Information-Directed Policy Search in Sparse-Reward Settings via the Occupancy Information Ratio*

Wesley A. Suttle
U.S. Army Research Laboratory
Adelphi, MD, USA
wesley.a.suttle.ctr@army.mil

Alec Koppel
J.P. Morgan AI Research
New York, NY, USA
alec.koppel@jpmchase.com

Ji Liu
Electrical and Computer Engineering
Stony Brook University
Stony Brook, NY, USA
ji.liu@stonybrook.edu

***Abstract*—We examine a new measure of the exploration/exploitation trade-off in reinforcement learning (RL) called the occupancy information ratio (OIR). We derive the Information-Directed Actor-Critic (IDAC) algorithm for solving the OIR problem, provide an overview of the rich theory underlying IDAC and related OIR policy gradient methods, and experimentally investigate the advantages of such methods. The central contribution of this paper is to provide empirical evidence that, due to the form of the OIR objective, IDAC enjoys superior performance over vanilla RL methods in sparse-reward environments.**

***Index Terms*—reinforcement learning, exploration vs. exploitation, sparse rewards**

I. Introduction

The field of reinforcement learning (RL) [1] has seen many attempts to address the exploration/exploitation trade-off by incentivizing exploration with various types of exploration bonus [2]–[5]. Few works have attempted to first directly quantify, then optimize the exploration/exploitation trade-off, however. Prior works in multi-armed bandits (MABs) and RL [6], [7] seek to balance the goals of exploration and exploitation 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. However, the same information-theoretic quantities and assumptions on problem structure that make their key insights possible also limit the practical utility of the information ratios proposed in [7] as tools for guiding action-selection. Due to these issues, the information ratio and its proxies explored in [7] suffer from tractability and scalability issues in realistic settings.

The occupancy information ratio (OIR) is a new RL objective that quantifies the exploration/exploitation trade-off using the ratio of the average cost of a policy to the entropy of its state occupancy measure. Occupancy measure entropy, or *occupancy information*, has recently been used as an optimization objective [8]–[10] that captures the informativeness of a policy by measuring the uniformity of its state space coverage. The occupancy information of a policy has the important advantage of being tractable for policy gradient methods, unlike the notions of information gain used in existing information ratios. It is well-known that policy gradient methods [11] are well-suited to problems with large, continuous spaces [3]–[5], and recent progress has also been made in providing global optimality guarantees for policy gradient methods [10], [12]–[15]. As we will show, the OIR is amenable to policy search in parameter space and enjoys a rich underlying theory with convergence guarantees, providing a promising proxy for the information ratios of [7].

Reward engineering is a challenging problem in RL with strong connections to the exploration/exploitation dilemma. Handcrafting dense rewards in complex environments can rapidly become intractable, so specifying tasks using sparse rewards is a common approach in many applications, such as robotics [16], [17]. Using sparse rewards introduces an exploration problem, however, as the agent is forced to systematically explore in order to uncover rewarding states that it can subsequently exploit [18], [19]. As recently demonstrated in the unsupervised learning community, intrinsic objectives like the occupancy information used in the OIR can be leveraged to explore effectively in the absence of rewards [9], [20], [21]. As we will experimentally illustrate, OIR-based methods provide a promising tool for intelligently balancing exploration with exploitation in sparse-reward settings.

In this work, we present the OIR and derive the Information-Directed Actor-Critic (IDAC) algorithm for optimizing it. Moreover, we provide results showing that this objective has no spurious extrema (Theorem 2), implying that IDAC exhibits convergence to global optimality as described in Theorems 3 and 4. Finally, we experimentally illustrate that the OIR yields policies that avoid spurious, suboptimal behavior in a variety of sparse-reward gridworld environments, whereas benchmarks exhibit a tendency to converge prematurely.

