Average-Reward Soft Actor-Critic

Anonymous authors
Paper under double-blind review

Keywords: average-reward, MaxEnt, entropy-regularization, actor-critic, deep RL.

Summary

The average-reward formulation of reinforcement learning (RL) has drawn increased interest in recent years for its ability to solve temporally-extended problems without relying on discounting. Meanwhile, in the discounted setting, algorithms with entropy regularization have been developed, leading to improvements over deterministic methods. Despite the distinct benefits of these approaches, deep RL algorithms for the entropy-regularized average-reward objective have not been developed. While policy-gradient based approaches have recently been presented for the average-reward literature, the corresponding actor-critic framework remains less explored. In this paper, we introduce an average-reward soft actor-critic algorithm to address these gaps in the field. We validate our method by comparing with existing average-reward algorithms on standard RL benchmarks, achieving superior performance for the average-reward criterion.

Contribution(s)

1. We generalize the soft actor-critic (SAC) algorithm from the discounted to the average-reward setting.
   Context: Haarnoja et al. (2018b) derived a MaxEnt RL algorithm, soft actor-critic, for the discounted setting. We derive theoretical results and implement new algorithmic techniques to adapt SAC to the average-reward setting.
2. We extend the policy improvement theorem to the entropy-regularized average-reward objective.
   Context: Previous work demonstrated the policy improvement theorem separately in discounted MaxEnt RL Haarnoja et al. (2018b) and average-reward (un-regularized) RL Zhang & Tan (2024). We close this gap by analyzing the theoretical properties of policy improvement in the entropy-regularized average-reward setting.
3. We experimentally demonstrate the advantage of our approach against available baselines in standard control environments.
   Context: We compare our algorithm with the state-of-the-art average-reward methods: ARO-DDPG (Saxena et al., 2023), ATRPO (Zhang & Ross, 2021), and APO (Ma et al., 2021).

Average-Reward Soft Actor-Critic

Anonymous authors
Paper under double-blind review

Abstract

The average-reward formulation of reinforcement learning (RL) has drawn increased interest in recent years for its ability to solve temporally-extended problems without relying on discounting. Meanwhile, in the discounted setting, algorithms with entropy regularization have been developed, leading to improvements over deterministic methods. Despite the distinct benefits of these approaches, deep RL algorithms for the entropy-regularized average-reward objective have not been developed. While policy-gradient based approaches have recently been presented for the average-reward literature, the corresponding actor-critic framework remains less explored. In this paper, we introduce an average-reward soft actor-critic algorithm to address these gaps in the field. We validate our method by comparing with existing average-reward algorithms on standard RL benchmarks, achieving superior performance for the average-reward criterion.

1 Introduction

A successful reinforcement learning (RL) agent learns from interacting with its surroundings to achieve desired behaviors, as encoded in a reward function. However, in "continuing" tasks, where the amount of interactions is potentially unlimited, the total sum of rewards received by the agent is unbounded. To avoid this divergence, a popular technique is to *discount* future rewards relative to current rewards. The framework of discounted RL enjoys convergence properties (Sutton & Barto, 2018; Kakade, 2003; Bertsekas, 2012), practical benefits (Schulman et al., 2016; Andrychowicz et al., 2020), and a plethora of useful algorithms (Mnih et al., 2015; Schulman et al., 2015; 2017; Hessel et al., 2018; Haarnoja et al., 2018b) making the discounted objective an obvious choice for the RL practitioner. Despite these benefits, the use of discounting introduces a (typically unphysical) hyperparameter γ which must be tuned for optimal performance. The difficulty in properly tuning the discount factor γ is illustrated in our motivating example, Figure 1. Furthermore, agents solving the discounted RL problem will fail to optimize for long-term behaviors that operate on timescales longer than those dictated by the discount factor, $(1-\gamma)^{-1}$. Moreover, recent work has argued that the discounted objective is not even a well-defined optimization problem (Naik et al., 2019). Importantly, despite most state-of-the-art algorithms operating within this discounted framework, their metric for performance is most often the total or average reward over trajectories, as opposed to the discounted sum, which they are designed to optimize. In such cases, the discounted objective is used as a crutch for optimizing the true object of interest: long-term average performance.

To address these issues, another objective for solving continuing tasks has been defined and studied (Schwartz, 1993; Mahadevan, 1996): the average-reward objective. Although it is arguably a more natural choice, it has less obvious convergence properties since the associated Bellman operators no longer possess the contraction property. Despite an ongoing line of work on the theoretical properties of the average-reward objective (Zhang et al., 2021; Wan, 2023), there remain a limited number of deep RL algorithms for this setting. Current algorithms beyond the tabular or linear settings focus on policy-gradient methods to develop deep actor-based models: (Zhang & Ross, 2021; Ma et al., 2021; Saxena et al., 2023). While these advancements represent a positive step toward solving the average-reward objective, there remains a need for alternative approaches for the problem of average-reward deep RL.

In both the discounted and average-reward scenarios, optimal policies are known to be deterministic (Mahadevan, 1996; Sutton & Barto, 2018). However, under various real-world circumstances (e.g. errors in the model, perception, and control loops), a deterministic policy can fail. In deployment, when RL agents face the sim-to-real gap, are transferred to other environments, or when perturbations arise (Haarnoja et al., 2017; 2018a; Eysenbach & Levine, 2022), fully-trained deterministic agents may be rendered useless. To address these important use-cases, it would be useful to have a stochastic optimal policy which is flexible and robust under uncertainty. Rather than using heuristics (e.g. ε-greedy, mixture of experts, Boltzmann) to generate a stochastic policy *post-hoc*, the original RL problem can be regularized with an entropy-based term that yields an optimal policy which is naturally stochastic. Implementing this entropy-regularized RL objective corresponds to additionally rewarding the agent (in proportion to a temperature parameter, β^{-1}) for using a policy which has a lower relative entropy (Levine, 2018), in the sense of Kullback-Leibler divergence. This formulation of entropy-regularized (often considered in the special case of maximum entropy or "MaxEnt"[1]) RL has led to significant developments in state-of-the-art off-policy algorithms (Haarnoja et al., 2017; 2018b;c).

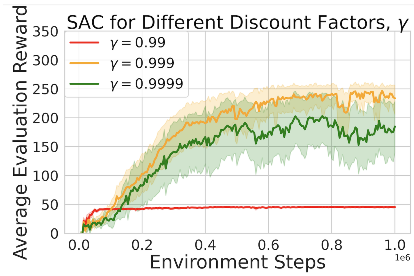

Figure 1: The Swimmer-v5 environment, often not included in Mujoco benchmarks (Franceschetti et al., 2022), is notoriously difficult for discounted methods to solve when the discount factor is not tuned over and set to its default value of $\gamma = 0.99$. Other discount-sensitive examples of environments have been discussed by Tessler & Mannor (2020). We find that after carefully tuning the discount factor, SAC can solve the task, but the solution is quite sensitive to the choice of γ. Each curve corresponds to an average over 30 random seeds, with the standard error indicated by the shaded region.

