) \leq \delta \cdot |\mathcal{Q}_\mathrm{crit}[x]|$.

In the next section, we consider an RL problem over the set of safe policies and dual relaxations thereof. To allow for general such relaxations, we introduce dynamic indicators.

Definition 3 (Dynamic indicator). *Let $c : \mathcal{Q} \to \mathbb{R}_{\geq 0}$ and the associated discounted risk*

$$\rho(x, \pi) = \sum_{t=0}^{\infty} \gamma^t c(X_t, A_t). \tag{8}$$

We say that c is a dynamic indicator of $\mathcal{X}_\mathrm{C}$ if, for all $x \in \mathcal{X}_\mathrm{V}$, $\mathbb{E}[\rho(x, \pi)] > 0$ if, and only if, $\pi \notin \Pi_\mathrm{V}(x)$.

The notion is independent of $\gamma \in (0, 1)$. A simple example is the composition of the indicator function of $\mathcal{X}_\mathrm{C}$ with the dynamics; it is a dynamic indicator of $\mathcal{X}_\mathrm{C}$ [3, Lemma 1]. While this one is always available, more elaborate dynamic indicators help penalize unviable states earlier in the Lagrangian relaxation and lower required penalties, eventually leading to better conditioning.

Remark 2 (Recovering from constraints violation). Our results hold in the two settings where visiting $\mathcal{X}_\mathrm{C}$ terminates the episode or not. The second case is fully consistent with the setup of infinite time-horizon RL that precedes. Then, actions taken from $\mathcal{X}_\mathrm{C}$ may map back into $\mathcal{X}_\mathrm{V}$: trajectories leaving $\mathcal{X}_\mathrm{V}$ may only return there *after* visiting $\mathcal{X}_\mathrm{C}$. We even have $\mathcal{X}_\mathrm{C} \cap \mathcal{X}_\mathrm{V} \neq \emptyset$ in general, and the intersection is composed of states with actions that map in $\mathcal{X}_\mathrm{V} \setminus \mathcal{X}_\mathrm{C}$. The first case, however, is not naturally framed in infinite time-horizon. Indeed, while adding an absorbing state with null reward and dynamic indicator as in [3] effectively cuts the sums in $G(x, \pi)$ and $\rho(x, \pi)$, the sum in $S(x, \pi)$ cannot be handled similarly without additional notation. In the interest of conciseness and clarity, we thus only introduce formally the case of non-terminal $\mathcal{X}_\mathrm{C}$. We emphasize that this is the more challenging case, as forbidding entropy collection after failure effectively further penalizes failure states.

4 Problem formulation

We consider a standard constrained RL problem with dynamics f, constraint set $\mathcal{X}_{\mathrm{C}}$, viability kernel $\mathcal{X}_{\mathrm{V}}$, return G, and entropy regularization with temperature $\alpha > 0$, as defined in Section 3:

$$\max_{\pi \in \Pi_{\mathrm{V}}} \bar{G}(x, \pi) + \alpha \bar{S}(x, \pi). \tag{9}$$

We investigate the following questions:

Question 1. In what sense can we interpret (9) as a robust control problem?

Question 2. Can we make (9) amenable to unconstrained algorithms?

We provide an empirical answer to Question 1 by identifying that the constraints repel trajectories of optimal controllers to an extent controlled by α, using tools from viability theory. The higher α, the stronger the repulsion. We then interpret this repulsion as a form of robustness to action noise, as the mode of the solution to (9) favors visiting states where adversarial action noise takes longer to bring the agent to states with constraints. We support this high-level interpretation with empirical demonstrations on toy examples and standard RL benchmarks. We then answer Question 2 through constraints penalties: we show that the solutions of (9) are approximated arbitrarily closely by solving a Lagrangian relaxation of the constraint $\pi \in \Pi_{\mathrm{V}}$. Provided that one can solve the resulting unconstrained problem in practice (using for instance classical RL algorithms such as SAC), our results provide a model-free way to approximate robustly-safe controllers arbitrarily closely with a tunable degree of robustness, as well as a clear interpretation of the temperature and penalty parameters.

