SIAM J. CONTROL OPTIM.
Vol. 62, No. 6, pp. 3145–3171

OCCUPANCY INFORMATION RATIO: INFINITE-HORIZON, INFORMATION-DIRECTED, PARAMETERIZED POLICY SEARCH*

WESLEY A. SUTTLE[†], ALEC KOPPEL[‡], AND JI LIU[§]

Abstract. In this work, we propose an information-directed objective for infinite-horizon reinforcement learning (RL), called the occupancy information ratio (OIR), inspired by the information ratio objectives used in previous information-directed sampling schemes for multi-armed bandits and Markov decision processes as well as recent advances in general utility RL. The OIR, composed of a ratio between the average cost of a policy and the entropy of its induced state occupancy measure, enjoys rich underlying structure and presents an objective to which scalable, model-free policy search methods naturally apply. Specifically, we show by leveraging connections between quasiconvex optimization and the linear programming theory for Markov decision processes that the OIR problem can be transformed and solved via convex optimization methods when the underlying model is known. Since model knowledge is typically lacking in practice, we lay the foundations for model-free OIR policy search methods by establishing a corresponding policy gradient theorem. Building on this result, we subsequently derive REINFORCE- and actor-critic-style algorithms for solving the OIR problem in policy parameter space. Crucially, exploiting the powerful *hidden quasiconvexity* property implied by our transformation of the OIR problem, we establish finite-time convergence of the REINFORCE-style scheme to global optimality and asymptotic convergence of the actor-critic-style scheme to (near) global optimality under suitable conditions. Finally, we experimentally illustrate the utility of OIR-based methods over vanilla methods in sparse-reward settings, supporting the OIR as an alternative to existing RL objectives.

Key words. reinforcement learning, policy gradient methods, nonconvex optimization

MSC codes. 93E03, 93E20, 93E35, 90-08, 65K05

DOI. 10.1137/22M1536650

1. Introduction.

The field of reinforcement learning (RL) [35] has seen many attempts to address the exploration/exploitation trade-off by incentivizing exploration via additive regularization; the hope is that, with more experience, the agent can improve its exploitation capabilities. Prior works on *information-directed* solution methods for multi-armed bandits (MABs) [29, 30] and Markov decision processes (MDPs) [22] instead seek to address this trade-off by minimizing an *information ratio* objective, defined as the ratio of cost incurred to information acquired. Importantly, when used as a tool for devising information-directed action-selection schemes, the specific form of these information ratio objectives leads to policies with improved data efficiency and improved regret bounds revealing the dependence of performance on information. Beyond the original works, the advantages of information ratio objectives have been

*Received by the editors November 23, 2022; accepted for publication (in revised form) July 30, 2024; published electronically November 21, 2024. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.
https://doi.org/10.1137/22M1536650
Funding: The research of the first author 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 the third author was supported in part by the National Science Foundation (NSF) under grant 2230101 and by the Air Force Office of Scientific Research (AFOSR) under award FA9550-23-1-0175. The second author contributed to this work in his previous role with ARL.
[†]U.S. Army Research Laboratory, Silver Spring, MD 20902 USA (wesley.a.suttle.ctr@army.mil).
[‡]J.P. Morgan AI Research, JP Morgan Chase, New York, NY 10017 USA (alec.koppel@jpmchase.com).
[§]Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA (ji.liu@stonybrook.edu).

analyzed in the frequentist bandit setting [15], as well as the more general linear partial monitoring setting [14]. In the RL setting, however, the same information-theoretic quantities and assumptions on problem structure that make these insights possible also limit the practical utility of the information ratios proposed in [22] as tools for guiding action-selection. In particular, the abstract learning targets, representation of cost in terms of regret, and mutual information formulation of information gain of a policy lead to difficulties in devising practical estimation procedures. Moreover, the practical schemes proposed in [22] rely on optimizing over the space $\mathcal{D}(\mathcal{A})$ of action distributions at each step, limiting their practical use to the finite action space setting. Due to these issues, the information ratio and its proxies explored in [22] suffer from tractability and scalability issues in realistic settings.