II. Problem Formulation

In this section we describe our problem setting and formulate the occupancy information ratio objective. We first define

*All proofs of the assertions in this paper are omitted due to space limitations, but are available upon request and will appear in a forthcoming publication. The research of W. A. Suttle was supported by the U.S. Army Research Laboratory (ARL) and was accomplished under Cooperative Agreements W911NF-21-2-0127 and W911NF-22-2-0003. The research of J. Liu was supported in part by ARL Cooperative Agreement W911NF-21-2-0098.

an underlying Markov decision process, then formulate the OIR as an objective to be optimized over it.

Markov Decision Processes. Consider a finite, average-cost MDP $(\mathcal{S}, \mathcal{A}, p, c)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p : \mathcal{S} \times \mathcal{A} \to \mathcal{D}(\mathcal{S})$ is the transition probability kernel mapping state-action pairs to distributions $\mathcal{D}(\mathcal{S})$ over $\mathcal{S}$, and $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ is the cost function. Let some parameterized family $\{\pi_\theta : \mathcal{S} \to \mathcal{D}(\mathcal{A})\}_{\theta \in \Theta}$ of policies be given, where $\Theta \subset \mathbb{R}^d$. For $\theta \in \Theta$, 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, assumed 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 $\mathcal{S} \times \mathcal{A}$. 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, define the state entropy by $H(d_\theta) = -\sum_s d_\theta(s) \log d_\theta(s)$. This quantity measures how well π_θ covers $\mathcal{S}$ in the long run.

Policy Gradients. Given an MDP $(\mathcal{S}, \mathcal{A}, p, c)$ and policy π_θ, define the relative state value function $V_\theta(s) = \sum_{t=0}^{\infty} \mathbb{E}_{\pi_\theta} [c(s, a) - J(\theta) \mid s_0 = s]$ and relative action value function $Q_\theta(s, a) = \sum_{t=0}^{\infty} \mathbb{E}_{\pi_\theta} [c(s, a) - J(\theta) \mid s_0 = s, a_0 = a]$. We are guaranteed by the policy gradient theorem [11] that, under certain conditions, $\nabla J(\theta) = \sum_s d_\theta(s) \sum_a Q_\theta(s, a) \nabla \pi_\theta(a|s) = \mathbb{E}_{\pi_\theta}\Big[(c(s, a) - J(\theta)) \nabla \log \pi_\theta(a|s)\Big]$.

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 used primarily to ensure the denominator remains positive. 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.

Remark 1. Though we stipulated that $\kappa > -\min_\theta H(d_\theta)$ in the definition of the OIR above, by letting $\kappa < -\max_\theta H(d_\theta)$, the expected cost $J(\theta)$ of the underlying MDP is instead treated as an expected *reward* to be maximized. Any algorithm for minimizing the OIR will therefore balance maximizing the reward $J(\theta)$ with maximizing the shifted entropy $|\kappa + H(d_\theta)|$, allowing the OIR framework to accommodate rewards by simply replacing the cost function c in the MDP with a reward function r, and choosing $\kappa < -\max_\theta H(d_\theta)$.

III. Algorithm

In this section we derive an actor-critic scheme for maximizing (1). We 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 1. We aim to perform policy gradient descent on (1), yet 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 [11] to estimate $\nabla J(\theta)$ and we can empirically estimate $J(\theta)$ and $H(d_\theta)$, it is not obvious how to estimate $\nabla H(d_\theta)$. Fortunately, the following extension of the classic policy gradient theorem allows us to tractably estimate $\nabla H(d_\theta)$.