5 Theoretical results

In this section, we explain on a high level why entropy regularization causes constraints to repel trajectories of optimal controllers and state our theoretical results on how to approximate (9) with a classical unconstrained problem. The proofs are in Appendix D.

5.1 Preserving future viable options

Explanation Our starting point to understand the claimed phenomenon is the observation that, for $x \in \mathcal{X}_{\mathrm{V}}$, the maximum immediate entropy achievable by a *safe* controller is limited by the number of unsafe actions available in x. Specifically, it follows immediately from properties of $\mathcal{H}$ that

$$\forall \pi \in \Pi_{\mathrm{V}},\ \mathcal{H}(\pi(\cdot \mid x)) \leq \ln|\mathcal{Q}_{\mathrm{V}}[x]|. \tag{10}$$

Since $\bar{S}$ is the (expected discounted) sum of the left-hand side of (10) along trajectories, it is meaningful that entropy-regularized, safe optimal controllers

avoid states for which this upper bound is low, i.e., where $|\mathcal{Q}_{\mathrm{V}}[x]|$ is low. On the other hand, completely forbidding actions leading to such states is also harmful, since it "propagates" the constraints backwards along trajectories, enforcing a similar upper bound on the immediate entropy obtainable in those previous states as well. In other words, entropy-regularized controllers limit the probability of actions that eventually lead to states with a low bound in (10), without completely avoiding such actions to avoid loss of immediate entropy. The more steps it takes to reach states with many constraints, the less pronounced this effect of the constraints is. It follows from this reasoning that trajectories that go towards states with many constraints generally have lower probability than trajectories that go away from them.

This discussion supports on a high level that entropy regularization with constraints promotes constraints avoidance by preserving the long-term number of future viable options. Next, we identify this behavior as a form of robustness to action noise of the mode policy. Indeed, the mode policy tends to minimize the long-term proportion of actions unavailable because of constraints, and thus the probability that action noise selects such an action is also approximately minimized. We leave a precise formalization of this idea to future work, and support it with empirical evidence in Section 6.

A metric of robustness This discussion highlights that, for any $\pi \in \Pi_{\mathrm{V}}$ and $x \in \mathcal{X}_{\mathrm{V}}$, the quantity $\bar{S}(x, \pi)$ captures the *long-term number of viable actions* that π considers from x. A controller achieving a high $\bar{S}(x, \pi)$ successfully avoids highly-constrained states. This motivates taking the cumulative entropy as a quantitative measurement of robustness, which enables comparing the robustness of controllers.

Definition 4. *We say that* $\pi_1 \in \Pi_{\mathrm{V}}$ *is* less S-robust *than* $\pi_2 \in \Pi_{\mathrm{V}}$*, and write* $\pi_1 \preceq \pi_2$*, if*

$$\bar{S}(x, \pi_1) \leq \bar{S}(x, \pi_2), \quad \forall x \in \mathcal{X}_{\mathrm{V}}. \tag{11}$$

Behavior for increasing temperatures For $\alpha = 0$, (9) recovers the constrained, unregularized problem

$$\max_{\pi \in \Pi_{\mathrm{V}}} \bar{G}(x, \pi). \tag{12}$$

We are then maximizing the return over viable policies with no concerns about robustness. As α increases, entropy is more and more prevalent in the objective of (9), whose solution converges to the maximum entropy policy $\pi^\star_{\mathrm{ent}}$

$$\pi^\star_{\mathrm{ent}} = \arg\max_{\pi \in \Pi_{\mathrm{V}}} \bar{S}(x, \pi), \quad \forall x \in \mathcal{X}_{\mathrm{V}}. \tag{13}$$

This is best seen through the soft-value function.

