A Value-Based Approach to Maximum Entropy Exploration

Jacob Adamczyk, Adam Kamoski, Rahul Kulkarni

Keywords: exploration, average-reward, MaxEnt, entropy-regularization

Summary

Efficient exploration remains a central challenge in reinforcement learning, serving as a useful pretraining objective for data collection, particularly when an external reward function is unavailable. A principled formulation of the exploration problem is to find policies that maximize the entropy of their induced steady-state visitation distribution, thereby encouraging uniform long-run coverage of the state space. Many existing exploration approaches require estimating state visitation frequencies through repeated on-policy rollouts, which can be computationally expensive. In this work, we instead consider an intrinsic average-reward formulation in which the reward is derived from the visitation distribution itself, so that the optimal policy maximizes steady-state entropy. An entropy-regularized version of this objective admits a spectral characterization: the relevant stationary distributions can be computed from the dominant eigenvectors of a problem-dependent transition matrix. This insight leads to a novel algorithm for solving the maximum entropy exploration problem, EVE (EigenVector-based Exploration), which avoids explicit visitation estimation and instead computes the solution through iterative updates, similar to a value-based approach. To address the original unregularized objective, we employ a posterior-policy iteration (PPI) approach, which monotonically improves the entropy. We prove that EVE's core update equation converges linearly under standard assumptions. We empirically demonstrate that EVE produces policies with competitive exploration performance relative to rollout-based baselines in deterministic grid-world environments.

Contribution(s)

1. This paper presents an algorithm (EVE) for maximum entropy exploration in the average-reward setting that computes the optimal exploratory policy using a novel update equation instead of rollouts.
   Context: Prior approaches to maximum entropy exploration typically require estimating visitation frequencies through repeated policy rollouts and reward iterations (Hazan et al., 2019). The proposed update equation requires a known forward and backward transition model to evaluate probability flows.
2. We prove that the EVE update rule is a contraction mapping, guaranteeing linear convergence to a unique (up to scaling) fixed point. Using posterior policy iteration, the EVE algorithm provably achieves monotonic entropy improvement.
   Context: We use existing techniques from non-linear Perron-Frobenius theory (Lemmens & Nussbaum, 2014) to prove convergence in projective space. Recent work by Hazan et al. (2019) gives convergence proofs for alternative algorithms in the same maximum entropy exploration setting.

A Value-Based Approach to Maximum Entropy Exploration

Jacob Adamczyk[1,2], **Adam Kamoski**[1], **Rahul Kulkarni**[1,2]
{jacob.adamczyk001, adam.kamoski001, rahul.kulkarni}@umb.edu

[1]**Department of Physics, University of Massachusetts Boston**
[2]**NSF Institute for Artificial Intelligence and Fundamental Interactions**

Abstract

Efficient exploration remains a central challenge in reinforcement learning, serving as a useful pretraining objective for data collection, particularly when an external reward function is unavailable. A principled formulation of the exploration problem is to find policies that maximize the entropy of their induced steady-state visitation distribution, thereby encouraging uniform long-run coverage of the state space. Many existing exploration approaches require estimating state visitation frequencies through repeated on-policy rollouts, which can be computationally expensive. In this work, we instead consider an intrinsic average-reward formulation in which the reward is derived from the visitation distribution itself, so that the optimal policy maximizes steady-state entropy. An entropy-regularized version of this objective admits a spectral characterization: the relevant stationary distributions can be computed from the dominant eigenvectors of a problem-dependent transition matrix. This insight leads to a novel algorithm for solving the maximum entropy exploration problem, EVE (EigenVector-based Exploration), which avoids explicit visitation estimation and instead computes the solution through iterative updates, similar to a value-based approach. To address the original unregularized objective, we employ a posterior-policy iteration (PPI) approach, which monotonically improves the entropy. We prove that EVE's core update equation converges linearly under standard assumptions. We empirically demonstrate that EVE produces policies with competitive exploration performance relative to rollout-based baselines in deterministic grid-world environments.

1 Introduction

Efficient exploration in reinforcement learning (RL) remains a long-standing open problem. A rich subfield has emerged to tackle this problem, with new research questions fueling new ideas such as entropy- and information-seeking agents (Tiomkin & Tishby, 2017), count-based systems (Bellemare et al., 2016; Ostrovski et al., 2017; Zhou et al., 2020; Lobel et al., 2023), curiosity-driven exploration (Pathak et al., 2017) and random network distillation (Burda et al., 2018). However, novel approaches and insights are still needed to address the fundamental question of exploration: *How can an agent learn to obtain uniform coverage in a new environment?*

An important class of approaches for exploration in RL focuses on objectives derived from visitation frequencies. Although efficient algorithms have been formulated and developed using these objectives (Hazan et al., 2019), such approaches are inherently computationally demanding. By design, the entropy objective in these approaches is dependent on the stationary visitation distribution induced by the policy itself. Evaluating this distribution typically requires estimating visitation frequencies through repeated rollouts. Because the objective is defined in terms of the policy-induced distribution, improving the policy requires repeatedly estimating this changing distribution, creat-

ing a circular dependency between policy updates and visitation estimates. In practice, this often necessitates on-policy sampling and can make optimization computationally expensive.

In this work, we address the exploration problem from a new perspective, deriving an update equation constructed to be consistent with the desired visitation distribution, using a backward transition model instead of rollouts. We arrive at an algorithm which uses a two-step temporal difference learning technique to learn the optimal policy directly.

Recent research has derived results that provide a framework for addressing the challenges noted. For systems with deterministic dynamics, Bayesian inference approaches can be used to solve the optimal control problem in entropy-regularized RL (Levine, 2018). In the average-reward setting, analytical expressions for the optimal policy and the associated state-action visitation distributions have been derived (Arriojas et al., 2023a). Our work leverages this distinctive spectral viewpoint, in which optimal policies and visitation distributions are characterized through eigenvectors of a tilted transition operator.

Building on the analytical results from Arriojas et al. (2023a), we obtain expressions for key quantities in exploration, including distributions in state-action space, an intrinsic reward function, and their relationship to the environment's transition dynamics. These analytic results lead to a new update equation that can be efficiently computed, based on a balancing of forward and backward probability flows. The resulting algorithm leads to policies that induce state-action visitation distributions with high entropy.

