Occupancy Information Ratio: Infinite-Horizon, Information-Directed, Parameterized Policy Search[*]

Wesley A. Suttle,[1] **Alec Koppel,** [2] **Ji Liu** [3]

[1]U.S. Army Research Laboratory [2]J.P. Morgan AI Research [3]Stony Brook University
wesley.a.suttle.ctr@army.mil, alec.koppel@jpmchase.com, ji.liu@stonybrook.edu

Abstract

We develop a new measure of the exploration/exploitation trade-off in infinite-horizon reinforcement learning (RL) problems called the occupancy information ratio (OIR), which is comprised of a ratio between the infinite-horizon average cost of a policy and the entropy of its induced long-term state occupancy measure. Modifying the classic RL objective in this way yields policies that strike an optimal balance between exploitation and exploration, providing a new tool for addressing the exploration/exploitation trade-off in RL. The paper develops for the first time policy gradient and actor-critic algorithms for OIR optimization based upon a new entropy gradient theorem, and establishes both asymptotic and non-asymptotic convergence results with global optimality guarantees. In experiments, these methodologies outperform several deep RL baselines in problems with sparse rewards.

Introduction

Stochastic optimization problems, where the goal is to optimize an objective that takes the form of an expectation, are ubiquitous in machine learning. Classic methods for analyzing stochastic optimization techniques like stochastic gradient descent frequently rely on the well-known connections between stochastic approximation (Robbins and Monro 1951) and dynamical systems (Borkar 2008). Policy optimization in reinforcement learning (RL) (Sutton and Barto 2018) can be viewed as a special form of stochastic optimization problem, where the expectation in the objective is taken over the trajectories induced by an agent's policy when interacting with its environment. In the model-free RL setting, where the agent's environment is unknown, however, the policy optimization problem features an additional, external imperative: the environment must be explored in order to learn an effective policy. This so-called *exploration/-exploitation* trade-off justifies attempts to alter the landscape of the underlying dynamical system by augmenting the corresponding optimization objective to appropriately balance exploration with exploitation.

In control of dynamical systems, this exploration/exploitation trade-off is often addressed by separating the steps of planning and control: first one conducts state estimation (system identification (Åström and Eykhoff 1971)) over a stochastic system, and then solves for actuation parameters in terms of it (Bertsekas and Shreve 1996). The utility of RL-based methods for control of dynamical systems is well-established (Recht 2019; Meyn 2022). Indeed, RL methods for control have been placed on solid theoretical footing for a variety of classic control problems, including for linear quadratic regulators (LQR) (Fazel et al. 2018; Bhandari and Russo 2019; Lale et al. 2022), the linear quadratic Gaussian (LQG) problem (Tang, Zheng, and Li 2021; Uehara et al. 2022), and ensuring safety (Berkenkamp et al. 2017; Cheng et al. 2019). Importantly, many of these works have recognized the importance of balancing exploration of the system with exploitation of current knowledge to achieve critical goals like safe state identification (Berkenkamp et al. 2017), improved sample efficiency (Bhandari and Russo 2019), and improved stability (Lale et al. 2022).

Recently, new approaches to addressing the exploration/-exploitation trade-off in the RL context have led to key theoretical advances. In the context of policy gradient methods, the recent works (Bhandari and Russo 2019; Agarwal et al. 2020a; Mei et al. 2020) have provided new insights into the relationships between system exploration, convergence rates, and global optimality. A key insight from these works is that the right kind of exploration accelerates convergence without sacrificing global optimality. The prior works (Russo and Van Roy 2014; Lu et al. 2021) in multi-armed bandits (MABs) and RL, on the other hand, seek to balance the goals of exploration and exploitation explicitly: by minimizing an *information ratio*, defined as the ratio of cost incurred – formulated as regret – to information acquired. A key insight of these works is that explicitly optimizing the rate of reward accrued per quantity information acquired about the system leads to more intelligent exploration behaviors and improved regret. Despite the advantages of information ratio-based techniques and the recent advances in the theory of policy gradient methods, however, the development of policy gradient methods for information ratio objectives remains unexplored.

Our goal is to develop information ratio optimization approaches for infinite-horizon RL problems that can operate

[*]All proofs and additional numerical studies have been omitted due to space considerations, but are available upon request and will be contained in a forthcoming publication.

in high-dimensional, possibly continuous spaces. The information ratio of (Lu et al. 2021) considers information gain of a policy over a fixed time-horizon, which in the infinite-horizon setting requires conditioning over an infinite trajectory. Moreover, the deep Q-learning-based methods proposed in (Lu et al. 2021) are inherently restricted to the finite action space case, limiting their scalability. To improve this scaling, operating in parameter space instead is required, for which policy gradient methods are most natural (Lillicrap et al. 2015; Schulman et al. 2017; Haarnoja et al. 2018). Our goal thus requires a definition of informativeness that is amenable to policy search in parameter space. Occupancy measure entropy has recently been used as an optimization objective (Hazan et al. 2019; Lee et al. 2019; Zhang et al. 2020a) that captures the amount of information about the environment that a policy provides through the Kullback–Leibler divergence of its state occupancy measure (also called state marginal distribution) from a uniform distribution. This motivates us to take occupancy measure entropy, or *occupancy information*, of a policy as the fundamental quantity defining its informativeness.

Based on this definition, we develop a new objective called the *occupancy information ratio*, or OIR, which captures the exploration/exploitation trade-off as defined by the ratio of long-term average cost to occupancy information of a policy. We underscore that **OIR is the first RL objective to concisely trade off exploration and exploitation and to which parameterized policies also naturally apply**, making it the first to scale well to large spaces. Consequently, we are able to derive policy gradient (PG) and actor-critic (AC) algorithms to optimize the OIR in parameter space. Moreover, through connections to quasiconcave programming and a novel application of the perspective transform that arises in fractional programming, we establish that this objective has no spurious extrema (Theorem 6), extending the hidden concavity arguments of (Zhang et al. 2020a) to *hidden quasiconcavity*. Hence our newly derived PG and AC algorithms exhibit convergence to global optimality as established in Theorems 8 and 11. In experiments, we illustrate that the OIR yields policies that avoid spurious, suboptimal behavior in practice, whereas benchmarks exhibit a tendency to become stuck. Due to the importance of the exploration/exploitation trade-off in achieving critical goals like safety, improved stability, and sample efficiency, we believe OIR-based methods are of interest to those working at the intersection of learning and dynamical systems.

Problem Formulation

In this section we describe our problem setting and formulate the occupancy information ratio objective. We first define an underlying Markov decision process, then formulate the OIR as an objective to be optimized over it.