Lemma 1. *Let a differentiable parameterized policy class $\{\pi_\theta\}_{\theta\in\Theta}$ be given and fix a policy parameter iterate θ_t at time-step t. We have $\nabla H(d_\theta)\big|_{\theta=\theta_t} = \mathbb{E}_{\pi_{\theta_t}}\Big[(-\log d_{\theta_t}(s) - H(d_{\theta_t}))\nabla \log \pi_{\theta_t}(a|s)\Big]$.*

This result implies the following OIR policy gradient theorem:

Theorem 1. *Let a differentiable parameterized policy class $\{\pi_\theta\}_{\theta\in\Theta}$ and a constant $\kappa \geq 0$ be given. We have $\nabla\rho(\theta_t) =$*

$$\mathbb{E}_{\pi_{\theta_t}}\left[\frac{\delta_t^J\big(\kappa + H(d_{\theta_t})\big) - J(\theta_t)\delta_t^H}{[\kappa + H(d_{\theta_t})]^2}\nabla \log \pi_{\theta_t}(a|s)\right], \tag{3}$$

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

Armed with Theorem 1, we now present an actor-critic algorithm for minimizing the OIR.

Information-Directed Actor-Critic. Information-Directed Actor-Critic (IDAC) is a variant of the classic actor-critic algorithm [22], [23] 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 $r_t(s, a) = -\log d_{\theta_t}(s)$ is a "shadow reward" associated with the gradient in Lemma 1. For ease of exposition, we assume access to an oracle DensityEstimator that returns the occupancy measure $d_\theta = \text{DensityEstimator}(\theta)$, given $\theta \in \Theta$. See Algorithm 1 for IDAC pseudocode.

IV. Theoretical Results

In this section we provide key results underpinning policy search for the OIR problem Taken together, these results demonstrate that our IDAC algorithm achieves global optimality under suitable conditions.

A. Stationarity Implies Global Optimality

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

$$\min_{\theta \in \Theta} \quad \rho(\theta) = \frac{J(\theta)}{\kappa + H(d_\theta)}. \tag{4}$$

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 $\theta \in \Theta$ 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 1. *The set Θ is compact. For any $s \in \mathcal{S}, a \in \mathcal{A}$, $\pi_\theta(a|s)$ is continuously differentiable on Θ, and the Markov chain induced by π_θ on $\mathcal{S}$ is ergodic.*

Algorithm 1 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 $s_0, \theta_0, \omega_0^J, \omega_0^H$, select μ_{-1}^H, $\mu_{-1}^J > 0$, and set $t \leftarrow 0$.
repeat
  Generate trajectory $\{(s_i, a_i)\}_{i=1,\ldots,K}$ using π_{θ_t}
  $\mu_t^J = (1-\tau)\mu_{t-1}^J + \tau \frac{1}{K}\sum_{i=1}^K c(s_i, a_i)$
  $d_{\theta_t} = \text{DENSITYESTIMATOR}(\theta_t)$
  $\mu_t^H = (1-\tau)\mu_{t-1}^H + \tau \frac{1}{K}\sum_{i=1}^K (-\log d_{\theta_t}(s_i))$
  for $i = 1, \ldots, K$ **do**
    Set $v_{\omega_t^J}(s_{K+1}) = v_{\omega_t^H}(s_{K+1}) = 0$
    $\delta_i^J = c(s_i, a_i) - \mu_t^J + v_{\omega_t^J}(s_{i+1}) - v_{\omega_t^J}(s_i)$
    $\delta_i^H = -\log d_{\theta_t}(s_i) - \mu_t^H + v_{\omega_t^H}(s_{i+1}) - v_{\omega_t^H}(s_i)$
    $\psi_i = \nabla \log \pi_{\theta_t}(a_i|s_i)$
  end for
  $\omega_{t+1}^J = \omega_t^J + \alpha \frac{1}{K}\sum_{i=1}^K \delta_i^J \nabla v_{\omega_t^J}(s_i)$
  $\omega_{t+1}^H = \omega_t^H + \alpha \frac{1}{K}\sum_{i=1}^K \delta_i^H \nabla v_{\omega_t^H}(s_i)$
  $\widehat{\nabla\rho(\theta_t)} = \frac{1}{[\kappa+\mu_t^H]^2}\frac{1}{K}\sum_{i=1}^K \left[\delta_i^J\left(\kappa + \mu_t^H\right) - \mu_t^J \delta_i^H\right]\psi_i$
  $\theta_{t+1} = \theta_t - \beta\widehat{\nabla\rho(\theta_t)}$
  $t \leftarrow t+1$
until convergence