We consider the "reward-free" setting: there are no extrinsic rewards and the agent's objective is to explore the environment uniformly. Our proposed algorithm, EVE ("EigenVector-based Exploration") uses the dominant eigenvectors of a tilted transition matrix to determine the unique policy that solves the entropy-regularized exploration problem. When an extrinsic reward function is provided, downstream tasks can then leverage the data collected by EVE to solve complex tasks with potentially sparse rewards. Alternatively, one might use EVE in the online setting to define the behavior policy. EVE shows that maximum-entropy exploration can be approached as a spectral fixed-point problem, allowing exploratory policies to be computed directly from the transition dynamics without estimating visitation frequencies through rollouts.

In the following sections, we provide necessary background material in reinforcement learning, highlighting the entropy-regularized average-reward objective before presenting our framework for calculating optimal policies. We then describe the corresponding algorithm, its convergence properties, and provide validating experiments. We end the paper by connecting to related work and discussing our results.

2 Background

Consider an environment with state space $\mathcal{S}$ and action space $\mathcal{A}$, and an agent interacting with the environment through a (generally stochastic) policy $\pi(a|s)$. The environment's transition dynamics $p(s'|s,a)$ describe the next state s' based on the current state s and action taken, a. In the following, we focus on the discrete state-action case for ease of exposition. The typical RL objective is to optimize the accumulated reward, $r(s,a)$. That is, to solve the following optimization problem:

$$J^* = \max_{\pi} \mathbb{E}_{\tau\sim p,\pi,\mu} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t), \tag{1}$$

where trajectories are sampled starting from an initial distribution μ and following dynamics p and policy π thereafter. Rewards are exponentially discounted over time with a discount factor $\gamma \in [0, 1)$. Much prior work in reinforcement learning considers the discounted objective in Equation (1); however, we argue that this choice is not suitable for the exploration problem. A discount factor inherently introduces a finite timescale for the agent's decision-making capacity: $T \sim (1-\gamma)^{-1}$. To solve the exploration problem, an agent should visit many states, even those beyond this artificially

chosen temporal horizon. To this end, we adopt an average-reward approach to the exploration problem. The average-reward RL objective has received growing interest in recent years (Mahadevan, 1996; Zhang et al., 2021b; Wan et al., 2021; Zhang & Ross, 2021; Ma et al., 2021; Hisaki & Ono, 2024; Adamczyk et al., 2025a; Rojas & Lee, 2026) due to its ability to solve long-horizon, continuing tasks without a discount factor. In the maximum entropy exploration problem, we argue that discounting may not always be appropriate: Using a discounted objective will result in entropy maximization of the *discounted* occupancy measure, whereas we may instead be interested in the undiscounted steady-state distribution, which cannot be easily computed with discounted approaches (Naik et al., 2019). In the present case, we therefore replace the discounted objective in Equation (1) by the *average reward* objective, defined by:

$$\rho^* = \max_\pi \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\tau\sim p,\pi,\mu} \sum_{t=1}^{T} r(s_t, a_t) = \max_\pi \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_{p,\pi}(s,a) r(s,a), \tag{2}$$

where ρ^* is the optimal "reward-rate". Assuming it exists (cf. the standard well-mixing assumptions, as in (Wan et al., 2021)), the reward-rate can be expressed as an expectation of rewards under the optimal policy's steady-state distribution (right-hand side of Equation (2)).

In this work, we are interested in solving the maximum (state-action) entropy problem. In this setting, "optimal exploration" corresponds to finding a policy that achieves a maximally uniform distribution over state-action pairs (Krass & Vrieze, 2002). Analogous to the formulation in (Hazan et al., 2019), we study the following objective:

$$\max_\pi \mathbb{H}(d_{p,\pi}) = \max_\pi \left(-\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_{p,\pi}(s,a) \log d_{p,\pi}(s,a) \right), \tag{3}$$

where $d_{p,\pi}$ represents the stationary distribution under the transition dynamics p and control policy π, obtained in the long-time limit. In other words, $d_{p,\pi}$ is the non-trivial eigenvector of the state-action transition matrix $p(s'|s,a)\pi(a'|s')$. Throughout, we will assume the transition dynamics is deterministic, aperiodic, and irreducible.

One approach to address this optimization problem is via reinforcement learning: By comparing Equation (2) to Equation (3), we see that one should assign the reward $r(s,a) = -\log d_{p,\pi}(s,a)$, to be consistent with the maximum entropy objective. With this identification, optimization of the average-reward objective is a way to solve the problem of maximum-entropy exploration using RL.

To leverage closed-form results in the average-reward setting, we include an entropy-regularization term: $\beta^{-1} D_{\mathrm{KL}}(\pi \,\|\, \pi_0)$, where π_0 is a chosen reference or prior policy (often taken to be uniform, in the case of "MaxEnt RL" (Haarnoja et al., 2018a)), and β is a real-valued "inverse temperature" (α^{-1} in e.g. (Haarnoja et al., 2018b)) responsible for regulating the importance of the Kullback-Leibler divergence, D_{KL} relative to the accrued rewards. The objective for the entropy-regularized average-reward problem can be expressed as:

$$\theta^* = \max_\pi \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\tau\sim p,\pi,\mu} \sum_{t=1}^{T} \left(r(s_t,a_t) - \beta^{-1} \log \frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)} \right). \tag{4}$$

Here, θ^* takes the role of the reward-rate ρ^*, but includes the entropic cost. A finite temperature $\beta < \infty$ can be used throughout, treating the regularized objective as a surrogate for the un-regularized solution of interest, with $\beta \to \infty$. In the following sections, we also propose two methods for solving the un-regularized problem.

Recent work by Arriojas et al. (2023a) provides a solution to the entropy-regularized average reward problem through the construction of a "tilted matrix". We summarize the main outcomes of their framework below. The key to solving the entropy-regularized average reward problem is the tilted matrix, defined as:

$$\widetilde{P}(s',a'|s,a) = p(s'|s,a)\pi_0(a'|s')e^{\beta r(s,a)}, \tag{5}$$

which combines the transition dynamics, prior policy, and reward function. The tilted matrix is sub-stochastic (column sums are less than one) and satisfies the Perron-Frobenius theorem. The corresponding dominant eigenvalue of $\widetilde{P}$ is $\lambda = \exp \beta\theta^*$, i.e. the exponential of the entropy-regularized average reward-rate, θ^*. These results can be understood as an extension of the linearly solvable framework of Todorov (2006). Furthermore, it was shown that the corresponding left eigenvector, satisfying $u^T \widetilde{P} = \lambda u^T$, encodes the optimal policy: $\pi^*(a|s) \propto \pi_0(a|s)u(s,a)$. The distinct, non-trivial right eigenvector v satisfying $\widetilde{P}v = \lambda v$ represents a "quasi-stationary distribution" (Méléard & Villemonais, 2012). These left and right eigenvectors are related to the backward and forward messages in the Bayesian interpretation, respectively, as discussed by Levine (2018); Arriojas (2022). Importantly, the steady-state distribution induced by the optimal policy π^* and transition dynamics p can be expressed through their (Hadamard) product:

$$d_{p,\pi^*}(s,a) = u(s,a)v(s,a). \tag{6}$$

This decomposition is similar to that given by Jain et al. (2023). However, our approach is Markovian, in the average-reward framework, and as shown below, yields a new update equation based on temporal differences. In the following, we discuss how the ideas from this section can be combined to unlock new insights and algorithms for the maximum-entropy exploration problem.