Despite the desirable features of both the average-reward and entropy-regularized objectives, an empirical study of the combination of these two formulations is limited, and no function-approximator algorithms exist yet for this setting. To address this, we propose a novel algorithm for average-reward RL with entropy regularization which is an extension of the discounted algorithm Soft Actor-Critic (SAC) (Haarnoja et al., 2018b;c).

Notably, our implementation requires minimal changes to common codebases, making it accessible for researchers and allowing for future extensions by the community.

2 Preliminaries

In this section, we discuss the background material necessary for the subsequent discussion. Let $\Delta(\mathcal{X})$ denote the probability simplex over the space $\mathcal{X}$. A Markov Decision Process (MDP) is modeled by a state space $\mathcal{S}$, action space $\mathcal{A}$, reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, transition dynamics $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and initial state distribution $\mu \in \Delta(\mathcal{S})$. The state space describes the set of possible configurations in which the agent (and environment) may exist. (This can be juxtaposed with the "observation" which encodes only the state information accessible to the agent. We will consider fully observable MDPs where state and observation are synonymous.) The action space is the set of controls available to the agent. Enacting control, the agent may alter its state. This change is dictated by the (generally stochastic) transition dynamics, p. At each discrete timestep, an action is taken and the agent receives a reward $r(s, a) \in \mathbb{R}$ from the environment.

We will make some of the usual assumptions for average-reward MDPs (Wan et al., 2021):

Assumption 1. The Markov chain induced by any stationary policy π is communicating.

Assumption 2. The reward function is bounded.

[1]MaxEnt refers to using a uniform prior policy. In that case, "low relative entropy" (with respect to a uniform prior) is equivalent to "high Shannon entropy". In this work, we consider the case of more general priors.

In solving an average-reward MDP, one seeks a control policy π which maximizes the expected *reward-rate*, denoted ρ^π. In the average-reward framework, such an objective reads:

$$\rho^\pi = \lim_{N\to\infty} \frac{1}{N} \mathop{\mathbb{E}}_{\tau\sim p,\pi,\mu} \left[\sum_{t=0}^{N-1} r(\mathbf{s}_t, \mathbf{a}_t) \right], \tag{1}$$

where the expectation is taken over trajectories generated by the dynamics p, control policy π, and initial state distribution μ.

The remaining non-scalar (that is, state-action-dependent) contribution to the value of a policy is called the average-reward differential bias function. Because of its analogy to the Q-function in discounted RL, we follow recent work (Zhang & Ross, 2021) and similarly denote it as:

$$Q^\pi_\rho(\mathbf{s}, \mathbf{a}) = \mathop{\mathbb{E}}_{\tau\sim p,\pi} \left[\sum_{t=0}^{\infty} r(\mathbf{s}_t, \mathbf{a}_t) - \rho^\pi \middle| \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a} \right]. \tag{2}$$

We will now introduce a variation of this MDP framework which includes an entropy regularization term. For notational convenience we refer to entropy-regularized average-reward MDPs as ERAR MDPs. The ERAR MDP constitutes the same ingredients as an average-reward MDP stated above, in addition to a pre-specified prior policy[2] $\pi_0 : \mathcal{S} \to \Delta(\mathcal{A})$ and "inverse temperature", β. The modified objective function for an ERAR MDP now includes a regularization term based on the relative entropy (Kullback-Leibler divergence), so that the agent now aims to optimize the expected *entropy-regularized reward-rate*, denoted θ^π:

$$\theta^\pi = \lim_{N\to\infty} \frac{1}{N} \mathop{\mathbb{E}}_{\tau\sim p,\pi,\mu} \left[\sum_{t=0}^{N-1} r(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{\beta} \log \frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)} \right], \tag{3}$$

$$\pi^*(a|s) = \operatorname*{argmax}_\pi \theta^\pi. \tag{4}$$

Assumption 1 implies the expression in Equation (3) is independent of the initial state-action and ensures the reward-rate is indeed a unique scalar. From hereon, we will simply write $\theta = \theta^{\pi^*}$ for the optimal entropy-regularized reward-rate for brevity. Comparing to Equation (1), this rate is seen to include an additional entropic contribution, the relative entropy between the control (π) and prior (π_0) policies.

Beyond a mathematical generalization from the MaxEnt formulation, the KL divergence term has also found use in behavior-regularized RL tasks, especially in the offline setting (Wu et al., 2019; Zhang & Tan, 2024) and has found growing interest in its application to large language models (LLMs) (Rafailov et al., 2024; Yan et al., 2024).

The corresponding differential entropy-regularized action-value function is then given by:

$$Q^\pi_\theta(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}_t, \mathbf{a}_t) - \theta^\pi + \mathop{\mathbb{E}}_{\tau\sim p,\pi} \left[\sum_{t=1}^{\infty} \left(r(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{\beta} \log \frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)} - \theta^\pi \right) \middle| \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a} \right]. \tag{5}$$

We have used the subscripts of θ and ρ in this section to distinguish the two value functions. In the following, we drop the θ subscript as we focus solely on the entropy-regularized objective. Similar to the notation for the average-reward rate, we make the notation compact, and write $Q(\mathbf{s}, \mathbf{a}) = Q^{\pi^*}_\theta(\mathbf{s}, \mathbf{a})$ as a shorthand.

3 Prior Work

Research on average-reward MDPs has a longstanding history, dating back to seminal contributions by Blackwell (1962) and later Mahadevan (1996), which laid the groundwork for future algorithmic

[2] For convenience we assume that π_0 has support across $\mathcal{A}$, ensuring the Kullback-Leibler divergence is always finite.

and theoretical investigations (Even-Dar et al., 2009; Abbasi-Yadkori et al., 2019; Abounadi et al., 2001; Neu et al., 2017; Wan et al., 2021). Due to their theoretical nature, these studies primarily focused on algorithms within tabular settings or under linear function approximation, possibly explaining the limited work on the average-reward problem in the deep RL community. However, recent work has begun to address this challenge by tackling deep average-reward RL (Zhang & Ross, 2021; Ma et al., 2021; Saxena et al., 2023) with methods based on the policy gradient algorithm (Sutton et al., 1999). Especially when tested on long-term optimization tasks, these studies have demonstrated superior performance of average-reward algorithms in the continuous control Mujoco benchmark (Todorov et al., 2012), compared to their discounted counterparts.