Markov Decision Processes. Consider an average-cost MDP described by the tuple $(\mathcal{S}, \mathcal{A}, p, c)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $p : \mathcal{S} \times \mathcal{A} \to \mathcal{D}(\mathcal{S})$ is the transition probability kernel mapping state-action pairs to distributions over the state space, and $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ is the cost function mapping state-action pairs to positive scalars. In this setting, at time-step t, the agent is in state s_t, chooses an action a_t according to a policy $\pi : \mathcal{S} \to \mathcal{D}(\mathcal{A})$ mapping states to distributions over $\mathcal{A}$, incurs cost $c(s_t, a_t)$, and then the system transitions into a new state $s_{t+1} \sim p(\cdot|s_t, a_t)$. Since we are interested in policy gradient methods, we give the following definitions with respect to a parameterized family $\{\pi_\theta : \mathcal{S} \to \mathcal{D}(\mathcal{A})\}_{\theta \in \Theta}$ of policies, where $\Theta \subset \mathbb{R}^d$ is some set of permissible policy parameters. Note that analogous definitions apply to any policy π. For any $\theta \in \Theta$, let $d_\theta(s) = \lim_{t\to\infty} P(s_t = s \mid \pi_\theta)$ denote the steady-state occupancy measure over $\mathcal{S}$ induced by π_θ, which we assume to be independent of the initial start-state. In addition, let $\lambda_\theta(s, a) = \lim_{t\to\infty} P(s_t = s, a_t = a \mid \pi_\theta)$ denote the state-action occupancy measure induced by π_θ over $\mathcal{S} \times \mathcal{A}$. Notice that $\lambda_\theta(s, a) = d_\theta(s)\pi_\theta(a|s)$. Furthermore, let $J(\theta) = \sum_s d_\theta(s) \sum_a \pi_\theta(a|s)c(s, a)$ denote the long-run average cost of using policy π_θ. Finally, given θ, define the entropy of the state occupancy measure induced by π_θ to be $H(d_\theta) = -\sum_s d_\theta(s) \log d_\theta(s)$. This quantity measures how well π_θ covers the state space $\mathcal{S}$ in the long run.

Occupancy Information Ratio. In this paper we consider the OIR objective

$$\rho(\theta) = \frac{J(\theta)}{\kappa + H(d_\theta)}, \tag{1}$$

where $\kappa > -\min_\theta H(d_\theta)$ is a user-specific constant, discussed in Remark 1. Given an MDP $(\mathcal{S}, \mathcal{A}, p, c)$, our goal is to find a policy parameter θ^* such that π_{θ^*} minimizes (1) *over the MDP*, i.e., subject to its costs and dynamics. As $J(\theta)$ and $H(d_\theta)$ are both long-run, infinite-horizon quantities, we regard (1) as an *infinite-horizon objective*.

Remark 1. *Since κ scales the relative importance of $H(d_\theta)$ in* (1)*, it can be viewed (and used) as a regularizer. When minimizing a function $f(x)$, one frequently considers a regularized objective function $f(x) + \kappa \|x\|$, where $\kappa \geq 0$. Here, the larger κ becomes, the more important the regularization term becomes with respect to the objective function. In contrast, for* (1)*, the relative importance of the entropy term actually diminishes as κ becomes larger: when κ is small, even minor changes in the value of $H(d_\theta)$ can have a large effect on the value of $\rho(\theta)$; when κ is large, on the other hand, even significant perturbations of the value of $H(d_\theta)$ have little effect on the value of $\rho(\theta)$.*

Remark 2. *Though we stipulated that $\kappa > -\min_\theta H(d_\theta)$ in the definition of the OIR above, letting $\kappa < -\max_\theta H(d_\theta)$ has an important interpretation as well. When $\kappa < -\max_\theta H(d_\theta)$ and $J(\theta) \geq 0$, for all $\theta \in \Theta$, clearly the OIR $\rho(\theta)$ will always be non-positive. Because of this, minimizing the OIR will in fact minimize the ratio of $-J(\theta)$ to the absolute value $|\kappa + H(d_\theta)|$. This means that the expected cost $J(\theta)$ of the underlying MDP is instead treated as an expected reward to be maximized, and any algorithm for minimizing the OIR will therefore balance maximizing the reward $J(\theta)$ with maximizing the shifted entropy $|\kappa + H(d_\theta)|$. This allows the OIR framework to accommodate rewards by simply replacing the cost function c in the MDP with a reward function r, and choosing $\kappa < -\max_\theta H(d_\theta)$.*

Policy Gradients

Sampling the gradient of (1) is not straightforward using existing tools, as obtaining stochastic estimates of $\nabla \rho(\theta)$ involves estimating

$$\nabla \rho(\theta) = \frac{\nabla J(\theta)(\kappa + H(d_\theta)) - J(\theta)\nabla H(d_\theta)}{[\kappa + H(d_\theta)]^2}. \tag{2}$$

Though we can use the classical policy gradient theorem [cf. Eq. (3)] to estimate $\nabla J(\theta)$ and we can empirically estimate $J(\theta)$ and $H(d_\theta)$, it is not obvious how to estimate $\nabla H(d_\theta)$. In what follows we prove an *entropy gradient theorem* that allows us to estimate $\nabla H(d_\theta)$ and consequently $\nabla \rho(\theta)$.

Policy Gradient Preliminaries. Given an MDP $(\mathcal{S}, \mathcal{A}, p, c)$ and policy π_θ, two important objects from the RL literature are the relative state value function $V_\theta(s) = \sum_{t=0}^{\infty} \mathbb{E}_{\pi_\theta}\left[c(s,a) - J(\theta) \mid s_0 = s\right]$ and the relative action value function $Q_\theta(s,a) = \sum_{t=0}^{\infty} \mathbb{E}_{\pi_\theta}\left[c(s,a) - J(\theta) \mid s_0 = s, a_0 = a\right]$. Under the assumption that $\pi_\theta(a|s)$ is differentiable in θ, for all $s \in \mathcal{S}, a \in \mathcal{A}$, classic policy gradient methods minimize $J(\theta)$ by taking stochastic gradient descent steps in the direction $-\nabla J(\theta)$. We are guaranteed by the policy gradient theorem (Sutton et al. 1999) that, under certain conditions,

$$\begin{aligned} \nabla J(\theta) &= \sum_s d_\theta(s) \sum_a Q_\theta(s,a) \nabla \pi_\theta(a|s) \\ &= \mathbb{E}_{\pi_\theta}\Big[(c(s,a) - J(\theta))\nabla \log \pi_\theta(a|s)\Big]. \end{aligned} \tag{3}$$

By following policy π_θ, we can sample from the right-hand side of (3) to estimate $\nabla J(\theta)$, then use this to perform stochastic gradient descent.