3 Results

We begin with the observation that, rather than using rollouts to estimate the policy-induced entropy, the steady-state distribution can be derived from the eigenvectors of the tilted matrix, attainable from off-policy methods (Adamczyk et al., 2025b). Given the linear-algebraic framework discussed above, we aim to find an intrinsic reward function which optimizes the average entropy-rate of the agent. Rather than using discounted occupancy measures, this framework allows the average entropy to be optimized directly. Comparing with Equation (4) we see that by choosing the reward function to be of the form

$$r(s,a) = -\log u(s,a)v(s,a), \tag{7}$$

the optimal reward-rate is exactly the desired entropy-rate. The Perron-Frobenius theorem implies that both vectors u and v are unique and strictly positive. The right eigenvector v can be computed from Equation (9) at any point to estimate the entropy of the distribution without any rollouts of the policy. This approach is therefore off-policy: for a fixed prior policy, data is not needed from the learned policy. This is in stark contrast to rollout-based methods that require sampling from the online policy. However, there are two caveats: (1) this is still an implicit solution, where the left and right eigenvectors u, v of the new tilted matrix must be self-consistent with the vectors used to define the reward function; and (2) an exact correspondence to the maximum entropy distribution only holds for the un-regularized problem, where $\beta \to \infty$. We address these issues in the following sections.

3.1 Self-Consistent Solution

To address the first problem (the self-referential nature of eigenvectors and reward functions), we must find left and right eigenvectors u, v which consistently satisfy the corresponding eigenvector equations and the reward function definition introduced in Section 2. After inserting the definition for the reward function (Equation (7)) into the eigenvector equations, we have for $\beta = 1$, (see Appendix A for further details):

$$\sum_{s',a'} u(s',a')P^{\pi_0}(s',a'|s,a) = \lambda u^2(s,a)v(s,a)\,, \tag{8}$$

$$\sum_{s,a} P^{\pi_0}(s',a'|s,a)\frac{1}{u(s,a)} = \lambda v(s',a')\,, \tag{9}$$

where $P^{\pi_0}(s', a'|s, a) = p(s'|s, a)\pi_0(a'|s')$ is the joint state-action transition dynamics and $0 < \lambda < 1$, the dominant (Perron-Frobenius) eigenvalue, encodes the optimal entropy-regularized reward rate, θ^*, as discussed above. Note that this optimal reward-rate θ^* includes a contribution from the average entropy achieved by the policy's steady-state distribution (our original objective) *and* the entropy of the optimal policy itself. In the next subsection, we discuss how to mitigate the latter term's effect.

We now note that Equation (9) allows one to eliminate the right eigenvector v. From an algorithmic standpoint, this reduces the required memory and simplifies the problem to estimating a single function. The relationship derived from the above eigenvector equations now provides our novel fixed point iteration scheme, which for general $\beta \neq 1$, reads:

$$u(s,a) \leftarrow \mathcal{T}u(s,a) \doteq \left(\frac{\left(\sum_{s',a'} u(s',a') P^{\pi_0}(s',a'|s,a)\right)^{\frac{1}{\beta}}}{\sum_{\bar{s},\bar{a}} P^{\pi_0}(s,a|\bar{s},\bar{a}) u(\bar{s},\bar{a})^{-\frac{1}{\beta}} \left(\sum_{s,a} u(s,a) P^{\pi_0}(s,a|\bar{s},\bar{a})\right)^{\frac{1-\beta}{\beta}}} \right)^{\frac{\beta}{1+\beta}} \tag{10}$$

where $(\bar{s}, \bar{a}) \to (s, a) \to (s', a')$ denote successive state-action pairs. This update equation combines information from the future (numerator) and the past (denominator). To gain a better intuition, we can rewrite this equation in log-space (motivated by the connection to the value function, since $q(s, a) = \beta^{-1} \log u(s, a)$), and highlight the simpler case of $\beta = 1$, finding

$$q(s,a) = \frac{1}{2} \log \left(\mathbb{E}_{a' \sim \pi_0} e^{q(s',a')} \right) - \frac{1}{2} \log \left(\sum_{\bar{s},\bar{a}} P(s,a|\bar{s},\bar{a}) e^{-q(\bar{s},\bar{a})} \right). \tag{11}$$

From here, we see that q satisfies a certain "soft flow" equation: q balances between the flows out of a state (first term) and the flows into the state (second term). The first term represents a soft maximum over the next state-action pairs while the second term represents a soft minimum over the preceding state-action pairs.

The derived update scheme simplifies the expensive iteration (computing optimal policy, rollouts, new reward function, etc.) into a single fixed-point equation. Interestingly, even without a discount factor, successive iterates of EVE maintain their magnitude. This stability is a result of Equation (10)'s convergence and the homogeneity of the updates (cf. Appendix A). Importantly, since we do not rely on discounting, this flow equally weights incoming (previous) and outgoing (future) transitions, ensuring that temporally distant states are visited with uniform probability, and not artificially weighted by the discounted measure. Upon convergence, Equation (10) therefore allows us to solve for the policy which maximizes entropy of the *undiscounted* distribution over state-action pairs.

The following theorem guarantees that, despite lacking an obvious contraction factor, (an m-step form of) the iteration in Equation (10) is indeed a contraction mapping and thus converges to a unique function, for all $\beta \geq 1$. Below, let $\kappa(A)$ denote the projective diameter of a positive matrix, A. We restrict the operation of $\mathcal{T}$ to the cone $\mathbb{R}_{>0}^{|\mathcal{S}||\mathcal{A}|}$ (the positive orthant).