Gaps therefore remain in the theory of information-directed methods under general function approximation. The work [26] proposes a variant of deep Q-learning that optimizes the ratio of Bellman error to a variance surrogate for information gain, with substantial performance gains in practice, suggesting developing the theory of information-directed schemes that can operate with parameterization is a worthy avenue of pursuit. New proxy objectives that tractably, scalably extend the spirit of the information-directed schemes of [29, 30, 22] to operate with function approximation and exhibit performance guarantees are therefore required. In order to achieve this, two issues must be addressed. First, in order to overcome the limited scalability inherent in value-based methods, operating in parameter space is required, for which policy gradient methods are most natural [20, 32, 9]. Recent theoretical progress has also been made in providing global optimality guarantees for policy gradient methods [4, 1, 23, 39, 3], strengthening the motivation for pursuing such methods. Second, to address the estimation issues associated with the notions of information gain used in [22], we need a definition of informativeness that is amenable to policy search in parameter space. Occupancy measure entropy has recently been used as an optimization objective [10, 19, 39] quantifying the amount of information about the environment that a policy provides through the Kullback–Leibler divergence of its state occupancy measure from a uniform distribution. Motivated by this, in this work we take occupancy measure entropy, or occupancy information, of a policy as the fundamental quantity defining its informativeness. Based on this definition, we develop and study a new RL 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.

Main contributions. Our main contributions are as follows. (1) We propose a new RL objective, the occupancy information ratio (OIR), that is both inspired by the information ratio objectives of [29, 30, 22] and amenable to solution via policy search. (2) Drawing on connections between quasiconvex optimization and the linear programming theory for MDPs, we derive a concave programming reformulation of the OIR optimization problem over the space of state-action occupancy measures, establishing underlying theory that we exploit to strengthen our subsequent convergence results. (3) We derive an OIR policy gradient theorem, then use it to develop OIR policy gradient algorithms: Information-Directed REINFORCE (ID-REINFORCE) and Information-Directed Actor-Critic (IDAC). (4) We establish corresponding convergence theory with three key results: (i) OIR policy optimization enjoys a powerful *hidden quasiconvexity* property guaranteeing its first-order stationary points are global optima; (ii) the gradient descent scheme underlying ID-REINFORCE enjoys a nonasymptotic, information-dependent convergence rate; (iii) IDAC converges with probability one to (a neighborhood of) a *global* optimum of the OIR problem. (5) We provide experimental results indicating that OIR-based methods are able to outperform vanilla RL methods in sparse-reward settings, providing auxiliary support for the study of the OIR as an independent RL objective.

It is important to note that while the technical motivation for the OIR objective stems from balancing explore-exploit issues via connections with the information ratio methods of [29, 30, 22], our main convergence theory is of an optimization flavor, in the sense that we provide asymptotic and nonasymptotic analysis of algorithms optimizing the OIR objective. An information-theoretic characterization of the resulting policies remains an important, open problem that we leave for future work. The key technical challenge in our results lies in handling the fractional form of the OIR objective, which has not been previously addressed in the literature. To overcome this challenge, we first characterize the *quasiconvex* structure of the OIR problem in section 3. Leveraging this structure, especially properties of the perspective transform familiar to the quasiconvex programming literature, we then extend the convex utility analysis of [39][1] to *quasiconvex* utilities, including the OIR, in sections 5.1–5.2. Finally, we extend the asymptotic actor-critic analyses of [5, 34] to our IDAC algorithm, taking special care to establish the requisite smoothness properties of the OIR gradient as well as asymptotic negligibility of corresponding, OIR-specific noise and error terms.

2. Problem formulation. We now describe our problem setting and formulate the OIR objective. We first define an underlying Markov decision process, then formulate the OIR as an objective to be optimized over it.

2.1. 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.

2.2. Occupancy information ratio. We consider the OIR objective

$$\rho(\theta) = \frac{J(\theta)}{\kappa + H(d_\theta)}, \tag{2.1}$$

where $\kappa > -\min_\theta H(d_\theta)$ is a user-specified constant chosen to ensure that the denominator in (2.1) remains strictly positive. Given an MDP $(\mathcal{S}, \mathcal{A}, p, c)$, our goal is to find

[1]While the term "hidden concavity" is used in [39] due to the authors' focus on maximization, we focus on minimization and will thus use the term "hidden convexity" when there is no risk of confusion.

a policy parameter θ^* such that π_{θ^*} minimizes (2.1) *over the MDP,* i.e., subject to its costs and dynamics. As $J(\theta)$ and $H(d_\theta)$ are both infinite-horizon quantities, we regard (2.1) as an *infinite-horizon objective.*

Remark 2.1. If we allow $\kappa < -\max_\theta H(d_\theta)$ and require $J(\theta) \geq 0$ for all $\theta \in \Theta$, minimizing (2.1) will in fact minimize the ratio of $-J(\theta)$ to the absolute value $|\kappa + H(d_\theta)|$. In this case, minimizing the OIR maximizes $J(\theta)/(|\kappa| - H(d_\theta))$. 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 $\kappa < -\max_\theta H(d_\theta)$.