In the deep average-reward RL literature, research has primarily focused on extending known algorithms from the discounted to the average-reward setting. For example, Zhang & Ross (2021) first provided an extension of the on-policy trust region method TRPO (Schulman et al., 2015) to the average-reward domain. To extend the classical discounted policy improvement theorem to this domain, they introduced a novel (double-sided) policy improvement bound based on Kémeny's constant (related to the Markov chain's mixing time). Experimentally, they illustrated the success of ATRPO against TRPO, especially for long-horizon tasks in the Mujoco suite. Shortly thereafter, (Ma et al., 2021) introduced an analogue of PPO (Schulman et al., 2017) for average-reward tasks with an extension of generalized advantage estimation (GAE) and addressing the problem of "value drift", again proving successful in experimental comparisons with PPO. Most recently, Saxena et al. (2023) continued this line of work by extending DDPG (Lillicrap et al., 2016) to the average-reward domain with extensive supporting theory, including finite-time convergence analysis. The authors also demonstrate the improved performance of their algorithm, ARO-DDPG, against the previously discussed methods, thereby demonstrating a new state-of-the-art algorithm for the average-reward objective.

In parallel, the discounted objective has included an entropy-regularization term, discussed in works such as (Todorov, 2006; 2009; Ziebart, 2010; Rawlik, 2013; Haarnoja et al., 2017; Geist et al., 2019) which to our knowledge has not yet been introduced in a deep average-reward algorithm. The included "entropy bonus" term in these methods has found considerable use in the development of both theory and algorithms in distinct branches of RL research (Haarnoja et al., 2018a; Eysenbach & Levine, 2022; Park et al., 2023). This innovation yields optimal policies naturally exhibiting stochasticity in continuous action spaces, which has led SAC (Haarnoja et al., 2018c) and its variants to become state-of-the-art solution methods for addressing the discounted objective.

However, there is limited work on the combination of average-reward and entropy-regularized methods, especially for deep RL. Recent work by Rawlik (2013); Neu et al. (2017); Rose et al. (2021); Li et al. (2022); Arriojas et al. (2023); Wu et al. (2024) set the groundwork for combining the entropy-regularized and average-reward formulations by providing supporting theory and validating experiments. We will leverage their results to address the problem of deep average-reward RL with entropy regularization, while introducing some new theoretical results. In the next section, we present our average-reward extension of soft actor-critic.

4 Proposed Algorithm

We begin with a brief discussion of soft actor-critic (SAC), for which we derive new theoretical results and provide an algorithm in the average-reward setting. SAC (Haarnoja et al., 2018b) relies on iteratively calculating a value (critic) of a policy (actor) and improving the actor through soft policy improvement (PI). In the discounted problem formulation, soft PI states that a new policy (denoted π') can be derived from the value function of a previous policy (π) with $\pi' \propto \exp \beta Q^{\pi}(\mathbf{s}, \mathbf{a})$, which is guaranteed to outperform the previous policy in the sense of (soft) Q-values: $Q^{\pi'}(\mathbf{s}, \mathbf{a}) > Q^{\pi}(\mathbf{s}, \mathbf{a})$ for all $\mathbf{s}, \mathbf{a}$ (cf. Lemma 2 of (Haarnoja et al., 2018b) for details). We will first show that an analogous result for policy improvement holds in the ERAR setting. Note that in the case of large state-action spaces, experimentally verifying such inequalities becomes in-tractable (Naik, 2024) and can be alleviated by instead comparing reward rates: scalar quantities which can (in principle) be efficiently evaluated with rollouts.

Since the value of a policy is now encoded in the entropy-regularized average reward rate θ^π and *not* in the differential value, the analogue to policy improvement ($Q^{\pi'} > Q^\pi$) is to establish the bound $\theta^{\pi'} > \theta^\pi$ for some construction of π' from π. Indeed, as we show, the same Boltzmann form over the differential value leads to soft PI in the ERAR objective. We later give some intuition on how this result can be understood as the limit $\gamma \to 1$ of SAC. After establishing PI and the related theory in this setting we will present our algorithm, denoted "ASAC" (for average-reward SAC, and following the naming convention of APO (Ma et al., 2021) and ATRPO (Zhang & Ross, 2021)).

4.1 Theory

As in the discounted case, it can be shown that the Q function for a fixed policy π satisfies a recursive Bellman backup equation[3]. This proposition was also derived in the concurrent work of Wu et al. (2024) which analyzed the ERAR problem in the inverse RL framework:

Proposition 1. *Let an ERAR MDP with reward function $r(\mathbf{s}, \mathbf{a})$, policy π and prior policy π_0 be given. Then the differential value of π, denoted $Q^\pi(\mathbf{s}_t, \mathbf{a}_t)$, satisfies*

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) - \theta^\pi + \mathop{\mathbb{E}}_{\mathbf{s}_{t+1} \sim p} V^\pi(\mathbf{s}_{t+1}), \tag{6}$$

with the entropy-regularized definition of state-value function

$$V^\pi(\mathbf{s}_t) = \mathop{\mathbb{E}}_{\mathbf{a}_t \sim \pi} \left[Q^\pi(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{\beta} \log \frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)} \right]. \tag{7}$$

For completeness, we give a proof of this result (and all others) in the Appendix. As in the discounted case, the proof exploits the recursive structure of Eq. (5).

As mentioned above, in the average reward formulation, the metric of interest is the reward-rate. Our policy improvement result thus focuses on increases in θ^π, generalizing the recent work of Zhang & Ross (2021) to the entropy-regularized setting. We find that the gap between any two entropy-regularized reward-rates can be expressed in the following manner:

Lemma 1 (ERAR Rate Gap)**.** *Consider two policies π, π' absolutely continuous w.r.t. π_0. Then the gap between their corresponding entropy-regularized reward-rates is:*

$$\theta^{\pi'} - \theta^\pi = \mathop{\mathbb{E}}_{\substack{\mathbf{s}_t \sim d_{\pi'} \\ \mathbf{a}_t \sim \pi'}} \left(A^\pi(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{\beta} \log \frac{\pi'(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)} \right), \tag{8}$$

where $A^\pi(\mathbf{s}_t, \mathbf{a}_t) = Q^\pi(\mathbf{s}_t, \mathbf{a}_t) - V^\pi(\mathbf{s}_t)$ is the advantage function of policy π and $d_{\pi'}$ is the steady-state distribution induced by π'.

As a consequence of this result, we find that with the proper choice of the updated policy π', the right-hand side of Equation (8) is guaranteed to be positive, implying that soft PI holds. Using the Boltzmann form of a policy (Haarnoja et al., 2018b) with the differential Q-values as the energy function and the appropriate prior distribution (π_0), gives the desired result:

[3] Equation (7) is an extension of V^π_{soft} in (Haarnoja et al., 2017) to the case of non-uniform prior policy.