Cross-Entropy Gradient. To estimate $\nabla \rho(\theta)$ we must know how to estimate $\nabla H(d_\theta)$. Fortunately, by using the relationship between entropy and cross-entropy, $\nabla H(d_\theta)$ can be estimated in a straightforward manner. Given two policy parameters θ and θ', define the cross-entropy between d_θ and $d_{\theta'}$ to be $CE(d_\theta, d_{\theta'}) = -\sum_s d_\theta(s) \log d_{\theta'}(s)$ and their Kullback-Leibler (KL) divergence to be $D_{KL}(d_\theta \mid\mid d_{\theta'}) = \sum_s \log\left(\frac{d_\theta(s)}{d_{\theta'}(s)}\right) d_\theta(s)$. Also recall the useful fact that $CE(d_\theta, d_{\theta'}) = H(d_\theta) + D_{KL}(d_\theta \mid\mid d_{\theta'})$. We now have:[1]

Lemma 1. *For any $\theta' \in \Theta$,*

$$\nabla H(d_\theta)\big|_{\theta=\theta'} = \nabla CE(d_\theta, d_{\theta'})\big|_{\theta=\theta'}. \tag{4}$$

This establishes an important fact: *we can estimate the entropy gradient $\nabla H(d_\theta)|_{\theta=\theta_t}$ by instead estimating the cross-entropy gradient $\nabla CE(d_\theta, d_{\theta_t})|_{\theta=\theta_t}$*. At first glance, this simply substitutes one problem for another. However, given a fixed θ_t, for any θ, we can use the classic policy gradient theorem (3) to obtain a tractable expression for $\nabla CE(d_\theta, d_{\theta_t})|_{\theta=\theta_t}$, as described next.

Entropy and OIR Policy Gradients. Our next results provide tractable gradient expressions enabling policy gradient algorithms for maximizing $H(d_\theta)$ and minimizing (1).

Theorem 2. *Let an MDP $(\mathcal{S}, \mathcal{A}, p, c)$ and a differentiable parametrized policy class $\{\pi_\theta\}_{\theta\in\Theta}$ be given, and recall the definition above of the state occupancy measure d_θ induced by π_θ on $\mathcal{S}$. Fix a policy parameter iterate θ_t at time-step t. The gradient $\nabla H(d_\theta)|_{\theta=\theta_t}$ [cf. (4)] with respect to the policy parameters θ of the state occupancy measure entropy $H(d_\theta)$, evaluated at $\theta = \theta_t$, satisfies $\nabla H(d_\theta)\big|_{\theta=\theta_t} =$*

$$\mathbb{E}_{\pi_{\theta_t}}\Big[\left(-\log d_{\theta_t}(s) - H(d_{\theta_t})\right)\nabla \log \pi_{\theta_t}(a|s)\Big]. \tag{5}$$

With Theorem 2 in hand, we have the following OIR policy gradient theorem:

Theorem 3. *Let an MDP $(\mathcal{S}, \mathcal{A}, p, c)$, a differentiable parametrized policy class $\{\pi_\theta\}_{\theta\in\Theta}$, and a constant $\kappa \geq 0$ be given, and recall the definitions of the average cost $J(\theta)$, state occupancy measure d_θ, and entropy $H(d_\theta)$. Fix a policy parameter iterate θ_t at time-step t. The gradient $\nabla \rho(\theta_t)$ [cf. (2)] with respect to the policy parameters θ of the OIR $\rho(\theta)$ [cf. (1)], evaluated at $\theta = \theta_t$, satisfies $\nabla \rho(\theta_t) =$*

$$\mathbb{E}_{\pi_{\theta_t}}\left[\frac{\delta_t^J\left(\kappa + H(d_{\theta_t})\right) - J(\theta_t)\delta_t^H}{[\kappa + H(d_{\theta_t})]^2}\nabla \log \pi_{\theta_t}(a|s)\right], \tag{6}$$

where $\delta_t^J = c(s,a) - J(\theta_t), \delta_t^H = -\log d_{\theta_t}(s) - H(d_{\theta_t})$.

The claim follows by combining equations (2) and (3) with Theorem 2. Armed with Theorem 3, we next develop policy gradient algorithms for minimizing the OIR.

Algorithms

In this section we derive two policy search schemes for minimizing (1). Throughout this section, we will assume that an average-cost MDP $(\mathcal{S}, \mathcal{A}, p, c)$ is fixed. The reward setting can be accommodated with minor changes by Remark 2.

Information-Directed REINFORCE. We present Information-Directed REINFORCE (ID-REINFORCE), which builds on the class REINFORCE algorithm (Williams 1992) to minimize the more complicated objective (1). At each time-step t, the algorithm generates a trajectory using the current policy π_{θ_t}. It then forms estimates of $J(\theta_t)$ and $H(d_{\theta_t})$ and in turn uses these to estimate $\nabla \rho(\theta_t)$ by leveraging (6). This gradient estimate is then used to update the policy parameters. Note that, in order to estimate $H(d_{\theta_t})$, it is necessary to first estimate d_{θ_t}. This task is addressed both implicitly and explicitly in previous works (Hazan et al. 2019; Lee et al. 2019; Zhang et al. 2020a). As in (Hazan et al. 2019), for ease of exposition we assume access to an oracle DensityEstimator that returns the occupancy measure $d_\theta = \text{DensityEstimator}(\theta)$ when provided with input policy parameter $\theta \in \Theta$. When $\mathcal{S}$ is finite and not too large, DensityEstimator can be implemented with a count-based estimator. We focus on this setting in this paper. Pseudocode for ID-REINFORCE can be found in Algorithm 1.

Information-Directed Actor-Critic. We next present the Information-Directed Actor-Critic (IDAC) algorithm, a variant of the classic actor-critic algorithm (Konda 2002; Bhatnagar et al. 2009) with two critics: the standard critic corresponding to average cost $J(\theta)$, and an entropy critic corresponding to the shadow MDPs $(\mathcal{S}, \mathcal{A}, p, r_t)$, $t \geq 0$, where

[1] For a function $f : \Theta \to \mathbb{R}$, we sometimes write $\nabla f(\theta)|_{\theta=\theta_t}$ to emphasize the fact that the gradient of f w.r.t. θ is being taken first, then subsequently evaluated at $\theta = \theta_t$.