Theorem 1. *Let the dynamics induced by $\pi_0(a|s)$ and $p(s'|s, a)$ be irreducible and aperiodic with index of primitivity m. The m-step mapping $u \leftarrow \mathcal{T}_m(u)$ extension of Equation (10) is a contraction under the projective metric, and converges linearly with rate $\kappa((P^{\pi_0})^m)$ when $\beta \geq 1$.*

The proof of this statement is given in Appendix A and follows from the properties of Hilbert's projective metric. Intuitively, this can be seen as a form of non-linear Perron-Frobenius theory (Lemmens & Nussbaum, 2012). The projective metric is a logarithmic form of the more well-known *span semi-norm* common in the average-reward literature (Zhang et al., 2021c).

3.2 The Un-regularized Objective

The entropy-regularized objective maintains an additional cost for deviating from a prior policy π_0. This regularization technique has proven beneficial in some contexts, like MaxEnt RL where π_0 is uniform Haarnoja et al. (2018a), offline RL where π_0 represents a behavior policy (Wu et al., 2019; Zhang & Tan, 2024), and LLM alignment (Rafailov et al., 2024; Yan et al., 2024). In the exploration problem, it may be of reasonable interest to use some (extrinsically defined) task's optimal policy as the prior. In this case, the agent will explore novel parts of the environment while staying "close" (in the sense of Kullback-Leibler divergence) to the prior policy, which may be an effective combination to satisfy the exploration-exploitation tradeoff. Though the regularized objective may be of independent interest, our primary objective is to obtain a solution to the maximum entropy exploration problem, as posed in Equation (3), without the additional regularization term. As such, maintaining such a regularization will bias the solution away from the pure MaxEnt solution.

In order to directly use the tilted matrix to determine the solution to the un-regularized problem, we need to take the limit $\beta \to \infty$. However, instead of increasing the inverse temperature β, one can alternatively anneal in the policy space by using an adaptive prior policy π_0, as introduced by Rawlik et al. (2012); Rawlik (2013). Adopting this approach, we will iterate the prior and posterior policies, by updating π_0 with the optimal policy resulting from the left eigenvector ($\pi^* \propto \pi_0 u$). This left eigenvector, for a given π_0, is found by solving for the fixed point in Equation (10). Intuitively, this method reduces the impact of the relative entropy regularization term, not by adjusting its weight β but instead by adjusting the reference policy. Upon convergence, the prior and optimal policies are the same, thereby incurring no additional entropic costs and only maximizing the reward function of interest (Eq. (7)). This method is known as *posterior policy iteration*, or PPI (Rawlik et al., 2012), and was also studied in the deep RL setting by Adamczyk et al. (2025b). In Appendix A.4 we prove that PPI provides monotonic entropy improvement in this setting, where the reward function also changes. We provide the corresponding pseudocode for EVE below.

Algorithm 1 EVE

Require: Transition matrix $P(s'|s,a)$, initial prior policy $\pi_0(a|s)$
Require: Inverse temperature $\beta(t) \geq 1$, maximum u iterations N, maximum PPI iterations T
for $t = 1$ **to** T **do**
  for $n = 1$ **to** N **do**
    $u \leftarrow \mathcal{T}(u; \beta(t))$ `// Update according to Equation (10)`
  Update optimal policy: $\pi^*(a|s) = \frac{\pi_0(a|s)u(s,a)}{\sum_a \pi_0(a|s)u(s,a)}$
  $\pi_0(a|s) \leftarrow \pi^*(a|s)$ `// Update prior policy`
Output: Exploration policy, π^*

4 Experiments

We use a tabular GridWorld environment to verify our theoretical results and test the proposed method (given in Algorithm 1). Specifically, we use deterministic dynamics with four discrete actions (in the cardinal directions). The initial state is denoted by a green circle and absorbing states are denoted by the grey region (transitions the agent back to the initial state). Although we use a tabular approach (with transition matrix given), the agent can alternatively learn a model (Sutton, 1991; Hafner et al., 2019; Moerland et al., 2023) during training. Notably, because of the form of the update in Equation (10), one must also model the backward transitions, or the "parent" state-action pairs.

As baselines, we consider the MaxEnt algorithm from Hazan et al. (2019) and a set of rollout-based techniques where a policy's steady-state distribution is calculated, and a new reward function is defined according to $r(s,a) = -\log d_\pi(s,a)$. Experimentally, we observe oscillatory behaviors in the rollout-based methods, as outlined in (Lee et al., 2019). To mitigate this effect, we introduce

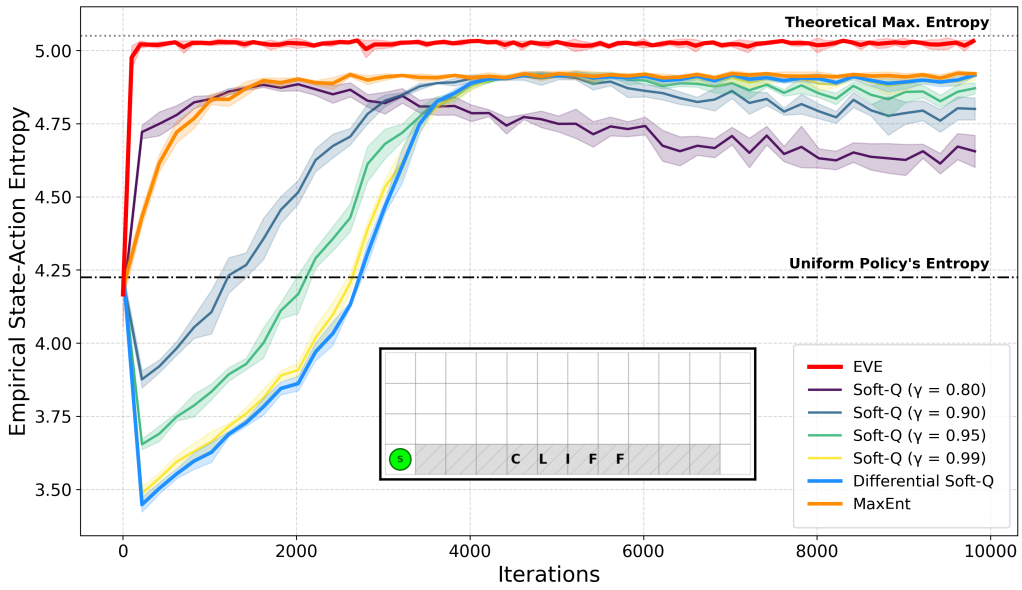

Figure 1: Compared to the baselines, EVE finds an exploratory policy that produces higher entropy after fewer iterations. (Inset) "CliffWorld" environment used. Agent can move in cardinal directions. Green circle denotes initial state; falling into the cliff resets the agent. Each line represents the mean over 5 independent initializations; shaded region denotes one standard deviation.

a learning rate (finetuned for each baseline) to slowly mix the consecutive reward functions. We also warmstart the RL loop with the previous value function. We found both of these steps were necessary for stable convergence of the baselines. We use a discounted approach, where the reward function is fed into a discounted soft Q learning solver ($\gamma \in \{0.8, 0.9, 0.95, 0.99\}$). We also use an average-reward differential soft Q learning algorithm to isolate the effects of discounting and our own update equation. The soft Q learning baselines use a fixed number of steps (50) before the reward function is updated and a linear schedule for $\beta \in \{1, 2, \ldots, 10\}$.