Theorem 1 (ERAR Policy Improvement)**.** *Let a policy π absolutely continuous w.r.t. π_0 and its corresponding differential value $Q^\pi(\mathbf{s}_t, \mathbf{a}_t)$ be given. Then, the policy*

$$\pi'(\mathbf{a}_t|\mathbf{s}_t) \doteq \frac{\pi_0(\mathbf{a}_t|\mathbf{s}_t)e^{\beta Q^\pi(\mathbf{s}_t,\mathbf{a}_t)}}{\int e^{\beta Q^\pi(\mathbf{s}_t,\mathbf{a}_t)} d\pi_0(\mathbf{a}_t|\mathbf{s}_t)} \tag{9}$$

achieves a greater entropy-regularized reward-rate. That is, $\theta^{\pi'} \geq \theta^\pi$, with equality only at convergence, when $\pi' = \pi = \pi^*$.

Upon convergence, Equation (8) is identically zero, with the optimal policy satisfying $\pi^* \propto \exp \beta A^*(\mathbf{s}_t, \mathbf{a}_t)$ as expected from the analogous discounted result. We note that the corresponding result in Lemma 2 of Haarnoja et al. (2018b) for SAC (which uses a uniform prior policy), involves the *total* value function. On the other hand, under the average-reward objective, the improved policy is calculated with the *differential* value function. Intuitively, this result can be understood as the $\gamma \to 1$ limit of PI for SAC. Numerically, this can be seen as setting $\gamma = 1$ and continuously subtracting the "extensive" contribution to the total value function throughout. This bulk contribution scales with the number of timesteps in an episode and is the result of accruing a per-timestep reward θ^π. Since the same term accrues in the state- and action-value functions, it cancels in the numerator and denominator of Equation (9). In the case of SAC, the bulk contribution (essentially $N\theta^\pi$, for $N \gg 1$) is included in the value function and so a discount factor $\gamma < 1$ is required to ensure that the total value function is bounded in the limit of large N (in the sense of Equation (3)). In contrast, for the case of ASAC, the bulk contribution is automatically excluded from the corresponding evaluation (by definition), and the differential value function remains bounded in the limit of large N, obviating the need to introduce a discount factor. This intuition can be formalized through a Laurent series expansion; cf. Mahadevan (1996).

To complete the discussion of convergence for ASAC, the policy evaluation (PE) step must also converge. To formulate this, we rely on the work of Wan et al. (2021) who give convergence proofs for average-reward policy evaluation.

Lemma 2 (ERAR Policy Evaluation)**.** *Consider a fixed policy π, for which θ^π of Equation* (1) *has been calculated (e.g. with direct rollouts). The iteration of Equations* (2) *and* (7) *converges to the entropy-regularized differential value of π: $Q^\pi(\mathbf{s}_t, \mathbf{a}_t)$.*

Proof. The proof follows from the convergence results established in the un-regularized case, e.g. Wan et al. (2021). Since the policy π is fixed (and $\pi \ll \pi_0$), the entropic cost $-\beta^{-1}\mathrm{KL}(\pi||\pi_0)$ is finite and can be absorbed into the reward function's definition: $r \leftarrow r - \beta^{-1}\mathrm{KL}(\pi||\pi_0)$, and the standard proof techniques apply. □

4.2 Implementation

As in SAC (Haarnoja et al., 2018b), we propose to interleave steps of policy evaluation (PE) and policy improvement (PI) using stochastic approximation to train the critic and actor networks, respectively. We use a deep neural net with parameters ψ, and denote Q_ψ as the "online" critic network (with trainable parameters), and denote $Q_{\bar{\psi}}$ as the "target" critic, updated periodically through Polyak averaging of the parameters. To implement a PI step, we use the KL divergence loss to update the parameters ϕ of an actor network π_ϕ based on the policy improvement theorem (Equation (9)):

$$\mathcal{L}_\phi = \sum_{\mathbf{s}_t \in \mathcal{B}} \mathrm{KL}\left(\pi_\phi(\cdot|\mathbf{s}_t) \middle|\middle| \frac{\pi_0(\cdot|\mathbf{s}_t)e^{\beta Q_\psi(\mathbf{s}_t,\cdot)}}{Z(\mathbf{s}_t)}\right) . \tag{10}$$

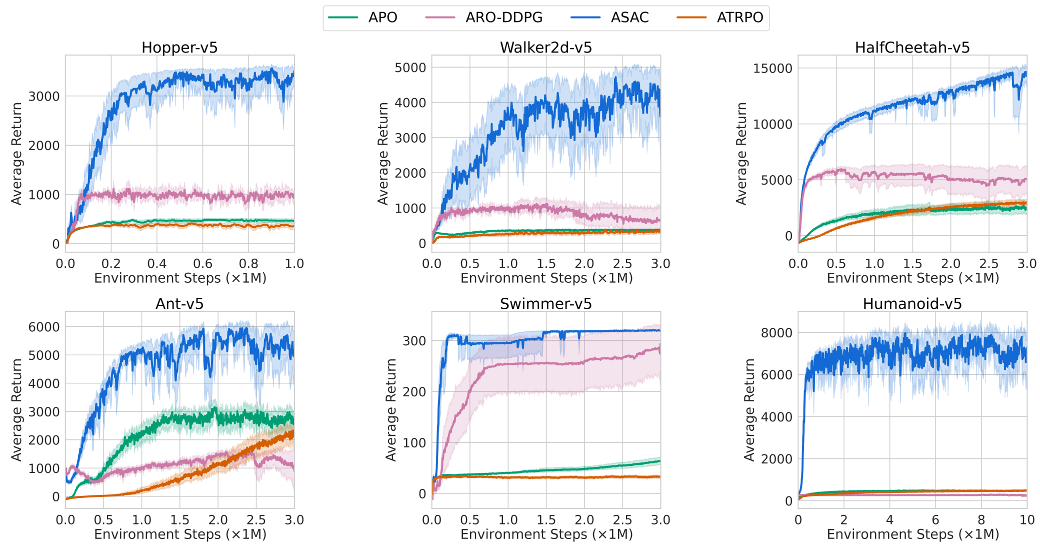

Figure 2: Training curves on continuous control benchmarks. We compare our algorithm, average-reward soft actor-critic (ASAC), with the following baselines: average-reward off-policy deep deterministic policy gradient (ARO-DDPG), average-reward trust-region policy optimization (ATRPO), and average-reward policy optimization (APO). ASAC learns the fastest with the best asymptotic performance. Each curve corresponds to an average over 20 random seeds, with standard errors indicated by the shaded region.