Algorithm 1: ID-REINFORCE

```
Initialization: Select rollout length $K$, step-sizes $\eta > 0$ and $\tau \in (0, 1]$, parametrized policy class $\{\pi_\theta\}_{\theta \in \Theta}$, and entropy additive constant $\kappa \geq 0$. Randomly sample $s_0$ and $\theta_0$, select $\mu^H_{-1}, \mu^J_{-1} > 0$, and set $t \leftarrow 0$.
repeat
    Generate trajectory $\{(s_i, a_i)\}_{i=1,\dots,K}$ using $\pi_{\theta_t}$
    $\widehat{J(\theta_t)} = \frac{1}{K}\sum_{i=1}^{K} c(s_i, a_i)$
    $\mu^J_t = (1-\tau)\mu^J_{t-1} + \tau\widehat{J(\theta_t)}$
    $d_{\theta_t} = \text{DensityEstimator}(\theta_t)$
    $\widehat{H(d_{\theta_t})} = \frac{1}{K}\sum_{i=1}^{K}(-\log d_{\theta_t}(s_i))$
    $\mu^H_t = (1-\tau)\mu^H_{t-1} + \tau\widehat{H(d_{\theta_t})}$
    for $i = 1, \dots, k$ do
        $\delta^J_i = c(s_i, a_i) - \mu^J_t$
        $\delta^H_i = -\log d_{\theta_t}(s_i) - \mu^H_t$
        $\psi_i = \nabla \log \pi_{\theta_t}(a_i|s_i)$
    end for
    $\widehat{\nabla\rho(\theta_t)} = \frac{1}{[\kappa+\mu^H_t]^2}\frac{1}{K}\sum_{i=1}^{K}\left[\delta^J_i\left(\kappa+\mu^H_t\right) - \mu^J_t\delta^H_i\right]\psi_i$
    $\theta_{t+1} = \theta_t - \eta\widehat{\nabla\rho(\theta_t)}$
    $t \leftarrow t+1$
until convergence
```

Algorithm 2: IDAC

```
Initialization: Select rollout length $K$, stepsize sequences $\{\alpha_t\}, \{\beta_t\}, \{\tau_t\}$, parametrized policy class $\{\pi_\theta\}_{\theta\in\Theta}$, parametrized critic class $\{v_\omega\}_{\omega\in\Omega}$, and entropy additive constant $\kappa \geq 0$. Randomly sample $s_0, \theta_0, \omega^J_0, \omega^H_0$, select $\mu^H_{-1}, \mu^J_{-1} > 0$, and set $t \leftarrow 0$.
repeat
    Generate trajectory $\{(s_i, a_i)\}_{i=1,\dots,K}$ using $\pi_{\theta_t}$
    $\mu^J_t = (1-\tau)\mu^J_{t-1} + \tau\frac{1}{K}\sum_{i=1}^{K} c(s_i, a_i)$
    $d_{\theta_t} = \text{DensityEstimator}(\theta_t)$
    $\mu^H_t = (1-\tau)\mu^H_{t-1} + \tau\frac{1}{K}\sum_{i=1}^{K}(-\log d_{\theta_t}(s_i))$
    for $i = 1, \dots, K$ do
        Set $v_{\omega^J_t}(s_{K+1}) = v_{\omega^H_t}(s_{K+1}) = 0$
        $\delta^J_i = c(s_i, a_i) - \mu^J_t + v_{\omega^J_t}(s_{i+1}) - v_{\omega^J_t}(s_i)$
        $\delta^H_i = -\log d_{\theta_t}(s_i) - \mu^H_t + v_{\omega^H_t}(s_{i+1}) - v_{\omega^H_t}(s_i)$
        $\psi_i = \nabla \log \pi_{\theta_t}(a_i|s_i)$
    end for
    $\omega^J_{t+1} = \omega^J_t + \alpha\frac{1}{K}\sum_{i=1}^{K}\delta^J_i\nabla v_{\omega^J_t}(s_i)$
    $\omega^H_{t+1} = \omega^H_t + \alpha\frac{1}{K}\sum_{i=1}^{K}\delta^H_i\nabla v_{\omega^H_t}(s_i)$
    $\widehat{\nabla\rho(\theta_t)} = \frac{1}{[\kappa+\mu^H_t]^2}\frac{1}{K}\sum_{i=1}^{K}\left[\delta^J_i\left(\kappa+\mu^H_t\right) - \mu^J_t\delta^H_i\right]\psi_i$
    $\theta_{t+1} = \theta_t - \beta\widehat{\nabla\rho(\theta_t)}$
    $t \leftarrow t+1$
until convergence
```

$r_t(s, a) = -\log d_{\theta_t}(s)$ is the shadow reward discussed in the proof of Theorem 2. We assume access to the DensityEstimator oracle throughout. For IDAC, we modify the classic actor-critic scheme by: (i) introducing an entropy critic to estimate the entropy gradient, and (ii) altering the policy update to take a gradient descent step in the direction $-\nabla\rho(\theta_t)$ instead of $-\nabla J(\theta_t)$. At time-step t, the algorithm computes two different TD errors: one corresponds to the critic for the MDP $(\mathcal{S}, \mathcal{A}, p, c)$, while the other corresponds to the critic for the shadow MDP $(\mathcal{S}, \mathcal{A}, p, r_t)$. Next, the cost and entropy critic TD errors are used to update their respective critics, which are in turn combined to perform the actor update. Pseudocode for IDAC can be found in Algorithm 2.

Density Estimation Issue. As discussed above, these algorithms estimate the state density with a count-based estimator, which can be inefficient in continuous, high-dimensional spaces. There are two main options for overcoming this issue. First, more sophisticated density estimation procedures may be used to directly estimate the state occupancy measure in continuous, higher-dimensional spaces. Recent works including (Hazan et al. 2019; Lee et al. 2019) have effectively leveraged kernel density estimation and variational autoencoders to maximize state entropy in the RL setting. These techniques can be extended to the OIR setting to allow our methods to handle continuous, higher-dimensional state spaces. Second, eliminating the need to perform density estimation altogether by directly estimating state entropy is another viable option. Indeed, recent works in unsupervised RL have leveraged particle-based entropy estimation techniques from statistics (Singh et al. 2003) to efficiently perform maximum state entropy exploration in continuous, high-dimensional domains (Yarats et al. 2021; Liu and Abbeel 2021; Mutti, Pratissoli, and Restelli 2021). These methods can be extended to obtain practical OIR methods that avoid density estimation altogether. Building on these techniques, the density estimation issue can be mitigated or eliminated, providing an important future direction.

Theoretical Results

In this section we provide the following key results underpinning policy search for the OIR problem: all stationary points of $\rho(\theta)$ are in fact global minimizers; the stochastic gradient descent scheme underlying ID-REINFORCE enjoys a non-asymptotic convergence rate depending on κ, the policy class, and ergodicity properties of the underyling MDP; IDAC enjoys asymptotic, almost sure (a.s.) convergence to a neighborhood of a stationary point. Taken together, these results prove that both algorithms converge to globally optimal solutions under suitable conditions.