Importantly, EVE does not require tracking a distribution or reward function, since these are built into the update equation, by design. As such, there are no oscillatory effects, so the corresponding learning rate can be removed altogether. The MaxEnt algorithm uses a convex combination of greedy policies to form a stochastic optimal policy, whereas the EVE policy is naturally stochastic at each iteration. As such, we circumvent the potentially large memory footprint of MaxEnt, which requires storing all previous policies.

Our experiments in Figure 1 show that EVE is able to find policies capable of efficient exploration that achieve the maximum possible entropy, $\log |\mathcal{S}||\mathcal{A}|$. To induce a uniform distribution, EVE's resulting policy is highly non-uniform, as the agent must navigate away from the cliff to continue exploring. The policies learned by rollout methods lead to lower entropy (worse coverage) and take more iterations to obtain. On the horizontal axis, "iterations" refers to the number of synchronous updates to the value function (i.e. value iteration steps). The code for our experiments can be found at `https://github.com/JacobHA/EVE`.

5 Related Work

Being a fundamental pillar of reinforcement learning, the issue of exploration has cemented itself as a cornerstone of theoretical and experimental RL literature, with ever growing interest, especially as environments become more complex and rewards more sparse. A prominent line of work formulates exploration as the maximization of entropy over state or state-action visitation distri-

butions. For instance, Hazan et al. (2019) demonstrate that maximizing the entropy of discounted occupancy measures can be solved as a convex program with polynomial sample complexity, marking an important step toward provably convergent exploration. Similarly, Jain et al. (2023) explore the maximum entropy problem, but with non-Markovian policies within a discounted framework. Touati & Ollivier (2021) use an offline approach that approximates the successor measure with a low-rank forward-backward decomposition. Our method is online, off-policy, and exact. The reliance on discounting in these approaches inherently skews the visitation measure away from the true long-run steady-state. Consequently, they are unable to leverage the exact spectral connection to the stationary distribution's eigenvectors that our undiscounted, average-reward formulation natively exploits. Table 1 of Mutti et al. (2022) provides a comprehensive overview of exploration methods.

Spectral methods have also been fruitfully applied to RL exploration in other contexts. Machado et al. (2017) utilize the eigenvectors of the successor representation to discover "eigenoptions" that encourage diverse exploratory behaviors. While conceptually related, EVE differs fundamentally in both objective and mechanism: rather than learning options to implicitly aid exploration, we focus on explicitly optimizing Equation (3), deriving a closed-form update from the eigenvectors of the tilted transition matrix.

Finally, the precise formulation of the entropy objective and its optimization landscape separates our work from previous approaches. While much of the literature focuses strictly on state entropy, recent studies emphasize the importance of maximizing entropy across the joint state-action space (Zhang et al., 2021a). Whereas Zhang et al. (2021a) tackles this using Rényi entropy within a total-reward framework, EVE employs Shannon entropy within the average-reward setting. Overall, many of these experimental methods use function approximation techniques and thus could offer a viable pathway for extending EVE to continuous, model-free problems.

6 Discussion

EVE offers a principled, computationally efficient solution to the exploration problem by directly optimizing the steady-state entropy of state-action visits uses a forward and backward transition model instead of costly rollouts or discounted surrogates. By leveraging recent analytical advances in entropy-regularized average-reward RL, we derived a self-consistent update mechanism for solving the MaxEnt exploration problem. Because our formulation is naturally embedded in the entropy-regularized setting, we can leverage the flexibility of the prior policy π_0, obtaining a smooth transition from a regularized exploration policy to the un-regularized maximum-entropy solution via PPI.

Beyond the new perspective and theoretical techniques provided for the pure exploration problem, EVE can also serve as a powerful pretraining objective Liu & Abbeel (2021). An agent initialized with this maximum-entropy policy uniformly covers the state-action space, which is especially valuable in sparse-reward environments. We believe that EVE has the potential to become a versatile building block for both reward-free data collection and downstream RL fine-tuning, thanks to its efficient learning.

6.1 Limitations and Future Work

The solution method proposed is designed to solve the maximum entropy problem. As such, it is not well-suited for "noisy TV" problems (Burda et al., 2018) which may instead require a more information-theoretic perspective on the intrinsic reward function. The current solution method only applies for the case of deterministic dynamics, based on the results in (Arriojas et al., 2023a); however an extension to stochastic dynamics exists and requires an additional loop over "biasing functions" (Arriojas et al., 2023b).

We also note that objectives beyond the Shannon entropy may be of interest (Hazan et al., 2019). When these alternate objectives depend on the stationary distribution, our results may also apply. For

example, the marginal-matching problem Lee et al. (2019) and state-entropy can also be addressed within this framework. We envision extensions of these ideas in model-based RL, where a learned (backward) dynamics model is used to estimate the EVE update equation.

References

Jacob Adamczyk, Volodymyr Makarenko, Stas Tiomkin, and Rahul V Kulkarni. Average-reward soft actor-critic. *Reinforcement Learning Journal*, 6:412–430, 2025a.

Jacob Adamczyk, Volodymyr Makarenko, Stas Tiomkin, and Rahul V Kulkarni. EVAL: EigenVector-based Average-reward Learning. *The Eighth Workshop on "Generalization in Planning" at the Association for the Advancement of Artificial Intelligence*, 2025b.

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

Argenis Arriojas, Jacob Adamczyk, Stas Tiomkin, and Rahul V. Kulkarni. Bayesian inference approach for entropy regularized reinforcement learning with stochastic dynamics. In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 99–109. PMLR, 31 Jul–04 Aug 2023b.

Argenis A. Arriojas. *Analytical Framework for Entropy Regularized Reinforcement Learning Using Probabilistic Inference*. PhD thesis, University of Massachusetts Boston, 2022.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

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, pp. 1861–1870. PMLR, 10–15 Jul 2018a.