Theorem 1. *Consider the soft-Q-value functions* Q_{ent} *and* Q_α *of* (13) *and* (9)*, respectively and for all* $\alpha \in \mathbb{R}_{\geq 0}$*. Then,* $\max_{\mathcal{Q}_{\mathrm{V}}} |\frac{1}{\alpha} Q_\alpha - Q_{\mathrm{ent}}| \to 0$ *as* $\alpha \to \infty$*.*

Corollary 1. *Denote by* $\pi^\star_{\mathrm{ent}}$ *and* $\pi^\star_\alpha$ *the solutions of* (13) *and* (9)*, respectively and for all* $\alpha \in \mathbb{R}_{\geq 0}$*. Then, the map* $\alpha \mapsto \pi^\star_\alpha$ *is monotonic for* $\preceq$ *and* $\max_{\mathcal{Q}_{\mathrm{V}}} |\pi^\star_\alpha - \pi^\star_{\mathrm{ent}}| \to 0$ *as* $\alpha \to \infty$*.*

Corollary 1 formalizes that the solution of (9) becomes more S-robust and approaches that of (13) as α increases. The mode of (9) thus gets robust to action noise by the preceding explanations and empirical evidence of Section 6.

5.2 Relaxing safety constraints with penalties

A practical consequence of our observation is that solving (9) yields controllers that preserve a safe distance to the constraints with high probability. An important drawback, however, is that the problem involves the viable set $\mathcal{Q}_{\mathrm{V}}$, which is unknown in model-free situations. We now leverage a Lagrangian relaxation of these viability constraints to make the problem amenable to model-free algorithms. The results in this section extend those of [3] to the case of an entropy-regularized objective.

In this section, we consider c a dynamic indicator function of $\mathcal{X}_{\mathrm{C}}$ and ρ the associated discounted risk (Definition 3). We are interested in the following penalized problem

$$\pi^{\star}_{\alpha,p} = \arg\max_{\pi \in \Pi} \bar{G}(x,\pi) + \alpha \bar{S}(x,\pi) - p\rho(x,\pi), \tag{14}$$

where $p \in \mathbb{R}_{\geq 0}$ is a penalty parameter. It is known that in the case $\alpha = 0$, (14) and (9) share the same solutions if p is large enough [3, Theorem 2]. Unfortunately, this result does not directly carry to the case $\alpha > 0$: from (6), $\pi^{\star}_{\alpha,p}(a \mid x) > 0$ for all $(x,a) \in \mathcal{Q}$, and thus in particular $\pi^{\star}_{\alpha,p} \notin \Pi_{\mathrm{V}}$. However, scaling the penalty remains possible if one accepts to trade viability for δ-safety.

Theorem 2. *For any $\delta > 0$, $\epsilon > 0$, and $\alpha > 0$, there exists $p^{\star} \in \mathbb{R}_{\geq 0}$ such that, for all $p > p^{\star}$, the optimal policy of (14) $\pi^{\star}_{\alpha,p}$ is δ-safe and*

$$\max_{\mathcal{Q}_{\mathrm{V}}} |\pi^{\star}_{\alpha,p} - \pi^{\star}_{\alpha}| < \epsilon. \tag{15}$$

Proof (Sketch of proof). The penalty enforces an upper-bound on the soft Q-value of state-actions in $\mathcal{Q}_{\mathrm{crit}}$ (Lemma 2). Values there thus decrease arbitrarily low as the penalty increases, while it remains lower-bounded on $\mathcal{Q}_{\mathrm{V}}$. This, in turn, shows δ-safety of $\pi^{\star}_{\alpha,p}$ for p large enough. Therefore, the value function of (14) approximates to that of (9) on $\mathcal{Q}_{\mathrm{V}}$, and $\pi^{\star}_{\alpha,p}$ gets arbitrarily close to $\pi^{\star}_{\alpha}$.