Stationarity Implies Global Optimality

As we will see, the OIR optimization problem enjoys a powerful *hidden quasiconcavity* property: under certain conditions on the set Θ and the policy class $\{\pi_\theta\}_{\theta\in\Theta}$, stationary points of $\rho(\theta)$ correspond to global optima of the OIR minimization problem

$$\min_{\theta \in \Theta} \quad \rho(\theta) = \frac{J(\theta)}{\kappa + H(d_\theta)}. \tag{7}$$

This result is surprising, as the objective function $\rho(\theta)$ is typically highly non-convex. Let $\Theta \subset \mathbb{R}^k$ be convex and let a parametrized policy class $\{\pi_\theta\}_{\theta\in\Theta}$ be given. Let $\lambda : \Theta \to \mathcal{D}(\mathcal{S} \times \mathcal{A})$ be a function mapping each parameter vector $\theta \in$

Θ to the state-action occupancy measure $\lambda(\theta) := \lambda_\theta := \lambda_{\pi_\theta}$ induced by the policy π_θ over $\mathcal{S} \times \mathcal{A}$. We make the following assumptions.

Assumption 4. *The set Θ is compact. For any $s \in \mathcal{S}, a \in \mathcal{A}$, the function $\pi_\theta(a|s)$ is continuously differentiable with respect to θ on Θ, and the Markov chain induced by π_θ on $\mathcal{S}$ is ergodic.*

Assumption 5. *The following statements hold:*
***1.** $\lambda(\cdot)$ gives a bijection between Θ and its image $\lambda(\Theta)$, and $\lambda(\Theta)$ is compact and convex.*
***2.** Let $h(\cdot) := \lambda^{-1}(\cdot)$ denote the inverse mapping of $\lambda(\cdot)$. $h(\cdot)$ is Lipschitz continuous.*
***3.** The Jacobian matrix $\nabla\lambda(\theta)$ is Lipschitz on Θ.*

We have the following theorem.

Theorem 6. *Let Assumptions 4 and 5 hold. Let θ^* be a stationary point of* (7)*, i.e., $\nabla\rho(\theta^*) = 0$. Then θ^* is globally optimal for* (7)*.*

This powerful hidden quasiconcavity property implies that any policy gradient algorithm that can be shown to converge to a stationary point of the OIR optimization problem $\min_{\theta\in\Theta} \rho(\theta)$ in fact converges to a global optimum. This greatly strengthens the convergence results provided next by guaranteeing that they apply to *global* optima. In contrast to the global optimality guarantees for tabular, softmax policy search established in (Bhandari and Russo 2019; Agarwal et al. 2020b; Mei et al. 2020; Zhang et al. 2020a; Bedi et al. 2021) using persistent exploration conditions, our result instead builds on hidden concavity arguments from (Zhang et al. 2020a), which apply to parameterized policies. However, Theorem 6 generalizes these results in important ways. First, it applies to ratio objectives, which have not been addressed in prior work. In addition, we establish hidden *quasiconcavity* for ratio objectives, not hidden *concavity*, which requires reformulation via a novel application of the perspective transform. Theorem 6 is thus a strict generalization of existing results for the landscape of RL objectives.

Non-Asymptotic Convergence Rate

Next, we establish a non-asymptotic convergence rate for the following projected gradient descent scheme for solving the OIR minimization problem (7):

$$\begin{aligned} \theta_{t+1} &= \mathrm{Proj}_\Theta\left(\theta_t - \eta\nabla\rho(\theta_t)\right) \qquad (8) \\ &= \arg\min_\theta [\rho(\theta_t) + \langle\nabla\rho(\theta_t), \theta - \theta_t\rangle + \frac{1}{2\eta}\left\|\theta - \theta_t\right\|^2], \end{aligned}$$

for a fixed stepsize $\eta > 0$, where Proj_Θ denotes euclidean projection onto Θ and the second equality holds by the convexity of Θ. Note that (8) is a reformulation of ID-REINFORCE with null gradient estimation error and projection onto the set Θ; we assume the projection operation for the purposes of analysis, and we discuss the gradient estimation issue at the end of this subsection.

Let $\Theta \subset \mathbb{R}^k$, $\{\pi_\theta\}_{\theta\in\Theta}$, and $\lambda : \Theta \to \mathcal{D}(\mathcal{S}\times\mathcal{A})$ be as in the previous section. Consider the mapping $\zeta : \mathcal{D}(\mathcal{S}\times\mathcal{A}) \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|+1}$, defined to be $\zeta(\lambda) = (\lambda/c^\top\lambda, 1/c^\top\lambda)$, where $c \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $c > 0$ is a vector of positive costs. Notice that, under the ergodicity conditions in Assumption 4 and properties of entropy, $\min_\theta \rho(\theta) > 0$ and $\max_\theta \rho(\theta) < \infty$. In addition to Assumptions 4 and 5, we will need the following.

Assumption 7. *$\nabla\rho(\theta)$ is Lipschitz and $L > 0$ is the smallest number such that $\left\|\nabla\rho(\theta) - \nabla\rho(\theta')\right\| \leq L\left\|\theta - \theta'\right\|$, for all $\theta, \theta' \in \Theta$.*

We have the following convergence rate result for the projected gradient descent scheme (8).

Theorem 8. *Let Assumptions 4, 5, and 7 hold. Let $D_\zeta = \max_{z,z'\in(\zeta\circ\lambda)(\Theta)} \left\|z - z'\right\|$ denote the diameter of the convex, compact set $(\zeta\circ\lambda)(\Theta)$. Define $M = \max_{\theta\in\Theta}\rho(\theta)$, $m = \min_{\theta\in\Theta}\rho(\theta)$, $K = \max\{m^2L, M^2m^2L\}$, and $L_1 = \max\{L, M^2L\}$. Then, with $\eta = 1/K$, for all $t \geq 0$,*

$$\rho(\theta_t) - \rho(\theta^*) \leq \frac{4M^2L_1\ell^2D_\zeta^2}{t+1}. \qquad (9)$$

Coupled with Theorem 6, this result provides a non-asymptotic convergence rate to *global optimality* for algorithms solving the OIR minimization problem (7).