Similar to SAC, the independence of parameters on the partition function Z allows us to simplify this loss expression to the more tractable form:

$$\mathcal{L}_\phi = \sum_{\mathbf{s}_t \in \mathcal{B}} \mathop{\mathbb{E}}_{\mathbf{a}_t \sim \pi_\phi} \left(\log \frac{\pi_\phi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)} - \beta^{-1} Q_\psi(\mathbf{s}_t, \mathbf{a}_t) \right) . \tag{11}$$

In practice, we also use the re-parameterization trick to efficiently propagate gradients through the actor model. After updating the actor via soft policy improvement, we update the critic (differential value) by performing a policy evaluation step with actions sampled from the current actor network. The mean squared error loss is calculated by comparing the expected Q-value to the right-hand side of Equation (6):

$$\mathcal{L}_\psi = \sum_{(\mathbf{s}_t, \mathbf{a}_t, r, \mathbf{s}_{t+1}) \sim \mathcal{B}} \left| Q_\psi(\mathbf{s}_t, \mathbf{a}_t) - \hat{y}(r, \theta; \bar{\psi}, \phi) \right|^2, \tag{12}$$

where $\hat{y}$ is the target value, defined as:

$$\hat{y}(r, \theta; \bar{\psi}, \phi) = r - \theta + \mathop{\mathbb{E}}_{\mathbf{a}_{t+1} \sim \pi_\phi(\cdot|\mathbf{s}_{t+1})} \left[Q_{\bar{\psi}}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) - \frac{1}{\beta} \log \frac{\pi_\phi(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})}{\pi_0(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})} \right].$$

To update the ERAR rate θ^π, we again bootstrap from Eq. (6). Specifically, we treat θ as a trainable parameter (using an Adam optimizer) and train it to minimize the residual error over a batch (using the same mini-batch as above) sampled from the replay buffer.

We adopt the double Q-learning paradigm (Fujimoto et al., 2018; Haarnoja et al., 2018b; Saxena et al., 2023) used in previous literature for reducing estimation bias: two critics are maintained, and the minimum Q-value is used at each state-action pair. Although the corresponding theory (Fujimoto et al., 2018) for the average-reward case has not been studied in detail, we found this to improve experimental performance Understanding the effect of estimation bias is an interesting line of study for future work.

Unique to the average-reward objective is the *family* of solutions to the Bellman equation. Rather than a unique solution, the average-reward Bellman equation gives the differential value function an additional degree of freedom: If $Q(\mathbf{s}, \mathbf{a})$ satisfies Eq. (5) then $Q(\mathbf{s}, \mathbf{a}) + c$ is also a solution for all $c \in \mathbb{R}$. Section 4.1 of (Ma et al., 2021) provides an interesting discussion on the learning of value functions with an additive bias and a related downstream "value drifting problem", which they correct with value-based regularization. Section 6 of (Wan et al., 2021) provides a discussion on learning centered value functions via an additionally learned corrective "value function" F. To correct for this additional degree of freedom in an off-policy way, we introduce a baseline for centering the value function. Since an entire family of value functions can solve the Bellman equation, to pin the value, we choose the solution which passes through the origin, by always subtracting the value $Q(s = 0, a = 0)$. This choice is arbitrary, but works well in practice. Compared to the proposed regularization, it does not require any additional hyperparameters. Since it is not centering the value function in the traditional sense, it does not require on-policy data, but in principle the constant shift can be recovered upon convergence via rollouts of the optimal policy.

Finally, in average-reward tasks with terminating states, previous work (Zhang & Ross, 2021) has introduced a "reset cost", giving a penalty to the agent for resetting the environment and treating the reset state $s \sim \mu(\cdot)$ as the next state to emulate a continuing task. Prior work has chosen a fixed reset cost (-100) which was found to work for the environments tested. However, it is not reasonable to expect such penalties to be effective for tasks with different reward scales or dynamics (cf. Humanoid results in Appendix D of (Zhang & Ross, 2021)). As such, we introduce a novel adaptive reset cost: To ensure the penalty for resetting is commensurate with the accrued rewards, we simply take the mean of all rewards in the current batch that do not correspond to termination. We use a rolling average (with the same learning rate as used for θ) to slowly adapt the penalty to the agent's policy. We note that learning (and even defining) an "optimal" reset cost is an open question, which calls for further study.

5 Experiments

To evaluate our new algorithm, we test ASAC on a set of locomotion environments of increasing complexity including HalfCheetah, Ant, Swimmer, Hopper, Walker2d, and Humanoid (all version 5) from the Gymnasium Mujoco suite (Todorov et al., 2012; Towers et al., 2024). We compare the performance (average evaluation return across 10 episodes) against the existing average-reward algorithms discussed in Section 3: APO, ATRPO, and ARO-DDPG. While the focus of this paper is on a comparison of algorithms for the average-reward criterion, we also provide a comparison to the discounted algorithm SAC in the Appendix. To alleviate the cost of hyperparameter tuning, we simply use the default values inherited from SAC. Further details on the implementation and hyperparameter selection can be found in Appendix 9. ASAC performs well compared to both off-policy (ARO-DDPG) and on-policy algorithms (ATRPO, APO). To maximize performance of the ARO-DDPG baseline, we found it beneficial to use a replay buffer of maximum length (equal to number of environment interactions). Compared to ASAC, the baselines fail to solve the task in a meaningful way on some environments (Walker, Ant, Humanoid), highlighting the importance of maximum-entropy approaches for high-dimensional locomotion tasks, especially in the average-reward setting. The results of these experiments are shown in Figure 2. Our experiments suggest that ASAC represents a new state-of-the-art algorithm for the average-reward setting.

6 Discussion

The motivation for developing novel algorithms for average-reward RL arises from the problems generally associated with discounting. When the RL problem is posed in the discounted framework, a discount factor $\gamma \in [0, 1)$ is a required input parameter. However, there is often no principled approach for choosing the value of γ corresponding to the specific problem being addressed. Thus, the experimenter must treat γ as a hyperparameter. This reduces the choice of γ to a trade-off between large values to capture long-term rewards and small values to capture computational efficiency which typically scales polynomially with the horizon, $H = (1-\gamma)^{-1}$ (Kakade, 2003).

It is important to note that the horizon H introduces a natural timescale to the problem, but this timescale may not be well-aligned with another timescale corresponding to the optimal policy: the mixing time of the induced Markov chain. For the discounted solution to accurately approximate the average-reward optimal policy, the discounting timescale (horizon) must be larger than the mixing time. Unfortunately, the estimation of the mixing time for the optimal dynamics can be challenging to obtain in the general case, even when the transition dynamics are known, making a principled use of discounting computationally expensive. Therefore, an arbitrary "sufficiently large" choice of γ is often made (sometimes dynamically (Wei et al., 2021; Koprulu et al., 2024)) without knowledge of the relevant problem-dependent timescale. This can be problematic from a computational standpoint as evidenced by recent work (Jiang et al., 2015; Schulman et al., 2017; Andrychowicz et al., 2020). These points are illustrated in Figure 1 which showed the performance of SAC for the Swimmer environment with different choices of γ. For the widely used choice $\gamma = 0.99$ the evaluation rewards are low relative to the optimal case, whereas the average rewards algorithms perform well (Fig. 2), highlighting the benefits of using the average-reward criterion.