2.3. OIR as a proxy objective for information-directed sampling. The general setting of [22] is a sequential decision-making problem where the goal is to balance optimizing a given objective with acquiring information about an abstract *learning target,* $\mathcal{X}$, through interactions with the environment, all while maintaining and updating some relevant *epistemic state,* $\mathcal{P}_t$. For example, $\mathcal{X}$ may denote the optimal policy for the objective or some suitable exploration scheme, while $\mathcal{P}_t$ might include policy and value function parameters at time t. Given some reward r, state s, and policy π, let $V_\pi(s)$ denote the value function starting from state s of policy π, and let $Q_\pi(s,a)$ denote the state-action value function for π starting from s,a. Define $V_*(s) = \max_\pi V_\pi(s), Q_*(s,a) = \max_\pi Q_\pi(s,a)$, and let $H(\mathcal{X}|\mathcal{P}_t)$ denote the conditional entropy, or remaining uncertainty, of the learning target given $\mathcal{P}_t$. Once the agent has successfully achieved its learning target, $H(\mathcal{X}|\mathcal{P}_t)$ will typically be small or zero. Given $\mathcal{P}_t$, horizon τ, and a candidate policy π, let $\mathcal{P}_{t+\tau}$ denote the epistemic state resulting from starting with $\mathcal{P}_t$ and using π for τ steps. Then $[H(\mathcal{X}|P_t) - H(\mathcal{X}|P_{t+\tau})]/\tau$ is the τ-step *information gain* resulting from following π. For a given candidate policy π, the τ-step information ratio of π at time t is defined in [22] as the ratio of its instantaneous squared shortfall to its τ-step information gain:

$$\Gamma^\pi_{\tau,t} = \frac{\mathbb{E}_\pi\left[V_*(s_t) - Q_*(s_t,a_t)\right]^2}{\left[H(\mathcal{X}|P_t) - H(\mathcal{X}|P_{t+\tau})\right]/\tau}. \tag{2.2}$$

For a candidate π, [30, 22] show that $\text{Regret}(T|\pi) = \sum_{t=0}^{T-1} \mathbb{E}_\pi\left[V_*(s_t) - Q_*(s_t,a_t)\right] \leq \sqrt{H(\mathcal{X}|P_0)\sum_{t=0}^{T-1}\Gamma^\pi_{\tau,t}}$. This bound suggests that, by choosing a policy minimizing (2.2) at each time step, overall regret can be minimized, leading to improved data efficiency due to intelligent information acquisition. However, several factors limit the tractability of the information ratio objective (2.2). First, the presence of V_*, Q_* renders explicit estimation of the numerator intractable. Similarly, the specific choice of $\mathcal{X}$, formulation of P_t, and choice of τ make estimation of the denominator difficult. Objective (2.2) is thus more useful as an archetype for proxy objectives than as an optimization objective itself. Several Q-learning–based schemes using such proxy objectives are accordingly proposed in [22], yet these are inherently restricted to the finite action space setting and the corresponding proxy objectives are not amenable to optimization using policy gradient-based methods, limiting scalability.

To obtain a proxy objective for (2.2) that is amenable to policy search, we must recast the components of (2.2) into policy search-friendly terms. We emphasize that, due to the abstract nature of the learning target $\mathcal{X}$ and epistemic states P_t and $P_{t+\tau}$ in (2.2), deriving a direct translation of this objective for the policy search setting and establishing analogues of the corresponding regret analysis of [22] are likely impractical. We instead focus on deriving a new proxy objective that is above all practically and theoretically tractable for parameterized policy search, yet still retains the spirit of (2.2) as a measure of the ratio of shortfall incurred to information gained. To achieve this, we first replace the squared shortfall in the numerator of (2.2) with the expected average cost, $J(\theta)$, of the candidate policy π_θ. By eliminating the dependence on the optimal value functions V_* and Q_* present in (2.2) and removing the square in the squared shortfall, we break the applicability of the regret analyses of [22]. In compensation, however, these steps enable our subsequent use of the policy gradient theorem, a key ingredient for policy search. We next recast the information gain in the denominator. Without prior knowledge of the optimal policy π^*, environmental exploration is a natural proxy for information gain. As discussed in the introduction, state occupancy measure entropy, $H(d_\theta)$, is widely used to quantify the exploration achieved by policy π_θ. For this reason, we replace the abstract τ-step information gain of (2.2) with the concrete state occupancy measure entropy term $\kappa + H(d_\theta)$, where we choose $\kappa > -\min_\theta H(d_\theta)$ to ensure that the denominator remains positive. Despite losing the generality of the information gain term of (2.2) and the applicability of the regret analysis of [22], the resulting OIR objective (2.1) is far more tractable for policy search, as will be seen next.