Remark 3. *When compared with the corresponding result in (Zhang et al. 2020a), to which it is closely related, the bound* (9) *of Theorem 8 contains an interesting dependence on the user-specified κ, the policy class $\{\pi_\theta\}_{\theta\in\Theta}$, and the underlying MDP. The presence of $M = \max_{\theta\in\Theta}\rho(\theta) = \max_\theta[J(\theta)/(\kappa + H(d_\theta))]$ in the bound* (9) *suggests that the convergence rate depends on the value of κ as well as the minimal possible value of $H(d_\theta)$ over $\theta \in \Theta$. To see why, let $C = \max_{\theta\in\Theta} J(\theta)$ and notice that*

$$M \leq \max_{\theta\in\Theta} \frac{C}{\kappa + H(d_\theta)} = \frac{C}{\kappa + \min_{\theta\in\Theta} H(d_\theta)}. \qquad (10)$$

On the one hand, when the MDP dynamics and policy class are such that $\min_{\theta\in\Theta} H(d_\theta)$ is large, then M will be closer to 0, yielding a tighter bound in (9)*. This suggests that it may be easier to optimize the OIR over MDPs and/or policy classes that tend to be "more ergodic". When both κ and $\min_{\theta\in\Theta} H(d_\theta)$ are close to 0, on the other hand, M may be very large, resulting in a looser bound in* (9)*. This highlights the usefulness of κ, as choosing larger κ values can be used to smooth the objective function $\rho(\theta)$ and thereby lead to stabler convergence when optimizing the OIR over MDPs and policy classes that tend to be "less ergodic".*

In the preceding theorem, we assume "exact policy gradient," or zero stochastic approximation error. This is limited to Theorem 8, whereas Theorem 11 below allows stochastic approximation error and Theorem 6 above is independent of estimation issues. Though this assumption is a drawback for Theorem 8, it allows us to succinctly focus on a core insight of this work: hidden quasiconcavity unlocks an information-dependent convergence rate to global optimality. We also note that, for REINFORCE-like algorithms like those considered in Theorem 8, long rollouts enable more accurate gradient estimates, for which the existing assumptions approximately apply. A precise treatment of gradient estimation error versus rollout length is an important direction future work, and can be achieved by extending the analysis in (Zhang et al. 2021) to the OIR problem.

Actor-Critic Convergence

We conclude this section by proving almost sure (a.s.) convergence of IDAC to a neighborhood of a stationary point of (7). By Theorem 6, this implies IDAC converges a.s. to a neighborhood of a *global* optimum. This is *much* stronger than existing asymptotic results for actor-critic schemes, which typically guarantee convergence to a neighborhood of a local optimum or saddle point (Bhatnagar et al. 2009; Zhang et al. 2020b; Agarwal et al. 2020b).

We analyze the algorithm as given in Algorithm 2 under the assumption that $\tau_t = \alpha_t$, for all $t \geq 0$, that $K = 1$, and with the addition of a projection operation to the policy update:

$$\theta_{t+1} = \Gamma\Big[\theta_t - \beta_t \frac{\delta_t^J(\kappa + \mu_t^H) - \mu_t^J \delta_t^H}{\left(\kappa + \mu_t^H\right)^2} \nabla \log \pi_{\theta_t}(a_t|s_t)\Big], \tag{11}$$

where $\Gamma : \mathbb{R}^d \to \Theta$ maps any parameter $\theta \in \mathbb{R}^d$ back onto the compact set $\Theta \subset \mathbb{R}^d$ of permissible policy parameters. This projection, which is common in the actor-critic and broader two-timescale stochastic approximation literatures (see, e.g., (Kushner and Yin 2003; Borkar 2008; Bhatnagar et al. 2009)) is for purposes of theoretical analysis, and is typically not needed in practice. In addition to Assumption 4, we impose the following:

Assumption 9. *Stepsizes* $\{\alpha_t\}, \{\beta_t\}$ *satisfy* $\sum_t \alpha_t = \sum_t \beta_t = \infty,\ \sum_t \alpha_t^2 + \beta_t^2 < \infty, \lim_t \frac{\beta_t}{\alpha_t} = 0$.

Assumption 10. *The value function approximators* v_ω *are linear, i.e.,* $v_\omega(s) = \omega^\top \phi(s)$*, where* $\phi(s) = [\phi_1(s) \cdots \phi_K(s)]^\top \in \mathbb{R}^K$ *is the feature vector associated with* $s \in \mathcal{S}$*. The feature vectors* $\phi(s)$ *are uniformly bounded for any* $s \in \mathcal{S}$*, and the feature matrix* $\Phi = [\phi(s)]_{s\in\mathcal{S}}^\top \in \mathbb{R}^{|\mathcal{S}|\times K}$ *has full column rank. For any* $u \in \mathbb{R}^K$*,* $\Phi u \neq \mathbf{1}$*, where* $\mathbf{1}$ *is the vector of all ones.*

Assumptions 4, 9, and 10 are standard in two-timescale convergence analyses for actor-critic algorithms (Bhatnagar et al. 2009). Moreover, we consider neural network parameterizations in our experiments and observe favorable convergence behavior, so we believe Assumption 10 can be relaxed. Consider the ordinary differential equation (ODE)

$$\dot{\theta} = \hat{\Gamma}(\nabla \rho(\theta)), \tag{12}$$

where $\hat{\Gamma}(\nabla\rho(\theta)) := \lim_{\eta\to 0^+} \left[\gamma\left(\theta + \eta\nabla\rho(\theta)\right) - \theta\right]/\eta$. We note here that (12) can be interpreted as the projected ODE $\dot{\theta} = \nabla\rho(\theta) + z(\theta)$, where $z(\theta)$ is the minimal force necessary to project θ back onto Θ. We now present the main result of this subsection, which establishes convergence of the actor-critic algorithm.

Theorem 11. *Let* $\mathcal{Z}$ *denote the set of asymptotically stable equilibria of the ODE* (12)*. Given any* $\varepsilon > 0$*, define* $\mathcal{Z}^\varepsilon = \{z \mid \inf_{z'\in\mathcal{Z}} \|z - z'\| \leq \varepsilon\}$*. For any* $\theta \in \Theta$*, let* $\varepsilon_\theta = (\epsilon_\theta^J\left[\kappa + H(d_\theta)\right] - J(\theta)\epsilon_\theta^H)/(\left[\kappa + H(d_\theta)\right]^2)$*. Under Assumptions 4, 9, and 10, given any* $\varepsilon > 0$*, there exists* $\delta > 0$ *such that, for* $\{\theta_t\}$ *obtained from Algorithm 2 with projection* (11)*, if* $\sup_t \|\epsilon_{\theta_t}\| < \delta$*, then* $\theta_t \to \mathcal{Z}^\varepsilon$ *a.s. as* $t \to \infty$*.*

Combined with Theorem 6, Theorem 11 establishes almost sure convergence of IDAC to a neighborhood of a *global* optimum of the OIR minimization problem (7). Note that if the linear approximation and features are expressive enough, then ε will be small or even zero.