Assumption 2. *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 2. *Let Assumptions 1 and 2 hold. Let θ^* be a stationary point of* (4)*, i.e., $\nabla\rho(\theta^*) = 0$. Then θ^* is globally optimal for* (4)*.*

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 [10], [12]–[15] using persistent exploration conditions, our result instead builds on hidden concavity arguments from [10], which apply to parameterized policies. However, Theorem 2 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 2 is thus a strict generalization of existing results for the landscape of RL objectives.

B. Non-Asymptotic Convergence without Approximation Error

Next, we establish a non-asymptotic convergence rate for the following projected gradient descent scheme for solving the OIR minimization problem (4):

$$\begin{aligned}\theta_{t+1} &= \text{Proj}_\Theta\left(\theta_t - \eta\nabla\rho(\theta_t)\right) \qquad (5)\\ &= \arg\min_\theta[\rho(\theta_t) + \langle\nabla\rho(\theta_t), \theta - \theta_t\rangle + \frac{1}{2\eta}\|\theta - \theta_t\|^2],\end{aligned}$$

for a fixed stepsize $\eta > 0$, where Proj_Θ denotes euclidean projection onto Θ and the second equality holds by the convexity of Θ. The updates (5) can be viewed as an idealized version of the gradient descent scheme underlying IDAC. We assume the projection operation, which is typically not needed in practice, for the purposes of analysis.

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. Under the ergodicity conditions in Assumption 1 and properties of entropy, $\min_\theta \rho(\theta) > 0$ and $\max_\theta \rho(\theta) < \infty$. In addition to Assumptions 1 and 2, we will need the following.

Assumption 3. *$\nabla\rho(\theta)$ is Lipschitz and $L > 0$ is the smallest number such that $\|\nabla\rho(\theta) - \nabla\rho(\theta')\| \leq L\|\theta - \theta'\|$, for all $\theta, \theta' \in \Theta$.*

We have the following convergence rate result for (5).

Theorem 3. *Let Assumptions 1, 2, and 3 hold. Let $D_\zeta = \max_{z,z'\in(\zeta\circ\lambda)(\Theta)}\|z - z'\|$ 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 $t \geq 0$,*

$$\rho(\theta_t) - \rho(\theta^*) \leq \frac{4M^2L_1\ell^2D_\zeta^2}{t+1}. \qquad (6)$$

Coupled with Theorem 2, this result provides a non-asymptotic convergence rate to *global optimality* for algorithms solving the OIR minimization problem (4).

C. Asymptotic Convergence with Approximation Error

We conclude this section by proving almost sure (a.s.) convergence of IDAC to a neighborhood of a stationary point of (4). By Theorem 2, this implies IDAC converges a.s. to a neighborhood of a *global* optimum. This result is stronger than typical asymptotic results for actor-critic schemes, which usually guarantee convergence to a neighborhood of a local optimum or saddle point [13], [23], [24].

We analyze the algorithm as given in Algorithm 1 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:

$$\theta_{t+1} = \Gamma\Big[\theta_t - \beta_t\frac{\delta_t^J(\kappa+\mu_t^H) - \mu_t^J\delta_t^H}{\left(\kappa+\mu_t^H\right)^2}\nabla\log\pi_{\theta_t}(a_t|s_t)\Big], \quad (7)$$

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. In addition to Assumption 1, we impose the following:

Assumption 4. *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 5. *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.*

We now present the main result of this subsection, which establishes convergence of the actor-critic algorithm. Consider the ordinary differential equation (ODE)

$$\dot{\theta} = \hat{\Gamma}(\nabla\rho(\theta)), \tag{8}$$

where $\hat{\Gamma}(\nabla\rho(\theta)) := \lim_{\eta\to 0^+} [\gamma(\theta + \eta\nabla\rho(\theta)) - \theta]/\eta$. We note here that (8) 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.