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*, 2018b.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pp. 2681–2691. PMLR, 2019.

Yukinari Hisaki and Isao Ono. RVI-SAC: Average reward off-policy deep reinforcement learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 18352–18373. PMLR, 21–27 Jul 2024.

Arnav Kumar Jain, Lucas Lehnert, Irina Rish, and Glen Berseth. Maximum state entropy exploration using predecessor and successor representations. *Advances in neural information processing systems*, 36:49991–50019, 2023.

Elon Kohlberg and John W Pratt. The contraction mapping approach to the Perron-Frobenius theory: Why Hilbert's metric? *Mathematics of Operations Research*, 7(2):198–210, 1982.

Dmitry Krass and O. J. Vrieze. Achieving target state-action frequencies in multichain average-reward Markov decision processes. *Mathematics of Operations Research*, 27(3):545–566, 2002.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press, 2012.

Bas Lemmens and Roger D Nussbaum. Birkhoff's version of Hilbert's metric and its applications in analysis. In *Handbook of Hilbert geometry*, pp. 275–303. European Mathematical Society-EMS-Publishing House GmbH, 2014.

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

Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. *Advances in Neural Information Processing Systems*, 34:18459–18473, 2021.

Sam Lobel, Akhil Bagaria, and George Konidaris. Flipping coins to estimate pseudocounts for exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 22594–22613. PMLR, 2023.

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. Main Track.

Marlos C. Machado, C. Rosenbaum, Xiaoxiao Guo, Miao Liu, G. Tesauro, and Murray Campbell. Eigenoption discovery through the deep successor representation. *International Conference on Learning Representations*, 2017.

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

Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023.

Mirco Mutti, Riccardo De Santi, and Marcello Restelli. The importance of non-markovianity in maximum state entropy exploration. In *International Conference on Machine Learning*, pp. 16223–16239. PMLR, 2022.

Sylvie Méléard and Denis Villemonais. Quasi-stationary distributions and population processes. *Probability Surveys*, 9, January 2012. ISSN 1549-5787.

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.

Georg Ostrovski, Marc G Bellemare, Aäron Oord, and Rémi Munos. Count-based exploration with neural density models. In *International conference on machine learning*, pp. 2721–2730. PMLR, 2017.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

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.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. *Proceedings of Robotics: Science and Systems VIII*, 2012.

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

Juan Sebastian Rojas and Chi-Guhn Lee. A differential perspective on distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

Stas Tiomkin and Naftali Tishby. A unified bellman equation for causal information and value in markov decision processes. *arXiv preprint arXiv:1703.01585*, 2017.

Emanuel Todorov. Linearly-solvable markov decision problems. *Advances in neural information processing systems*, 19, 2006.

Ahmed Touati and Y. Ollivier. Learning one representation to optimize all rewards. *Neural Information Processing Systems*, 2021.

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.

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.

Chuheng Zhang, Yuanying Cai, Longbo Huang, and Jian Li. Exploration by maximizing rényi entropy for reward-free rl framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10859–10867, 2021a.

Shangtong Zhang, Yi Wan, Richard S Sutton, and Shimon Whiteson. Average-reward off-policy policy evaluation with function approximation. In *international conference on machine learning*, pp. 12578–12588. PMLR, 2021b.

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, 2021c.

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.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502. PMLR, 2020.

A Proofs

A.1 Derivation of Equation (9)

Recall the definition of the tilted matrix, $\widetilde{P}$, from Equation (5). We substitute in the self-consistent reward function $r(s,a) = -\log u(s,a)v(s,a)$, where u and v are the left and right eigenvectors of $\widetilde{P}$, respectively.

$$\lambda u(s,a) = \sum_{s',a'} u(s',a') \frac{P^{\pi_0}(s',a'|s,a)}{(v(s,a)u(s,a))^\beta} \quad \text{and} \quad \lambda v(s',a') = \sum_{s,a} v(s,a) \frac{P^{\pi_0}(s',a'|s,a)}{(v(s,a)u(s,a))^\beta},$$

which can be written as:

$$\sum_{s',a'} u(s',a') P^{\pi_0}(s',a'|s,a) = \lambda u^{1+\beta}(s,a) v^\beta(s,a) \tag{12}$$

$$\sum_{s,a} P^{\pi_0}(s',a'|s,a) u^{-\beta}(s,a) v^{1-\beta}(s,a) = \lambda v(s',a'). \tag{13}$$

A.2 Derivation of Equation (10)

For ease of notation, let $i, j,$ and k be vector indices which correspond to the sequential state-action pairs $(\bar{s},\bar{a})$, (s,a), and (s',a'), respectively. Accordingly, let P_{ji} denote $P(s,a|\bar{s},\bar{a})$, and let P_{kj} denote $P(s',a'|s,a)$. Equations (12) and (13) are written more compactly as:

$$\sum_k u_k P^{\pi_0}_{kj} = \lambda u_j^{1+\beta} v_j^\beta \tag{14}$$

$$\sum_i P^{\pi_0}_{ji} u_i^{-\beta} v_i^{1-\beta} = \lambda v_j, \tag{15}$$

where for Equation (14) we have also shifted the timestep to align with the right eigenvector in Equation (15). Now, by solving for the right eigenvector in terms of the left, we have, for the two state-action pairs,

$$v_j = \left(\frac{1}{\lambda}\frac{(u^T P)_j}{u_j^{\beta+1}}\right)^{\frac{1}{\beta}} \quad \text{and} \quad v_i^{1-\beta} = \left(\frac{1}{\lambda}\frac{(u^T P)_i}{u_i^{\beta+1}}\right)^{\frac{1-\beta}{\beta}}. \tag{16}$$

Substituting these expressions into Equation (15), we have

$$\sum_i P^{\pi_0}_{ji} u_i^{-\beta} \left(\frac{1}{\lambda}\frac{(u^T P)_i}{u_i^{\beta+1}}\right)^{\frac{1-\beta}{\beta}} = \lambda \left(\frac{1}{\lambda}\frac{(u^T P)_j}{u_j^{\beta+1}}\right)^{\frac{1}{\beta}} \tag{17}$$

The eigenvalue cancels, and we can solve for $u_j = u(s,a)$ as:

$$u_j^{(1+\beta)/\beta} = \frac{\left(\sum_k u_k P^{\pi_0}_{kj}\right)^{1/\beta}}{\sum_i P^{\pi_0}_{ji} u_i^{-1/\beta} (u^T P)_i^{(1-\beta)/\beta}} \tag{18}$$

or, written in terms of state-action pairs, for $\beta = 1$,

$$u(s,a) = \sqrt{\frac{\sum_{s',a'} u(s',a') P(s',a'|s,a)}{\sum_{\bar{s},\bar{a}} u^{-1}(\bar{s},\bar{a}) P(s,a|\bar{s},\bar{a})}}. \tag{19}$$

A.3 Convergence of Equation (10)

Our convergence result requires use of the m-step operator since the corresponding matrix $(P^{\pi_0})^m$ is strictly positive, which ensures the projective diameter κ is finite, which in turn supplies us with a contraction (rather than simply a non-expansive) mapping for $\mathcal{T}$.