Experiments

The experiments presented in this section demonstrate that OIR policy gradient methods avoid spurious behavior, while state-of-the-art methods can become overconfident and settle into suboptimality. In particular, when the reward signal is sparse, OIR methods can lead to improved performance when compared with vanilla RL methods. For Fig. 1, we first compared a neural network version of IDAC with the Stable Baselines 3 implementations (Raffin et al. 2019) environment depicted in Fig. 2 in the appendix; for Fig. 3 presented in the appendix, we also compared tabular versions of IDAC and vanilla actor-critic with softmax policies on the `GridWorld2` depicted in Fig. 2. Details on the gridworld implementations are provided in the appendix.

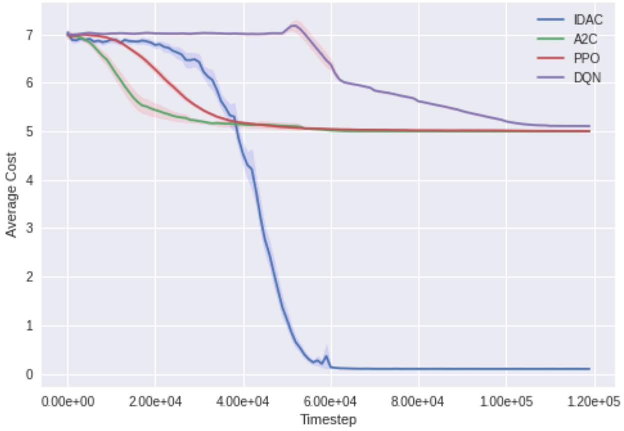

Figure 1: Comparison of neural network IDAC with common deep RL methods on the sparse-reward `LargeGridWorld`. Plots give means and 95% confidence intervals. Optimal average cost is 0.1. Training took place over 1e+6 timesteps; no further improvement occurred beyond timestep 1.2e+5.

In all cases, the vanilla methods prematurely converge to suboptimal policies, whereas the OIR-based methods solve the problem. This illustrates that, in sparse-reward environments, the inherent skepticism of OIR-based policy gradient methods can lead to improved performance.

Conclusion

In this paper we have addressed the exploration/exploitation trade-off in reinforcement learning via a new RL objective: the OIR. Interesting future directions include clarifying the relationship between optimal solutions to the OIR and vanilla problems, development of continuous-spaces version of IDAC, and thorough empirical evaluation of deep RL variants of IDAC on a range of benchmark problems.

References

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2020a. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, 64–66. PMLR.

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2020b. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, 64–66. PMLR. ISBN 2640-3498.

Åström, K. J.; and Eykhoff, P. 1971. System identification – a survey. *Automatica*, 7(2): 123–162.

Bedi, A. S.; Parayil, A.; Zhang, J.; Wang, M.; and Koppel, A. 2021. On the Sample Complexity and Metastability of Heavy-tailed Policy Search in Continuous Control. *arXiv preprint arXiv:2106.08414*.

Berkenkamp, F.; Turchetta, M.; Schoellig, A.; and Krause, A. 2017. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30.

Bertsekas, D. P.; and Shreve, S. E. 1996. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific.

Bhandari, J.; and Russo, D. 2019. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*.

Bhatnagar, S.; Sutton, R.; Ghavamzadeh, M.; and Lee, M. 2009. Natural actor-critic algorithms. *Automatica*, 45(11): 2471–2482.

Borkar, V. S. 2008. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press.

Cheng, R.; Orosz, G.; Murray, R. M.; and Burdick, J. W. 2019. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3387–3395.

Fazel, M.; Ge, R.; Kakade, S.; and Mesbahi, M. 2018. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, 1467–1476. PMLR.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 1861–1870. PMLR.

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

Konda, V. 2002. *Actor-Critic Algorithms*. Ph.D. thesis, MIT.

Kushner, H. J.; and Yin, G. G. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer-Verlag New York.

Lale, S.; Azizzadenesheli, K.; Hassibi, B.; and Anandkumar, A. 2022. Reinforcement learning with fast stabilization in linear dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, 5354–5390. PMLR.

Lee, L.; Eysenbach, B.; Parisotto, E.; Xing, E.; Levine, S.; and Salakhutdinov, R. 2019. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

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

Lu, X.; Van Roy, B.; Dwaracherla, V.; Ibrahimi, M.; Osband, I.; and Wen, Z. 2021. Reinforcement Learning, Bit by Bit. *arXiv preprint arXiv:2103.04047*.

Mei, J.; Xiao, C.; Szepesvári, C.; and Schuurmans, D. 2020. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, 6820–6829. PMLR.

Meyn, S. 2022. *Control Systems and Reinforcement Learning*. Cambridge University Press.

Mutti, M.; Pratissoli, L.; and Restelli, M. 2021. Task-agnostic exploration via policy gradient of a non-parametric state entropy estimate. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9028–9036.

Raffin, A.; Hill, A.; Ernestus, M.; Gleave, A.; Kanervisto, A.; and Dormann, N. 2019. Stable baselines3. *GitHub repository*.

Recht, B. 2019. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2: 253–279.

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Russo, D.; and Van Roy, B. 2014. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27: 1583–1591.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, A.; and Demchuk, E. 2003. Nearest neighbor estimates of entropy. *American journal of mathematical and management sciences*, 23(3-4): 301–321.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT Press.

Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 99: 1057–1063.

Tang, Y.; Zheng, Y.; and Li, N. 2021. Analysis of the optimization landscape of linear quadratic gaussian (lqg) control. In *Learning for Dynamics and Control*, 599–610. PMLR.

Uehara, M.; Sekhari, A.; Lee, J. D.; Kallus, N.; and Sun, W. 2022. Provably efficient reinforcement learning in partially observable dynamical systems. *arXiv preprint arXiv:2206.12020*.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3): 229–256.

Yarats, D.; Fergus, R.; Lazaric, A.; and Pinto, L. 2021. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, 11920–11931. PMLR.

Zhang, J.; Bedi, A. S.; Wang, M.; and Koppel, A. 2021. Beyond Cumulative Returns via Reinforcement Learning over State-Action Occupancy Measures. In *2021 American Control Conference*, 894–901.

Zhang, J.; Koppel, A.; Bedi, A. S.; Szepesvári, C.; and Wang, M. 2020a. Variational Policy Gradient Method for Reinforcement Learning with General Utilities. *Advances in Neural Information Processing Systems*, 33: 4572–4583.

Zhang, K.; Koppel, A.; Zhu, H.; and Başar, T. 2020b. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612.