In this work, we have developed a framework for combining the benefits of the average-reward approach with entropy regularization. In particular, we have focused on extensions of the discounted algorithm SAC to the average-reward domain. By leveraging the connection of the ERAR objective to the soft discounted framework, we have presented the first solution to ERAR MDPs in continuous state and action spaces by use of function approximation. Our experiments suggest that ASAC compares favorably in several respects to their discounted counterparts: stability, convergence speed, and asymptotic performance. Our algorithm leverages existing codebases allowing for a straightforward and easily extendable implementation for solving the ERAR objective.

7 Future Work

The current work suggests multiple extensions for future exploration. Beginning with the average-reward extension of SAC (Haarnoja et al., 2018b), further developments have been made (Haarnoja et al., 2018c) including automated temperature adjustment, which we foresee as a straightforward extension for future work. As a value-based technique, other ideas from the literature such as TD(n), REDQ (Chen et al., 2021), DrQ (Kostrikov et al., 2020), combating estimation bias (Hussing et al., 2024), or dueling architectures (Wang et al., 2016) may be included. From the perspective of sampling, the calculation of θ can likely benefit from more complex replay sampling, e.g. PER (Schaul et al., 2015). An important contribution for future work is studying the sample complexity and convergence properties of the proposed algorithm. We believe that the average-reward objective with entropy regularization is a fruitful direction for further research and real-world application, with this work addressing a gap in the existing literature.

References

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvári, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pp. 3692–3702. PMLR, 2019.

J. Abounadi, D. Bertsekas, and V. S. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001. DOI: 10.1137/S0363012999361974.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. What matters in on-policy reinforcement learning? a large-scale empirical study. *arXiv preprint arXiv:2006.05990*, 2020.

Argenis Arriojas, Jacob Adamczyk, Stas Tiomkin, and Rahul V. Kulkarni. Entropy regularized reinforcement learning using large deviation theory. *Phys. Rev. Res.*, 5:023085, May 2023.

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.

David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, pp. 719–726, 1962.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith W. Ross. Randomized ensembled double q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2021.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Benjamin Eysenbach and Sergey Levine. Maximum entropy RL (provably) solves some robust RL problems. In *International Conference on Learning Representations*, 2022.

Maël Franceschetti, Coline Lacoux, Ryan Ohouens, Antonin Raffin, and Olivier Sigaud. Making reinforcement learning work on swimmer. *arXiv preprint arXiv:2208.07587*, 2022.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 06–11 Aug 2017.

Tuomas Haarnoja, Vitchyr Pong, Aurick Zhou, Murtaza Dalal, Pieter Abbeel, and Sergey Levine. Composable deep reinforcement learning for robotic manipulation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6244–6251. IEEE, 2018a.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018b.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018c.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Marcel Hussing, Claas Voelcker, Igor Gilitschenski, Amir-massoud Farahmand, and Eric Eaton. Dissecting deep rl with high update ratios: Combatting value overestimation and divergence. *arXiv preprint arXiv:2403.05996*, 2024.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1181–1189, 2015.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University College London, 2003.

Cevahir Koprulu, Po-han Li, Tianyu Qiu, Ruihan Zhao, Tyler Westenbroek, David Fridovich-Keil, Sandeep Chinchali, and Ufuk Topcu. Dense dynamics-aware reward synthesis: Integrating prior experience with demonstrations. *arXiv preprint arXiv:2412.01114*, 2024.

Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Tianjiao Li, Feiyang Wu, and Guanghui Lan. Stochastic first-order methods for average-reward markov decision processes. *arXiv preprint arXiv:2205.05800*, 2022.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In Yoshua Bengio and Yann LeCun (eds.), *International Conference on Learning Representations*, 2016.

Xiaoteng Ma, Xiaohang Tang, Li Xia, Jun Yang, and Qianchuan Zhao. Average-reward reinforcement learning with trust region methods. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2797–2803. International Joint Conferences on Artificial Intelligence Organization, 8 2021. DOI: 10.24963/ijcai.2021/385. Main Track.

Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22:159–195, 1996.

Sanjoy K Mitter and NJ Newton. The duality between estimation and control. *Published in Festschrift for A. Bennoussan*, 2000.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Abhishek Naik. Reinforcement learning for continuing problems using average reward. 2024.

Abhishek Naik, Roshan Shariff, Niko Yasui, Hengshuai Yao, and Richard S Sutton. Discounted reinforcement learning is not an optimization problem. *arXiv preprint arXiv:1910.02140*, 2019.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware unsupervised skill discovery. *arXiv preprint arXiv:2302.05103*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Konrad Cyrus Rawlik. *On probabilistic inference approaches to stochastic optimal control*. PhD thesis, The University of Edinburgh, 2013.

Dominic C Rose, Jamie F Mair, and Juan P Garrahan. A reinforcement learning approach to rare trajectory sampling. *New Journal of Physics*, 23(1):013013, 2021.

Naman Saxena, Subhojyoti Khastagir, NY Shishir, and Shalabh Bhatnagar. Off-policy average reward actor-critic with deterministic policy search. In *International Conference on Machine Learning*, pp. 30130–30203. PMLR, 2023.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Anton Schwartz. A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the tenth international conference on machine learning*, volume 298, pp. 298–305, 1993.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Chen Tessler and Shie Mannor. Reward tweaking: Maximizing the total reward while planning for short horizons. *arXiv preprint arXiv:2002.03327*, 2020.

Evangelos A Theodorou and Emanuel Todorov. Relative entropy and free energy dualities: Connections to path integral and kl control. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 1466–1473. IEEE, 2012.

Emanuel Todorov. Linearly-solvable Markov decision problems. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483, 2009.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments, 2024.

Yi Wan. *Learning and Planning with the Average-Reward Formulation*. PhD thesis, University of Alberta, 2023.

Yi Wan, Abhishek Naik, and Richard S Sutton. Learning and planning in average-reward Markov decision processes. In *International Conference on Machine Learning*, pp. 10653–10662. PMLR, 2021.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pp. 1995–2003. PMLR, 2016.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward MDPs with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021.

Feiyang Wu, Jingyang Ke, and Anqi Wu. Inverse reinforcement learning with the average reward criterion. *Advances in Neural Information Processing Systems*, 36, 2024.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Xue Yan, Yan Song, Xidong Feng, Mengyue Yang, Haifeng Zhang, Haitham Bou Ammar, and Jun Wang. Efficient reinforcement learning with large language model priors. *arXiv preprint arXiv:2410.07927*, 2024.