Lemma 1. *Let the prior policy during posterior policy iteration (PPI), be denoted $\pi_0(t)$. Then, the index of primitivity, $m(t) = m$, for the matrix $P^{\pi_0(t)}$ is a constant.*

Notably, this result also holds when a softened version of PPI is used: i.e. if a rolling average of policies is used rather than "hard updates", the result remains, because the primitivity is a topological property of the connectivity of the underlying state-action graph. The value of m is based only on the minimum number of steps for which there is a positive probability of transitioning between any two state-action pairs. Since iterations of PPI yield policies with the same support (i.e. absolute continuity holds: $\pi_0(t+1) \ll \pi_0(t)$), various policies cannot affect the number of steps required, only the corresponding probabilities.

Theorem 1. *Let the dynamics induced by $\pi_0(a|s)$ and $p(s'|s,a)$ be irreducible and aperiodic with index of primitivity m. The m-step mapping $u \leftarrow \mathcal{T}_m(u)$ extension of Equation (10) is a contraction under the projective metric, and converges linearly with rate $\kappa((P^{\pi_0})^m)$ when $\beta \geq 1$.*

Proof. We focus on the case of "m-step returns", where the one-step transition matrix P^{π_0} is replaced by $M \doteq (P^{\pi_0})^m$, where m is the smallest integer (number of steps in the MDP) such that $M > 0$. Since $M > 0$, its projective diameter $\Delta(M) = \max_{i,j} d_H(Me_i, Me_j)$ is finite, and by the Birkhoff–Hopf theorem the linear map $x \mapsto Mx$ contracts with coefficient $\tau_m = \tanh\left(\frac{\Delta(M)}{4}\right) < 1$ (cf. "Contraction" section below).

We follow the techniques discussed in Lemmens & Nussbaum (2014). Let $\mathcal{T}(u)$ denote the operator defined by Equation (10), swapping $P_0^\pi \to M$. First we show that $\mathcal{T}$ is positive, homogeneous, and monotone for all $\beta \geq 1$.

Positive Positivity follows trivially since we operate in the positive orthant, so $u(s,a) > 0$ is strictly positive and the application of the positive matrix M maintains positivity. Both the numerator and denominator in Equation (10) are strictly positive for any $u > 0$ and all real values of β.

Homogeneous To show that the operator is homogeneous, we evaluate $\mathcal{T}(cu)$ for a constant $c > 0$:

$$[\mathcal{T}(cu)]_j = \left(\frac{(M(cu))_j^{\frac{1}{\beta}}}{\sum_i M_{ji} (cu)_i^{-\frac{1}{\beta}} (M(cu))_i^{\frac{1-\beta}{\beta}}} \right)^{\frac{\beta}{1+\beta}}$$

Factoring the scalar c out of the linear matrix multiplications yields:

$$[\mathcal{T}(cu)]_j = \left(\frac{c^{\frac{1}{\beta}} (Mu)_j^{\frac{1}{\beta}}}{\sum_i M_{ji} c^{-\frac{1}{\beta}} u_i^{-\frac{1}{\beta}} c^{\frac{1-\beta}{\beta}} (Mu)_i^{\frac{1-\beta}{\beta}}} \right)^{\frac{\beta}{1+\beta}}$$

Next, we consolidate the scalar c in the denominator by combining its exponents throughout:

$$[\mathcal{T}(cu)]_j = \left(c^{\frac{1+\beta}{\beta}} \right)^{\frac{\beta}{1+\beta}} [\mathcal{T}(u)]_j = c[\mathcal{T}(u)]_j$$

Thus, the fixed-point operator $\mathcal{T}$ is homogeneous of degree 1 for all $\beta > 0$.

Monotonicity A sufficient condition for the monotonicity of $\mathcal{T}$ requires $\frac{\partial[\mathcal{T}(u)]_k}{\partial u_i} \geq 0$ for all $u > 0$ and $M \geq 0$. Let $[\mathcal{T}(u)]_k = \left(\frac{N_k(u)}{D_k(u)}\right)^{\delta}$, where:

$$\delta = \frac{\beta}{1+\beta}, \quad N_k(u) = (Mu)_k^{\frac{1}{\beta}}, \quad D_k(u) = \sum_j M_{kj} u_j^{-\frac{1}{\beta}} (Mu)_j^{\frac{1-\beta}{\beta}}.$$

Since $\delta > 0$ for all $\beta > 0$, it is sufficient to show that $\frac{\partial N_k(u)}{\partial u_i} \geq 0$ and $\frac{\partial D_k(u)}{\partial u_i} \leq 0$. For the numerator, applying the chain rule gives:

$$\frac{\partial N_k(u)}{\partial u_i} = \frac{1}{\beta}(Mu)_k^{\frac{1-\beta}{\beta}} M_{ki} \geq 0.$$

For $D(u)$, applying the product and chain rules yields:

$$\frac{\partial D_k(u)}{\partial u_i} = \sum_j M_{kj} \left[\left(-\frac{1}{\beta}\right) u_j^{-\frac{1+\beta}{\beta}} (Mu)_j^{\frac{1-\beta}{\beta}} \delta_{ji} + u_j^{-\frac{1}{\beta}} \left(\frac{1-\beta}{\beta}\right) (Mu)_j^{\frac{1-2\beta}{\beta}} M_{ji} \right].$$

Since this term is in the denominator, the derivative must be non-positive for any $M > 0$. Hence, the scalar coefficients of both terms inside the sum must be non-positive, $\beta \geq 1$.