5.3 Safe policies from the relaxed problem

It follows from Theorem 2 that the solution $\pi^{\star}_{\alpha,p}$ to the penalized problem (14) is a δ-safe policy and the *mode* of $\pi^{\star}_{\alpha,p}$ is safe if the penalty is sufficient.

Corollary 2. *Under the same notations as Theorem 2, there exists $\bar{\delta} \in (0,1)$ such that, if $\delta \in (0,\bar{\delta})$, then the policy $\hat{\pi}_{\alpha,p}$ following the mode (7) of $\pi^{\star}_{\alpha,p}$ is safe.*

Proof. This directly follows from Theorem 2.

We empirically investigate the robustness of this policy in the next section.

Conclusion on the questions Finally, we are able to answer Questions 1 and 2 based on the following arguments. Entropy regularization in the presence of constraints biases the learning problem towards policies that avoid constraints to preserve a high number of viable options, with the temperature coefficient monotonically controlling the degree of S-robustness. Furthermore, the viability constraints of (9) can be relaxed by a Lagrangian formulation at the price of trading viability for δ-safety. Specifically, the solution of (14) approximates arbitrarily closely that of (9), provided that the penalty is sufficiently high. In particular, penalties above a finite threshold recover the mode of (9) exactly and the policy following that mode is therefore safe. Put together, these results provide a model-free way to approximate safe and robust controllers with tunable degrees of robustness.

6 Empirical results

We provide in this section the empirical evidence that entropy regularization with constraints yields policies whose mode avoids constraints and is robustly safe under increased action noise, and that penalties enable approximating these constraints. We start with a discrete grid world, where we can solve the constrained problem numerically, to showcase how constraints repel trajectories in the presence of entropy regularization. Second, we introduce failure penalties to reveal how they recover the constraints. Finally, we illustrate the claimed robustness to increased action noise on MuJoCo benchmarks[3]. These experimental results confirm our interpretation of the two hyperparameters: penalties control the probability of failure, while the temperature controls the degree of robustness.

6.1 Cliff walking

Our gridworld (Fig. 1) is an adaptation of the cliff environment [23, Example 6.6]. Three states in the middle of the bottom row represent the cliff; the failure set $\mathcal{X}_{\mathrm{C}}$ the agent should robustly avoid. The right column represents the target of escaping the cliff. The failure and target states are invariant under all actions. Otherwise, the dynamics follow the direction of the chosen action, or map back into the current state if the agent hits a border. Actions outside of the cliff or target get a -1 reward.

Interaction of constraints and entropy The constrained version of the environment — the fenced cliff — only offers three actions to an agent neighboring the cliff, imposing a lower upper-bound on the entropy in those states as per (10). This observation is key in understanding why entropy regularization avoids states with unviable actions, yielding robustness (Fig. 1.d).

[3] The code to reproduce results is available at `www.github.com/Data-Science-in-Mechanical-Engineering/entropy_robustness`.

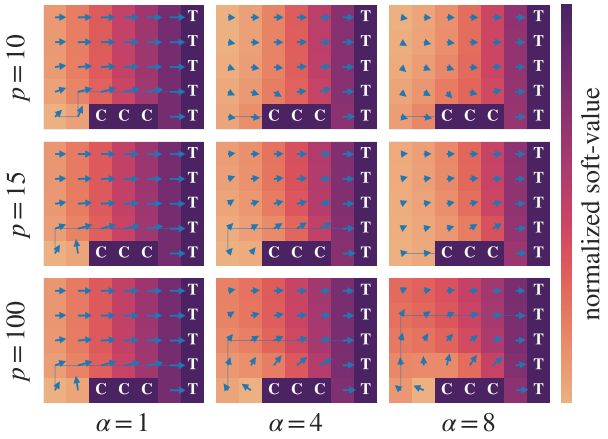


Fig. 2. Unconstrained cliff — Safety and robustness as functions of α and p: Safety and robustness can be achieved by penalizing (p) the constraints $\mathcal{X}_C$ and adjusting the temperature (α).