Theorem 4. *Let* $\mathcal{Z}$ *denote the set of asymptotically stable equilibria of the ODE* (8). *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* $\varepsilon_\theta = (\epsilon_\theta^J [\kappa + H(d_\theta)] - J(\theta)\epsilon_\theta^H)/([\kappa + H(d_\theta)]^2)$. *Under Assumptions 1, 4, and 5, given any* $\varepsilon > 0$, *there exists* $\delta > 0$ *such that, for* $\{\theta_t\}$ *obtained from Algorithm 1 with projection* (7), *if* $\sup_t \|\epsilon_{\theta_t}\| < \delta$, *then* $\theta_t \to \mathcal{Z}^\varepsilon$ *a.s. as* $t \to \infty$.

Combined with Theorem 2, Theorem 4 establishes almost sure convergence of IDAC to a neighborhood of a *global* optimum of the OIR minimization problem (4). Note that if the linear approximation and features are expressive enough, then ε will be small or even zero.

V. Experiments

The experiments presented in this section demonstrate that, when the reward signal is sparse, OIR methods can lead to improved performance when compared with vanilla RL methods. To illustrate this, we present two different sets of experiments on gridworld environments of varying complexity. In the first set of experiments, we compared tabular implementations of IDAC and vanilla actor-critic (AC) on three relatively small gridworlds. For the second set of experiments we compared a neural network version of IDAC with the Stable Baselines implementations[1] of A2C, DQN, and PPO on a larger, more complex gridworld. In all cases, the vanilla methods prematurely converge to suboptimal policies, whereas IDAC solves the problem.

A. Setup

Environments. As pictured in Figure 2, each gridworld has designated start and goal states s_{start} and s_{goal}, and a set of blocked states the agent cannot enter. Episodes are of fixed length K. In a given state s, the agent chooses an action $a \in \mathcal{A}(s) \subset \{\text{stay}, \text{move up}, \text{move down}, \text{move left}, \text{move right}\}$,

[1]https://stable-baselines3.readthedocs.io/en/master

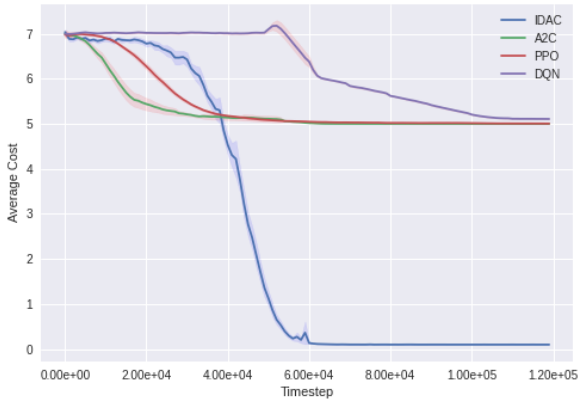

Fig. 1: Comparison of neural network IDAC with common deep RL methods on the sparse-reward `LargeGridWorld`. Plots give means and 95% confidence intervals. Optimal average cost is 0.1. Training took place over 1e+6 timesteps; no further improvement occurred beyond timestep 1.2e+5.

where $\mathcal{A}(s)$ is the set of actions not running into a blocked state or off the grid when executed in s. Transitions are deterministic. The cost function is 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 $0 < c_{\text{goal}} < c_{\text{allowed}} < c_{\text{blocked}}$.