Contraction To analyze the convergence of $\mathcal{T}$, we operate in the projective space of strictly positive vectors $\mathbb{R}^n_{>0}$. The *projective metric* (also "Hilbert's projective metric") is defined for any $x, y \in \mathbb{R}^n_{>0}$ as:

$$d_H(x, y) \doteq \ln\left(\max_i \frac{x_i}{y_i}\right) - \ln\left(\min_i \frac{x_i}{y_i}\right). \tag{20}$$

This metric is naturally scale-invariant, meaning $d_H(cx, y) = d_H(x, y)$ for any scalar $c > 0$, making it an appropriate metric for the average-reward setting where the differential value function is well-defined up to constant shifts (multiplicative constants in the exponential space that we operate under). We rely on three standard properties of d_H for element-wise operations:

1. **Exponentiation:** $d_H(x^p, y^p) = |p| d_H(x, y)$ for any $p \in \mathbb{R}$.
2. **Multiplication/Division:** $d_H(x \circ u, y \circ w) \leq d_H(x, y) + d_H(u, w)$, where $\circ$ is the Hadamard (element-wise) product.
3. **Linear Contraction (Birkhoff-Hopf Theorem):** For any strictly positive matrix $M > 0$, the linear map $x \to Mx$ acts as a strict contraction: $d_H(Mx, My) \leq \tau d_H(x, y)$, where the Birkhoff contraction coefficient is $\tau = \tanh\left(\frac{\Delta(M)}{4}\right) < 1$, and $\Delta(M)$ is the projective diameter of M.

Further discussion and details of this metric can be found in, e.g., Kohlberg & Pratt (1982).

As established in prior sections, $\mathcal{T}$ is positive, homogeneous of degree 1, and monotonically order-preserving for $\beta \geq 1$ and thus satisfies the hypotheses of Birkhoff's convergence theorem. We can track the Lipschitz constant of $\mathcal{T}$ under d_H.

We decompose the operator $\mathcal{T}(u)$ into its numerator $N(u)$, denominator $D(u)$, and outer exponent $\delta = \frac{\beta}{1+\beta}$, evaluating the metric contraction rate of each part sequentially. Let $x, y \in \mathbb{R}^n_{>0}$ be two positive vectors with initial projective distance $\Delta = d_H(x, y)$.

First, for the numerator $N(u) = (Mu)^{\frac{1}{\beta}}$, we use the rules for positive operators and exponentiation written above:

$$d_H(N(x), N(y)) \leq \frac{1}{\beta} d_H(Mx, My) \leq \frac{\tau}{\beta}\Delta.$$

Second, we evaluate the intermediate vector $f(u)$ inside the denominator's sum, defined element-wise as $f_j(u) = u_j^{-\frac{1}{\beta}}(Mu)_j^{\frac{1-\beta}{\beta}}$. Using the triangle inequality for element-wise multiplication, the scaling property of exponentiation and simplifying:

$$d_H(f(x), f(y)) \leq \left|-\frac{1}{\beta}\right| d_H(x, y) + \left|\frac{1-\beta}{\beta}\right| d_H(Mx, My).$$
$$d_H(f(x), f(y)) \leq \frac{1}{\beta}\Delta + \frac{\beta-1}{\beta}\tau\Delta = \left(\frac{1+\tau(\beta-1)}{\beta}\right)\Delta.$$

The denominator $D(u) = Mf(u)$ applies the linear operator M to $f(u)$, contracting the distance by another factor of τ:

$$d_H(D(x), D(y)) \leq \tau d_H(f(x), f(y)) \leq \tau\left(\frac{1+\tau(\beta-1)}{\beta}\right)\Delta.$$

Combining the numerator and denominator via element-wise division (cf. the Multiplication/Division rule above), the ratio $R(u) = \frac{N(u)}{D(u)}$ yields:

$$\begin{aligned} d_H(R(x), R(y)) &\leq d_H(N(x), N(y)) + d_H(D(x), D(y)) \\ &\leq \left[\frac{\tau}{\beta} + \frac{\tau}{\beta} + \frac{\tau^2(\beta-1)}{\beta}\right]\Delta = \left(\frac{2\tau+\tau^2(\beta-1)}{\beta}\right)\Delta. \end{aligned}$$

Finally, applying the outer exponent $\delta = \frac{\beta}{1+\beta}$ gives the global contraction rate $C(\beta, \tau)$ for the full operator $\mathcal{T}$:

$$\begin{aligned} d_H(\mathcal{T}(x), \mathcal{T}(y)) &\leq \frac{\beta}{1+\beta}\left(\frac{2\tau+\tau^2(\beta-1)}{\beta}\right)\Delta \\ &= \frac{2\tau+\tau^2(\beta-1)}{1+\beta}\Delta, \end{aligned}$$

where the coefficient $C(\beta, \tau) = \frac{2\tau+\tau^2(\beta-1)}{1+\beta} < 1$ since $\tau < 1$.

Therefore, by the Banach fixed-point theorem, the sequence of projective distances $d_H(\mathcal{T}^k(x), \mathcal{T}^k(y)) \leq C(\beta, \tau)^k\Delta$ vanishes as $k \to \infty$, proving convergence to a unique positive projective fixed point u^*. Because the index of primitivity m remains constant across different prior policies during Posterior Policy Iteration (cf. Lemma 1), the rate holds under PPI. Thus, the m-step version of Equation (10) converges linearly under the Hilbert projective metric. We note that the solution to the m-step version of Equation (10) (i.e. replacing $P^{\pi_0} \to (P^{\pi_0})^m$) is the same as the $m = 1$-step version, since both matrices have the same eigenvector u. Empirically, we find that the standard $m = 1$ form also converges as shown in Figure 1.

□

A.4 Posterior Policy Iteration Proof

Here we show that PPI iterations lead to a monotonically increasing entropy-rate.

Proof. Consider, at the n^{th} iteration of PPI, the prior policy π_n and corresponding posterior policy $\pi_{n+1} \propto \pi_n(a|s)u(s, a)$. Denote the entropy-regularized objective as $J(\pi, \pi_n) = \mathbb{E}_\pi(r - \beta^{-1}D_{\text{KL}}(\pi \,\|\, \pi_n))$. Clearly $J(\pi_n, \pi_n) = \rho(\pi_n)$ (average-reward, i.e. entropy-rate, for policy π_n). Now $J(\pi_{n+1}, \pi_n) \geq J(\pi_n, \pi_n) = \rho(\pi_n)$ by definition, since π_{n+1} is the maximizer of the objective. Since the Kullback-Liebler divergence is non-negative, we also have $\rho(\pi_{n+1}) \geq J(\pi_{n+1}, \pi_n)$. Combining the two inequalities, we have $\rho(\pi_{n+1}) \geq \rho(\pi_n)$: in other words, the entropy-rate is monotonically increasing as a function of n. At both steps, equality is obtained only if $\pi_n = \pi_{n+1}$ (at convergence). □