Indeed, when maximizing entropy only (Fig. 1.d), the optimal policy favors transitioning away from states neighboring the constraints due to the aforementioned upper bound on immediate entropy. In turn, the immediate entropy of the policy in the 2-step neighbors is also reduced since some transitions are less desirable. The same logic applies recursively "outwards" from states with unviable actions, and the policy generally pushes trajectory away from the constraints; that is, towards the top corners. When initialized on the right, the policy aims at reaching the invariant target states where full entropy is available. When initialized on the left, the mode of the policy favors the top-left corner to avoid the low entropy of states close to the constraints, overcoming the long-term benefit of the target state. This trade-off between short- and long-term entropy depends on the discount factor γ.

In contrast, finite temperatures (Fig. 1.a–c) further encourage reaching the goal state to avoid the negative reward. The agent thus takes more risks to collect rewards while preserving some distance from the constraints. This trade-off between performance and robustness is controlled by the temperature parameter α: high values favor entropy (and, thus, robustness by what precedes), whereas lower ones favor performance. While high robustness may be desirable, it comes at the price of suboptimality. Too high a temperature may entirely prevent task completion for the mode policy if the path thereto is inherently risky, leading to unsuccessful learning outcomes due to poor choice of hyperparameters.

Interaction of penalties and entropy Sufficient penalties enable solving the constrained problem (Fig. 2), consistently with Theorem 2. The example shows the robustness–performance trade-off with different temperatures and penalties. Importantly, entropy and penalties are now competing, and any fixed penalty is

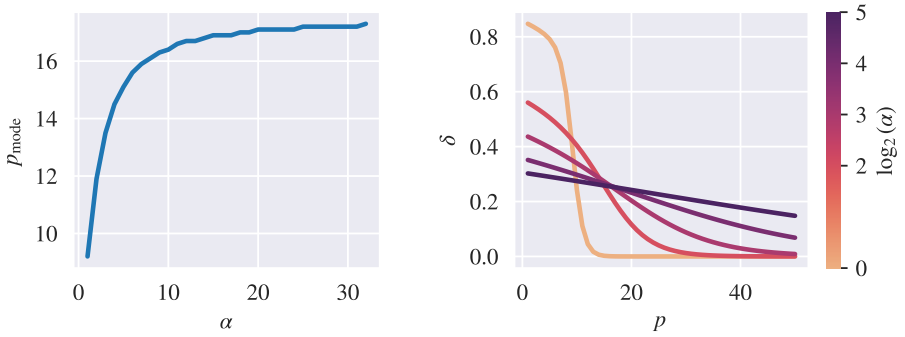


Fig. 3. Effect of the temperature on the minimum safe penalty p_{mode} (left) and δ-safety (right) on the cliff: *Left:* The minimum penalty such that the mode of the stochastic policy is safe. *Right:* The minimum δ such that the policy is δ-safe as functions of p and α. Policies get safer as p increases, but less safe as α does.

eventually overcome by high temperatures, degrading safety (Fig. 3). The penalty thus needs to scale with the temperature to ensure δ-safety with a low δ.

The minimum sufficient value for the penalty depends not only on α, but also on other hyperparameters such as the reward function, discount factor, and dynamic indicator. For instance, if the dynamic indicator is simply the indicator function of the constraints set, then the minimum penalty scales exponentially with the longest trajectory contained in $\mathcal{X} \setminus (\mathcal{X}_{\text{V}} \cup \mathcal{X}_{\text{C}})$. Other choices of dynamic indicators may improve this dependency by incurring the penalty earlier in the trajectory, but choosing the penalty remains a problem-specific concern. While theory suggests picking it as high as possible, too high a penalty may introduce numerical instabilities when combined with value function approximators outside of tabular methods. We refer to [3] for an extended discussion.