Sheng Zhang, Zhe Zhang, and Siva Theja Maguluri. Finite sample analysis of average-reward TD learning and Q-learning. *Advances in Neural Information Processing Systems*, 34:1230–1242, 2021.

Yiming Zhang and Keith W Ross. On-policy deep reinforcement learning for the average-reward criterion. In *International Conference on Machine Learning*, pp. 12535–12545. PMLR, 2021.

Zhe Zhang and Xiaoyang Tan. An implicit trust region approach to behavior regularized offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16944–16952, 2024.

Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

Supplementary Materials

The following content was not necessarily subject to peer review.

8 Proofs

Lemma 1 (ERAR Backup Equation)**.** *Let an ERAR MDP be given with reward function $r(\mathbf{s}, \mathbf{a})$, fixed evaluation policy π and prior policy π_0. Then the differential value of π, $Q^\pi(\mathbf{s}_t, \mathbf{a}_t)$, satisfies*

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) - \theta^\pi + \mathbb{E}_{\mathbf{s}_{t+1}\sim p} V^\pi(\mathbf{s}_{t+1}), \tag{13}$$

with the entropy-regularized definition[4] *of state-value function*

$$V^\pi(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t\sim\pi}\left[Q^\pi(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{\beta}\log\frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}\right]. \tag{14}$$

Proof. We begin with the definitions for the current state-action and for the next state-action value functions, respectively:

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) - \theta^\pi + \mathop{\mathbb{E}}_{p,\pi}\left[\sum_{k=1}^{\infty}\left(r(\mathbf{s}_{t+k}, \mathbf{a}_{t+k}) - \frac{1}{\beta}\log\frac{\pi(\mathbf{a}_{t+k}|\mathbf{s}_{t+k})}{\pi_0(\mathbf{a}_{t+k}|\mathbf{s}_{t+k})} - \theta^\pi\right)\right],$$
$$Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) = r(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) - \theta^\pi + \mathop{\mathbb{E}}_{p,\pi}\left[\sum_{k=2}^{\infty}\left((\mathbf{s}_{t+k}, \mathbf{a}_{t+k}) - \frac{1}{\beta}\log\frac{\pi(\mathbf{a}_{t+k}|\mathbf{s}_{t+k})}{\pi_0(\mathbf{a}_{t+k}|\mathbf{s}_{t+k})} - \theta^\pi\right)\right].$$

Re-writing $Q^\pi(\mathbf{s}_t, \mathbf{a}_t)$ by writing out the first term in the infinite sum and highlighting the terms of $Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})$ in blue,

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) - \theta^\pi + \mathop{\mathbb{E}}_{p,\pi}\Bigg[r(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) - \frac{1}{\beta}\log\frac{\pi(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})}{\pi_0(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})} - \theta^\pi + \sum_{k=2}^{\infty}\left(r(\mathbf{s}_{t+k}, \mathbf{a}_{t+k}) - \frac{1}{\beta}\log\frac{\pi(\mathbf{a}_{t+k}|\mathbf{s}_{t+k})}{\pi_0(\mathbf{a}_{t+k}|\mathbf{s}_{t+k})} - \theta^\pi\right)\Bigg],$$
$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) - \theta^\pi + \mathop{\mathbb{E}}_{\mathbf{s}_{t+1}\sim p, \mathbf{a}_{t+1}\sim\pi}\left[Q^\pi_\theta(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) - \frac{1}{\beta}\log\frac{\pi(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})}{\pi_0(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})}\right].$$

Identifying the entropy-regularized state value function (as in the discounted setting) $V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t\sim\pi}\left[Q^\pi(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{\beta}\log\frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}\right]$ completes the proof. ☐

Lemma 1 (ERAR Rate Gap)**.** *Consider two policies π, π' absolutely continuous w.r.t. π_0. Then the gap between their corresponding entropy-regularized reward-rates is:*

$$\theta^{\pi'} - \theta^\pi = \mathop{\mathbb{E}}_{\substack{\mathbf{s}_t\sim d_{\pi'}\\ \mathbf{a}_t\sim\pi'}}\left(A^\pi(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{\beta}\log\frac{\pi'(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}\right), \tag{15}$$

where $A^\pi(\mathbf{s}_t, \mathbf{a}_t) = Q^\pi(\mathbf{s}_t, \mathbf{a}_t) - V^\pi(\mathbf{s}_t)$ is the advantage function of policy π and $d_{\pi'}$ is the steady-state distribution induced by π'.

[4] Equation (14) is an extension of V^π_{soft} in Haarnoja et al. (2017) to the case of a non-uniform prior policy.

Proof. Working from the right-hand side of the equation,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{s}_t\sim d_{\pi'},\mathbf{a}_t\sim\pi'}\left(A^\pi(\mathbf{s}_t,\mathbf{a}_t)-\frac{1}{\beta}\log\frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}\right)=\mathop{\mathbb{E}}_{\mathbf{s}_t\sim d_{\pi'},\mathbf{a}_t\sim\pi'}\left(Q^\pi(\mathbf{s}_t,\mathbf{a}_t)-V^\pi(\mathbf{s}_t)-\frac{1}{\beta}\log\frac{\pi'(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}\right)\\
&=\mathop{\mathbb{E}}_{\mathbf{s}_t\sim d_{\pi'},\mathbf{a}_t\sim\pi'}\left(r(\mathbf{s}_t,\mathbf{a}_t)-\theta^\pi+\mathop{\mathbb{E}}_{\mathbf{s}_{t+1}\sim p}V^\pi(\mathbf{s}_{t+1})-V^\pi(\mathbf{s}_t)-\frac{1}{\beta}\log\frac{\pi'(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}\right)\\
&=\theta^{\pi'}-\theta^\pi+\mathop{\mathbb{E}}_{\mathbf{s}_t\sim d_{\pi'},\mathbf{a}_t\sim\pi'}\left(\mathop{\mathbb{E}}_{\mathbf{s}_{t+1}\sim p(\cdot|\mathbf{s}_t,\mathbf{a}_t)}V^\pi(\mathbf{s}_{t+1})-V^\pi(\mathbf{s}_t)\right)\\
&=\theta^{\pi'}-\theta^\pi.
\end{aligned}
$$

where we have used the definition

$$\theta^{\pi'}=\mathop{\mathbb{E}}_{\mathbf{s}_t\sim d_{\pi'},\mathbf{a}_t\sim\pi'}\left(r(\mathbf{s}_t,\mathbf{a}_t)-\frac{1}{\beta}\log\frac{\pi'(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}\right), \tag{16}$$

and

$$\mathop{\mathbb{E}}_{\mathbf{s}_t\sim d_{\pi'}}\mathop{\mathbb{E}}_{\mathbf{a}_t\sim\pi'}\mathop{\mathbb{E}}_{\mathbf{s}_{t+1}\sim p}V^\pi(\mathbf{s}_{t+1})=\mathop{\mathbb{E}}_{\mathbf{s}_t\sim d_{\pi'}}V^\pi(\mathbf{s}_t), \tag{17}$$

which follows given that $d_{\pi'}$ is the stationary distribution. In other words, $d_{\pi'}$ is an eigenvector of the transition operator $p(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t)\cdot\pi'(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})$. □