Implementation. For the first set of experiments, we implemented a tabular version of Algorithm 1. In order to have a baseline to compare against, we also implemented vanilla average-cost AC. For both algorithms, we used tabular softmax policies: $\pi_\theta(a_i|s) = \exp(\theta^T\psi(s,a_i))/(\sum_j \exp(\theta^T\psi(s,a_j)))$, where $\theta \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ and $\psi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ maps each state-action pair to a unique standard basis vector $e_k \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$, where e_k has a 1 in its kth entry and 0 everywhere else. We similarly used tabular representation for the value functions: $v_\omega(s) = \omega^T\phi(s)$, where $\omega \in \mathbb{R}^{|\mathcal{S}|}$ and $\phi : \mathcal{S} \to \mathbb{R}^{|\mathcal{S}|}$ maps each state s_i to a unique standard basis vector e_i.

For the second set of experiments, we implemented IDAC with a categorical policy using two-layer, fully connected neural networks for both the policy and value functions, and we compared against the Stable Baselines implementations of A2C, DQN, and PPO with two-layer, fully connected neural networks for all policies and value function approximators.

B. Tabular Experiment Results

Figures 3, 4, and 5 compare IDAC and vanilla AC on the `GridWorld` environments with $c_{\text{goal}} = 1, c_{\text{allowed}} = 10$, and $c_{\text{blocked}} = 100$. To generate these figures, 15 instances of each algorithm were run on the environment, the average cost and entropy were computed for each episode, and the sample means and 95% confidence intervals for the cost, entropy, and corresponding OIR over the 15 runs were used to generate

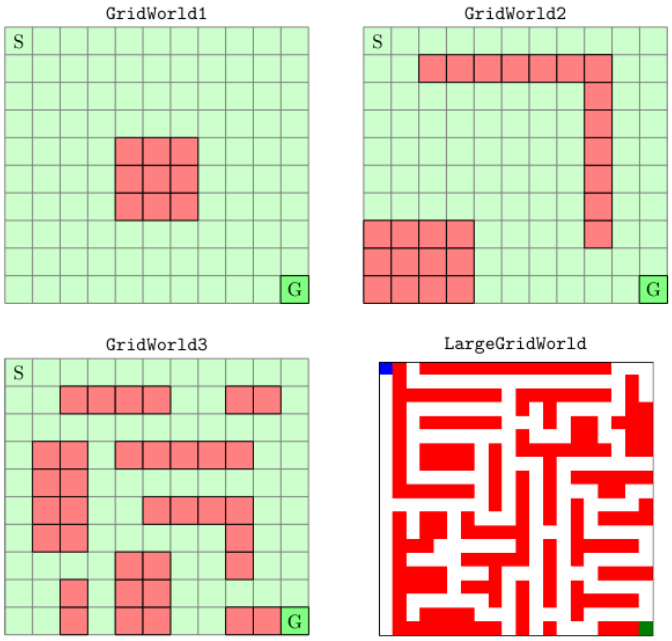

Fig. 2: `GridWorld` environments. For `Gridworlds 1`, `2`, and `3`, the start state is S and goal state is G. Shaded regions represent blocked states. For `LargeGridWorld`, the blue square is the start state and the green square is the goal state.

the learning curves. Hyperparameters [cf., Algorithm 1] were chosen through trial and error. As the figures show, the OIR algorithm outperforms the vanilla algorithm in every case.

In all three figures, the average costs indicate that both algorithms quickly learn to avoid actions moving off the grid or into blocked states, decreasing to a value around 10. In all cases, vanilla AC gets stuck near 10 for the remainder of training. This corresponds to taking allowed actions, but not attaining the goal state. The IDAC algorithm, on the other hand, clearly spends an increasing amount of time in the goal state, since its cost decreases well below 10. Vanilla actor-critic converges fairly quickly to a policy visiting only a small subset of the available states, indicating that vanilla AC's policies become deterministic before the state space has been sufficiently explored. IDAC, in contrast, maintains policies with higher state occupancy measure entropy early on, only decreasing as the algorithm discovers the goal state, then seeks to strike the right balance between cost and entropy. Finally, in all cases IDAC naturally minimizes the OIR, while vanilla AC consistently increases it.