6.2 Reinforcement learning benchmarks

We now illustrate on standard RL benchmarks that this constraints avoidance translates into increased robustness to action noise. For this, we train entropy-regularized agents on two popular MuJoCo benchmarks, namely the `Pendulum-v1` and the `Hopper-v4` environment [27], with various temperatures. We then evaluate the mode of the learned policy under additional external action noise, whereas training is noise-free. The action noise is sampled from a uniform distribution $\mathcal{U}(-\epsilon, \epsilon)$. For each value of the temperature, we evaluate the frequency of successful constraints avoidance over 100 episodes. Further details on the setup are in Appendix B and additional results are in Appendix A.

Consistently with our theoretical results, we find that (i) entropy-regularization decreases the return by avoiding high-value states with many unviable actions; and (ii) the mode of entropy-regularized policies is more robust to disturbances as the training temperature increases.

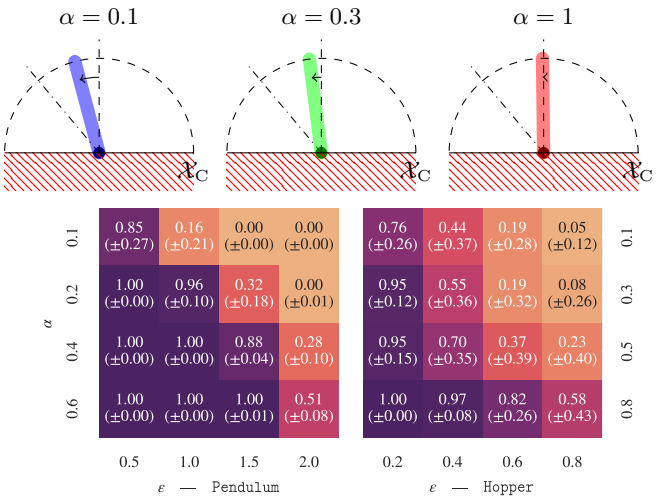


Fig. 4. Learning robust policies with SAC *Top:* With a target angle at 40° (dashed-dotted line) the agent learns to stabilize at different angles depending on the training temperature. For higher temperatures, the agent stabilizes the pendulum further away from the failure set $\mathcal{X}_C$. *Bottom:* Rate of successful failure avoidance on the disturbed `Pendulum-v1` (left heat map) and `Hopper-v4` (right heat map) environments. As the temperature increases the mode of the stochastic policy is robust to higher levels of action noise ϵ.

Pendulum We modify the `Pendulum-v1` environment as follows to incorporate robustness concerns: (i) the initial state is the still, upright position; (ii) the constraints consist of angles with magnitude beyond 90° and the penalty is 90; and (iii) the reward is the squared angular difference to a target angle of 40°, which is outside of the viability kernel since the agent exerts bounded torque.

The results are shown in Fig. 4. All policies lean towards the target state but avoid leaving the viability kernel and reaching the constraints. The sufficient penalty emulates the boundary of the viability kernel, which reduces the effective number of available actions when leaning to one side. This pushes entropy-regularized policies away from the target state, and they learn to stabilize angles closer to 0 as α increases — the maximum entropy policy keeps the pendulum upright. The results show a robustness–performance trade-off between staying upright and leaning as far as possible towards the target, which is controlled by the temperature α. Furthermore, the mode of the entropy-regularized policy can cope with significantly higher action disturbances when trained with higher temperatures.

Hopper We repeat the same experiment as in the previous section for a modified `Hopper-v4` environment [27]. We modify the environment by penalizing the "unhealthy" states with a penalty of $p = 500$. The results are shown in Fig. 4. Increasing the temperature improves the robustness to additional action noise.

However, the learned gait is slower, hinting at a performance–robustness trade-off for this environment (see Appendix A.2). Interestingly, as the temperature is increased, the training finds two distinct robust behaviors: one is the intended hopping forward; the other is standing still and only collecting the healthy reward, which is arguably the most robust behaviour.