APPENDIX

Experiments, Continued

Environments

Each gridworld is composed of an $n \times m$ grid of states, $\mathcal{S} = \{0, \ldots, n-1\} \times \{0, \ldots, m-1\}$, along with a designated start state s_{start}, designated goal state s_{goal}, and a set $B \subset \mathcal{S}$ of blocked states which the agent is not permitted to enter. Episodes are of fixed length K, and the agent begins each episode in state s_{start}. In a given state $s = (i, j)$, the agent chooses an action $a \in \{\text{stay}, \text{ up}, \text{ down}, \text{ left}, \text{ right}\}$. The agent then attempts to move in the direction corresponding to the action selected: if the selected action would move the agent off the grid or into a blocked state, the agent remains in s; otherwise, the agent moves into (or remains in) the state corresponding to the action selected. For example, if $a = \text{up}$ is chosen, the agent attempts to move to state $s' = (i, j-1)$. If s' is off the grid (i.e. $j - 1 < 0$) or $s' \in B$, the agent remains in s. Otherwise, the agent transitions to s'. Finally, let $\mathcal{A}(s)$ denote the set of all actions at s that do not lead off the grid or into a blocked state; the cost function is then given by:

$$c(s,a) = \begin{cases} c_{\text{goal}} & \text{if } s = s_{\text{goal}} \text{ and } a \in \mathcal{A}(s), \\ c_{\text{allowed}} & \text{if } s \neq s_{\text{goal}} \text{ and } a \in \mathcal{A}(s), \\ c_{\text{blocked}} & \text{if } a \notin \mathcal{A}(s), \end{cases}$$

where $0 < c_{\text{goal}} < c_{\text{allowed}} < c_{\text{blocked}}$. A policy minimizing $J(\theta)$ will move as quickly as possible to s_{goal} while always choosing actions within $\mathcal{A}(s)$. Because of this, when a problem is small enough that the agent can reach the goal state quickly and remain in it for most of the episode, the optimal average cost should be close to 1. A policy minimizing $\rho(\theta)$, on the other hand, will seek to balance minimizing $J(\theta)$ with maximizing $H(d_\theta)$, while avoiding actions $a \notin \mathcal{A}(s)$.

For the first set of experiments, we considered the `Gridworld1`, `2`, and `3` environments shown in Figure 2. The `LargeGridWorld` environment used in the second experiment is also depicted in Figure 2.

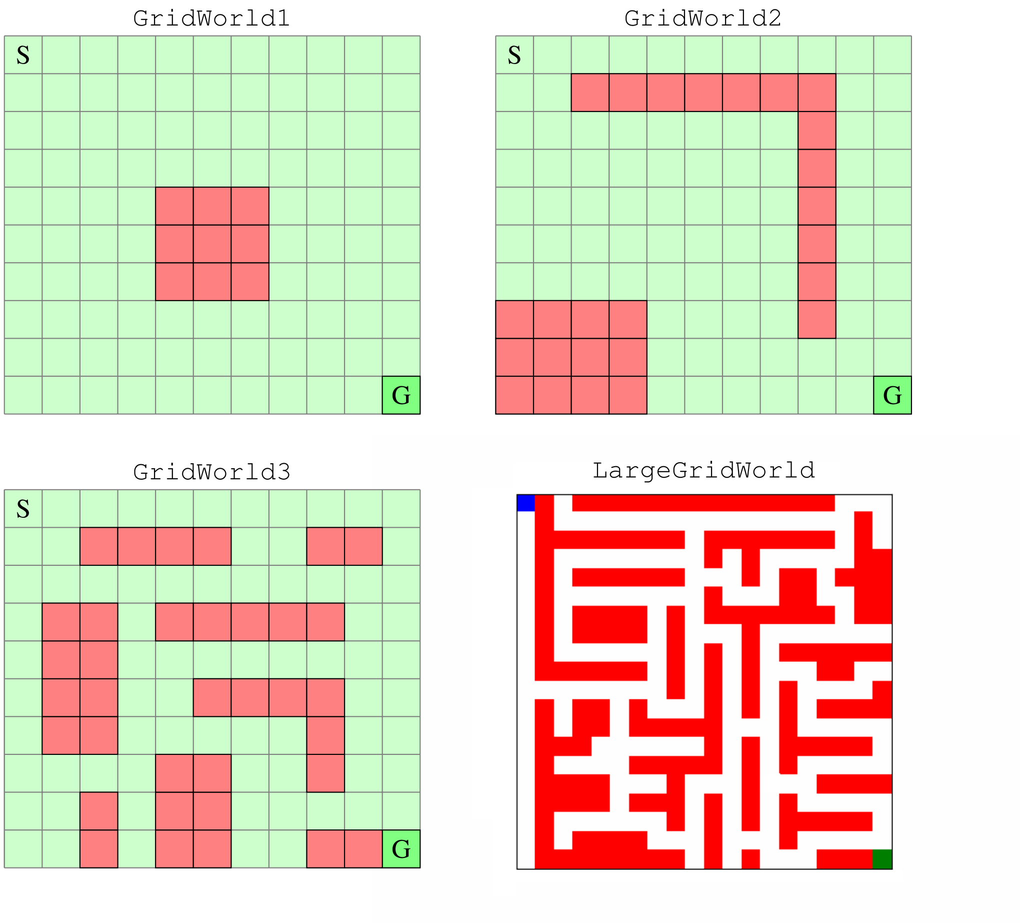

Figure 2: `GridWorld` environments. For `Gridworlds 1`, `2`, and `3`, the start state is S and goal state is G. Shaded regions represent blocked states B. For the `LargeGridWorld` environment, the blue square is the start state and the green square is the goal state.

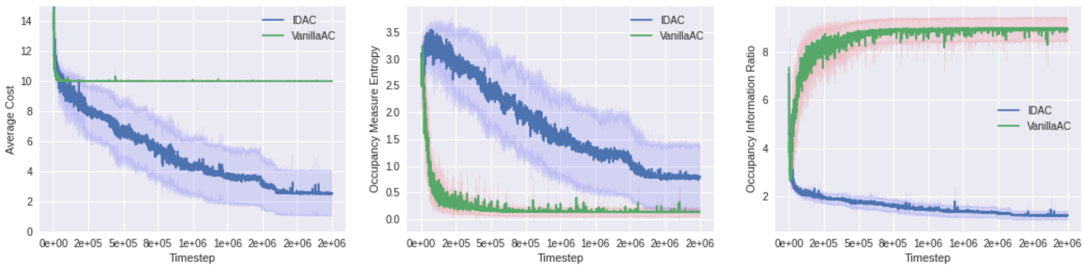

Figure 3: Comparison of tabular, softmax policy versions of vanilla actor-critic and IDAC with $\kappa = 1.0$ on `GridWorld2`. Optimal average cost is 1. Cost: both algorithms decrease to 10; vanilla AC gets stuck, IDAC's cost decreases well below 10. Entropy: vanilla AC's policy becomes deterministic, while IDAC maintains higher-entropy policies. OIR: IDAC minimizes the OIR, vanilla AC increases it.