C. Neural Network Experiment Results

Figure 1 compares the performance of neural IDAC and A2C, DQN, and PPO on `LargeGridWorld` with $c_{\text{goal}} = 0.1, c_{\text{allowed}} = 5$, and $c_{\text{blocked}} = 10$. To generate the data for these figures, we first trained 48 instances of neural IDAC with different random seeds. We next trained 15 instances of each of the A2C, DQN, and PPO algorithms on the environment. For each algorithm, the average cost was computed for each episode, and the sample means and 95% confidence intervals were used to create the learning curves. We note that, out of the 48 IDAC trials, 40 succeeded in finding the goal state, while 8 failed, an 83% success rate. To create Figure 1, we randomly selected 15 of the 40 successful runs of IDAC to compare with A2C, DQN, and PPO. As the figure illustrates, IDAC outperformed all three. Furthermore, none of A2C, DQN, and PPO found the goal state after 1×10^6 timesteps. Again, hyperparameters were chosen through trial and error.

As in the tabular experiments, all algorithms quickly learn to avoid blocked actions. In the case of A2C and PPO, this leads to an average cost of exactly 5, while for DQN the cost remains slightly above 5 due to exploration noise lower bounded by 0.05. Though the optimal cost is 0.1, once they have converged to these values, they remain there for the remainder of training, reflecting premature convergence before sufficient exploration has been achieved. Meanwhile, since neural IDAC is minimizing $\rho(\theta)$ instead of $J(\theta)$, it swiftly locates the goal state and finds an optimal policy with average cost 0.1. This illustrates that, in sparse-reward environments, OIR-based policy gradient methods can lead to improved performance over vanilla techniques.

VI. Conclusion

In this paper we have addressed the exploration/exploitation trade-off in reinforcement learning with sparse rewards 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

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

[2] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

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

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

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

[6] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1583–1591, 2014.

[7] X. Lu, B. Van Roy, V. Dwaracherla, M. Ibrahimi, I. Osband, and Z. Wen, "Reinforcement learning, bit by bit," *arXiv preprint arXiv:2103.04047*, 2021.

[8] E. Hazan, S. Kakade, K. Singh, and A. Van Soest, "Provably efficient maximum entropy exploration," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2681–2691.

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

[10] J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvári, and M. Wang, "Variational policy gradient method for reinforcement learning with general utilities," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4572–4583, 2020.

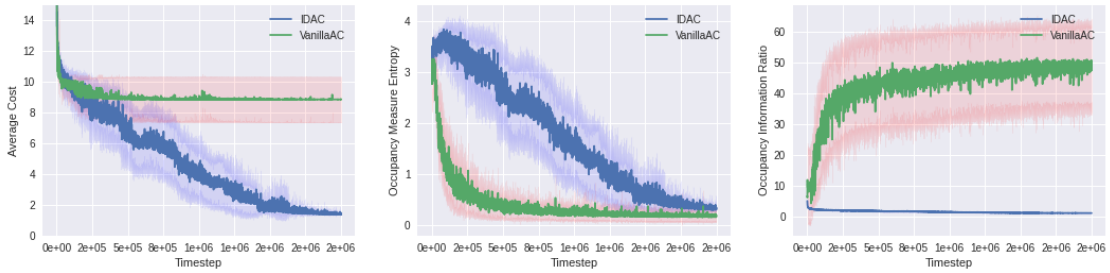

Fig. 3: Comparison of IDAC with $\kappa = 1.0$ and vanilla actor-critic on `GridWorld1`.