Our experiments inform hyperparameter settings for RL practitioners: while entropy regularization leads to robustly safe policies, high temperatures (or minimum entropy constraints [28]) can make parts of the state space unreachable, lead to conservative policies, and may even entirely prevent task completion as seen in the Hopper example.

7 Conclusion

We study the interaction between entropy regularization and state constraints in RL and reveal empirically that this favors policies that are constraints-avoiding and robust to increased action noise, as they preserve an expected long-term number of viable actions. We also show both in theory and in practice how to approximate the constraints with failure penalties. In particular, the mode of the policy — which is often what is deployed after training completion — is recovered exactly by penalties above a finite threshold.

The connection between entropy regularization with constraints and control-theoretic robustness is novel, to the best of our knowledge. This study identifies the phenomenon, its relevance for RL, and opens many interesting avenues for future work. A particularly promising one is the systematic study of the identified robustness. Indeed, we hypothesize that entropy regularization with constraints induces a kind of *soft constraints tightening*; that is, restricts the optimization domain to controllers that go away from the constraints with at least some given probability. Such a result would enable identifying "softly invariant sets": subsets of the viability kernel that are control invariant under a robustly safe controller (but not directly under entropy-regularized controllers, as they have full action support) and contain the entropy-regularized controller's trajectory with high probability. This would draw a clear theoretical bridge between entropy-regularized constrained RL and robust control through constraints tightening. An alternative would be identifying a noise model to which the modes of entropy-regularized, constrained policies are robust. More generally, it would be interesting to find other regularization terms that promote robustness and that are amenable to RL beyond the cumulative entropy, following ideas from [7]. Such regularizers could enable novel robustness properties with rigorous guarantees, and perhaps help training policies that are less sensitive to the sim-to-real gap.

In the meantime, we expect our findings to inform practitioners when applying RL algorithms such as SAC. While entropy regularization has mainly been developed as an exploration mechanism [2], it biases the policy to robustness to action noise if one uses constraints penalties. This understanding enables principled decisions when tuning the temperature and penalties, for instance by

discouraging the common practice of annealing the temperature if robustness is a concern.

Acknowledgments. The authors thank Zeheng Gong for help with the empirical results. Simulations were performed with computing resources granted by RWTH Aachen University under project rwth1626. This work has been supported by the Robotics Institute Germany, funded by BMBF grant 16ME0997K.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Moos, J., Hansel, K., Abdulsamad, H., Stark, S., Clever, D., Peters, J.: Robust reinforcement learning: A review of foundations and recent advances. Machine Learning and Knowledge Extraction **4**(1), 276–315 (2022). https://doi.org/10.3390/make4010013
2. Haarnoja, T., Zhou, A., Abbeel, P., 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)
3. Massiani, P.F., Heim, S., Solowjow, F., Trimpe, S.: Safe Value Functions. IEEE Transactions on Automatic Control (2023)
4. Haarnoja, T., Pong, V., Zhou, A., Dalal, M., Abbeel, P., Levine, S.: Composable deep reinforcement learning for robotic manipulation. In: IEEE International Conference on Robotics and Automation. pp. 6244–6251 (2018). https://doi.org/10.1109/ICRA.2018.8460756
5. Haarnoja, T., Ha, S., Zhou, A., Tan, J., Tucker, G., Levine, S.: Learning to walk via deep reinforcement learning. In: Proceedings of Robotics: Science and Systems (2019). https://doi.org/10.15607/RSS.2019.XV.011
6. Eysenbach, B., Levine, S.: Maximum entropy RL (provably) solves some robust RL problems. In: International Conference on Learning Representations (2022)
7. Geist, M., Scherrer, B., Pietquin, O.: A theory of regularized Markov decision processes. In: Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 2160–2169 (2019)
8. Heim, S., Badri-Spröwitz, A.: Beyond basins of attraction: Quantifying robustness of natural dynamics. IEEE Transactions on Robotics **35**(4), 939–952 (2019)
9. Heim, S., Rohr, A., Trimpe, S., Badri-Spröwitz, A.: A Learnable Safety Measure. In: Conference on Robot Learning. pp. 627–639. PMLR (May 2020)
10. Aubin, J.P., Bayen, A.M., Saint-Pierre, P.: Viability theory: new directions. Springer Science & Business Media (2011)
11. Brunke, L., Greeff, M., Hall, A.W., Yuan, Z., Zhou, S., Panerati, J., Schoellig, A.P.: Safe learning in robotics: From learning-based control to safe reinforcement learning. Annual Review of Control, Robotics, and Autonomous Systems **5**, 411–444 (2022)
12. Achiam, J., Held, D., Tamar, A., Abbeel, P.: Constrained policy optimization. In: International conference on machine learning. pp. 22–31. PMLR (2017)
13. Kerrigan, E.C., Maciejowski, J.M.: Soft constraints and exact penalty functions in model predictive control. In: Control 2000 Conference, Cambridge. pp. 2319–2327 (2000)
14. Xing, A.Q., Wang, C.L.: Applications of the exterior penalty method in constrained optimal control problems. Optimal Control Applications and Methods **10**(4), 333–345 (1989)