Theorem 1 (ERAR Policy Improvement)**.** *Let a policy π absolutely continuous w.r.t. π_0 and its corresponding differential value $Q^\pi(\mathbf{s}_t,\mathbf{a}_t)$ be given. Then, the policy*

$$\pi'(\mathbf{a}_t|\mathbf{s}_t)\doteq\frac{\pi_0(\mathbf{a}_t|\mathbf{s}_t)e^{\beta Q^\pi(\mathbf{s}_t,\mathbf{a}_t)}}{\int e^{\beta Q^\pi(\mathbf{s}_t,\mathbf{a}_t)}d\pi_0(\mathbf{a}_t|\mathbf{s}_t)} \tag{18}$$

achieves a greater entropy-regularized reward-rate. That is, $\theta^{\pi'}\geq\theta^\pi$, with equality only at convergence, when $\pi'=\pi=\pi^*$.

Proof. Let π' be defined as above. Then

$$\frac{1}{\beta}\log\frac{\pi'(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}=Q^\pi(\mathbf{s}_t,\mathbf{a}_t)-\frac{1}{\beta}\log\mathop{\mathbb{E}}_{a\sim\pi_0}e^{\beta Q^\pi(\mathbf{s}_t,\mathbf{a}_t)}. \tag{19}$$

Using Lemma 1,

$$
\begin{aligned}
\theta^{\pi'}-\theta^\pi&=\mathop{\mathbb{E}}_{s\sim d_{\pi'},a\sim\pi'}\left(A^\pi(\mathbf{s}_t,\mathbf{a}_t)-\frac{1}{\beta}\log\frac{\pi'(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}\right)\\
&=\mathop{\mathbb{E}}_{s\sim d_{\pi'},a\sim\pi'}\left(Q^\pi(\mathbf{s}_t,\mathbf{a}_t)-V^\pi(s)-\frac{1}{\beta}\log\frac{\pi'(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}\right)\\
&=\mathop{\mathbb{E}}_{s\sim d_{\pi'},a\sim\pi'}\left(\frac{1}{\beta}\log\mathop{\mathbb{E}}_{a\sim\pi_0}e^{\beta Q^\pi(\mathbf{s}_t,\mathbf{a}_t)}-V^\pi(s)\right)\geq 0\,,
\end{aligned}
$$

where the last line follows from the variational formula Mitter & Newton (2000); Theodorou & Todorov (2012),

$$\frac{1}{\beta}\log\mathop{\mathbb{E}}_{a\sim\pi_0}e^{\beta Q^\pi(\mathbf{s}_t,\mathbf{a}_t)}=\sup_\pi\mathop{\mathbb{E}}_{a\sim\pi}\left(Q^\pi(\mathbf{s}_t,\mathbf{a}_t)-\frac{1}{\beta}\log\frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}\right). \tag{20}$$

□

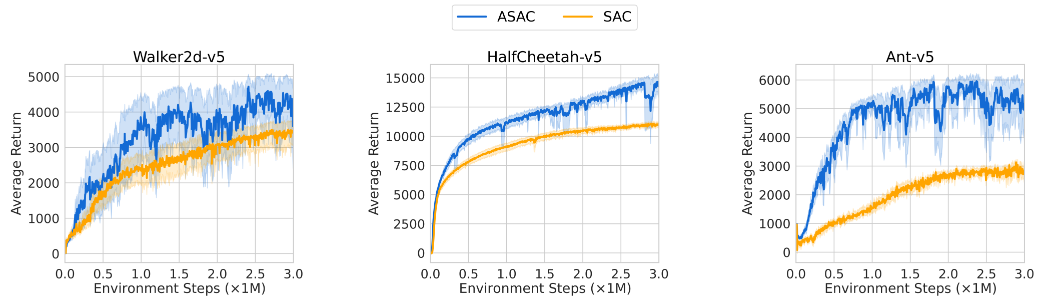

Figure 3: Comparison to SAC shows that our average-reward extension outperforms the original discounted SAC on the environments tested. We note that the reward values are different than in earlier environment versions (as used in e.g. Haarnoja et al. (2018b)), as the result of an updated reward function and bug fixes (including changes to contact forces, control costs), described in detail here: `https://farama.org/Gymnasium-MuJoCo-v5_Environments`.

9 Implementation Details

For all SAC runs, we used Raffin et al. (2021) implementation of SAC with hyperparameters (beyond the default values) shown below in Section 9.1. The finetuned runs here took ~ 3000 GPU hours for all environments, ran on a variety of RTX series and A100 GPUs. Each run requires roughly $\sim 1 - 10$ GB of RAM.

9.1 Hyperparameters

In addition to the methods discussed in the main text, we also use gradient clipping (on critic network only), with the maximum gradient norm of 10 for all experiments.

For all ASAC experiments, we use the same hyperparameters as Haarnoja et al. (2018b): batch size of 256, replay buffer size of $1\,000\,000$, hidden dimension of 256 for each of 2 hidden layers (actor and critic networks), Polyak averaging with coefficient 0.005, train frequency and gradient steps of 1 (train for one gradient step at each environment step). We use the Adam optimizer for actor, critic, and reward-rate with learning rates $10^{-4}, 5 \times 10^{-4}, 5 \times 10^{-3}$. We clip the critic network gradients with a maximum norm of 10. In all environments (for SAC and ASAC) we use $\beta = 5$, except for Swimmer and Humanoid, for which we use $\beta = 20$. Note that this is in line with the "reward scale" used in (Haarnoja et al., 2018b). We found that hyperparameter sweeps can give better performance for individual environments, but these choices gave a strong performance universally. We found the replay buffer size to be a sensitive hyperparameter for ARO-DDPG, in particular for maintaining its asymptotic performance. We chose the largest replay buffer for ARO-DDPG (equivalent to total environment interactions), but further tuning is left to future work as it is an expensive environment-dependent operation. We also note that beyond the default hyperparameters for ASAC described above, we did not perform any tuning, showcasing ASAC's robustness to hyperparameter choice. Future work may entail an extensive hyperparameter sweep and sensitivity analysis to further understand the robustness and maximize performance across various environments.