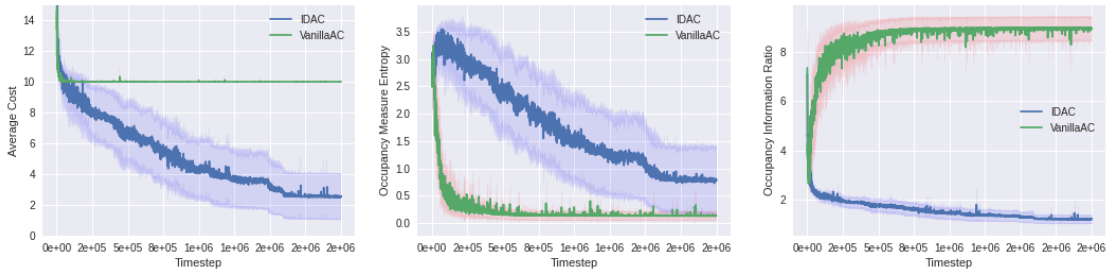

Fig. 4: Comparison of IDAC with $\kappa = 1.0$ and vanilla actor-critic on `GridWorld2`.

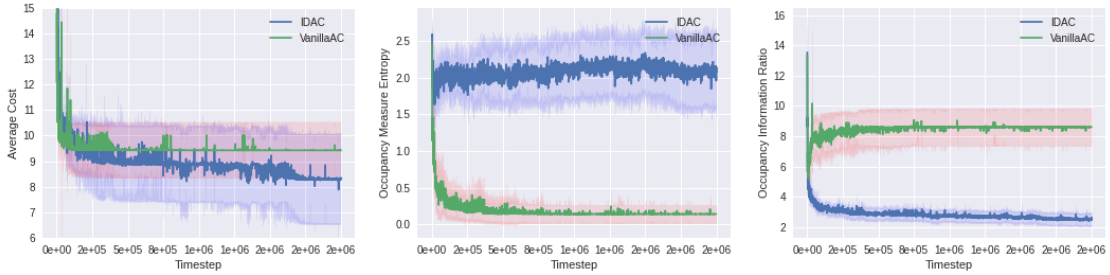

Fig. 5: Comparison of IDAC with $\kappa = 1.0$ and vanilla actor-critic on `GridWorld3`.

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

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

[13] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "Optimality and approximation with policy gradient methods in Markov decision processes," in *Conference on Learning Theory*. PMLR, 2020, pp. 64–66.

[14] J. Mei, C. Xiao, C. Szepesvári, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6820–6829.

[15] A. S. Bedi, A. Parayil, J. Zhang, M. Wang, and A. Koppel, "On the sample complexity and metastability of heavy-tailed policy search in continuous control," *arXiv preprint arXiv:2106.08414*, 2021.

[16] M. Vecerik, T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. Heess, T. Rothörl, T. Lampe, and M. Riedmiller, "Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards," *arXiv preprint arXiv:1707.08817*, 2017.

[17] A. Singh, L. Yang, K. Hartikainen, C. Finn, and S. Levine, "End-to-end robotic reinforcement learning without reward engineering," *arXiv preprint arXiv:1904.07854*, 2019.

[18] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 6292–6299.

[19] M. Riedmiller, R. Hafner, T. Lampe, M. Neunert, J. Degrave, T. Wiele, V. Mnih, N. Heess, and J. T. Springenberg, "Learning by playing solving sparse reward tasks from scratch," in *International conference on machine learning*. PMLR, 2018, pp. 4344–4353.

[20] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, "Reinforcement learning with prototypical representations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 920–11 931.

[21] H. Liu and P. Abbeel, "Behavior from the void: Unsupervised active pre-training," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 459–18 473, 2021.

[22] V. Konda, "Actor-critic algorithms," Ph.D. dissertation, MIT, 2002.

[23] S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee, "Natural actor-critic algorithms," *Automatica*, vol. 45, no. 11, pp. 2471–2482, 2009.

[24] K. Zhang, A. Koppel, H. Zhu, and T. Başar, "Global convergence of policy gradient methods to (almost) locally optimal policies," *SIAM Journal on Control and Optimization*, vol. 58, no. 6, pp. 3586–3612, 2020.