15. Zhou, K., Doyle, J., Glover, K.: Robust and optimal control. Prentice Hall (1996)
16. Grüne, L., Pannek, J.: Nonlinear Model Predictive Control. Springer, 2 edn. (2017)
17. Limon, D., Alamo, T., Raimondo, D.M., De La Peña, D.M., Bravo, J.M., Ferramosca, A., Camacho, E.F.: Input-to-state stability: a unifying framework for robust model predictive control. Nonlinear Model Predictive Control: Towards New Challenging Applications pp. 1–26 (2009)
18. Bansal, S., Chen, M., Herbert, S., Tomlin, C.J.: Hamilton-jacobi reachability: A brief overview and recent advances. In: Conference on Decision and Control. pp. 2242–2253 (2017)
19. Calafiore, G.C., Campi, M.C.: The scenario approach to robust control design. IEEE Transactions on automatic control **51**(5), 742–753 (2006)
20. Morimoto, J., Doya, K.: Robust reinforcement learning. Neural Computation **17**(2), 335–359 (2005). https://doi.org/10.1162/0899766053011528
21. Pinto, L., Davidson, J., Sukthankar, R., Gupta, A.: Robust adversarial reinforcement learning. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 2817–2826 (2017)
22. Tessler, C., Efroni, Y., Mannor, S.: Action robust reinforcement learning and applications in continuous control. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 6215–6224 (2019)
23. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
24. Fox, R., Pakman, A., Tishby, N.: Taming the noise in reinforcement learning via soft updates. In: 32nd Conference on Uncertainty in Artificial Intelligence. pp. 202–211 (2016)
25. Nachum, O., Norouzi, M., Xu, K., Schuurmans, D.: Bridging the gap between value and policy based reinforcement learning. Advances in neural information processing systems **30** (2017)
26. Massiani, P.F., Heim, S., Trimpe, S.: On exploration requirements for learning safety constraints. In: Learning for Dynamics and Control. pp. 905–916. PMLR (2021)
27. Towers, M., Terry, J.K., Kwiatkowski, A., Balis, J.U., Cola, G.d., Deleu, T., Goulão, M., Kallinteris, A., KG, A., Krimmel, M., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J.J., Shen, A.T.J., Younis, O.G.: Gymnasium (Mar 2023). https://doi.org/10.5281/zenodo.8127026, https://zenodo.org/record/8127025
28. Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al.: Soft actor-critic algorithms and applications. arXiv preprint arXiv:1812.05905 (2018)
29. Huang, S., Dossa, R.F.J., Ye, C., Braga, J., Chakraborty, D., Mehta, K., Araújo, J.G.: Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. Journal of Machine Learning Research **23**(274), 1–18 (2022), http://jmlr.org/papers/v23/21-1342.html
30. Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 5026–5033. IEEE (2012)