3. Elements of OIR optimization. We now turn to the problem of optimizing the OIR defined in (2.1). First, we build on parallels with linear programming solutions to MDPs and quasiconvex programming to transform the nonconvex problem of minimizing (2.1) into a concave program over the space of state-action occupancy measures. This endows the OIR problem with the powerful *hidden quasiconvexity* property (cf. section 5.1) that we exploit to strengthen the convergence results for our policy gradient algorithms in section 5. Second, we lay the groundwork for model-free policy search methods developed in section 4 by deriving a policy gradient theorem for $\nabla\rho(\theta)$.

3.1. Concave reformulation. Given an average-cost MDP $(\mathcal{S}, \mathcal{A}, p, c)$ and a policy π, let $\lambda_\pi \in \mathcal{D}(\mathcal{S} \times \mathcal{A})$ denote the state-action occupancy measure induced by π on $\mathcal{S} \times \mathcal{A}$, i.e., $\lambda_{sa} = \lim_{t\to\infty} P(s_t = s, a_t = a \mid \pi)$. As discussed in section 8.8 of [27], if we have access to p and c, an optimal state value function can be obtained by solving a related linear program, (P). This is useful, as the existence of weakly polynomial-time algorithms for solving linear programs [12, 11] ensures the problem can be solved efficiently. Furthermore, the state-action occupancy measure λ^* of the optimal policy π^* for $(\mathcal{S}, \mathcal{A}, p, c)$ can be obtained by solving the following linear program, which is dual to (P): $\min_{\lambda\geq 0}\{J(\lambda) = c^T\lambda \mid \sum_{s,a}\lambda_{sa} = 1$ and $\sum_a \lambda_{sa} = \sum_{s',a} p(s|s',a)\lambda_{s'a}\ \forall s \in \mathcal{S}\}$. Call this dual linear program (D). The constraints ensure that the decision variables λ give a valid state-action occupancy measure for the MDP. Given a feasible solution λ to (D), $J(\lambda)$ is clearly the expected long-run average cost of following a policy that induces λ. Further, the policy π_λ defined by $\pi_\lambda(a|s) = \frac{\lambda_{sa}}{\sum_{a'}\lambda_{sa'}}$ induces λ (see Thm. 8.8.2 in [27]). Thus, once the optimal λ^* is obtained by solving (D), the corresponding policy π_{λ^*} is optimal for $(\mathcal{S}, \mathcal{A}, p, c)$.

An analogous problem, (Q), can be used to minimize (2.1) over $(\mathcal{S}, \mathcal{A}, p, c)$:

$$(Q)\qquad \begin{aligned} \min_{\lambda\geq 0}\quad & \rho(\lambda) = \frac{J(\lambda)}{\kappa + \widehat{H}(\lambda)} \\ \text{s.t.}\quad & \sum_{s,a}\lambda_{sa} = 1 \\ & \sum_a \lambda_{sa} = \sum_{s',a} p(s|s',a)\lambda_{s'a} \quad \forall s \in \mathcal{S}, \end{aligned}$$

where $\widehat{H}(\lambda) = H(d^\lambda)$ denotes the entropy of $d^\lambda \in \mathcal{D}(\mathcal{S})$ given by $d_s^\lambda = \sum_a \lambda_{sa}$. Furthermore, in the standard definition of the function $H(d)$, for any $d_i = 0$, we take

$d_i \log d_i = \lim_{d_i \to 0^+} d_i \log d_i = 0$, so $H(d)$ is always well-defined and finite for $d \geq 0$ (see, e.g., [8]). Similarly, we take $d_s^\lambda \log d_s^\lambda = 0$ whenever $d_s^\lambda = 0$, so that $\widehat{H}(\lambda)$ is well-defined for $\lambda \geq 0$. To ensure that the objective of (Q) is well-defined we assume the following.

Assumption 3.1. For all λ feasible to (Q), d^λ has at least two nonzero entries.

This ensures $\rho(\lambda_\pi)$ is well-defined for any π and is weaker than the ergodicity conditions frequently encountered in the RL literature (cf. Assumption 5.1).

Since the feasible region of (Q) corresponds to precisely those state-action occupancy measures achievable over $(\mathcal{S}, \mathcal{A}, p, c)$, solving (Q) yields the state-action occupancy measure minimizing $\rho(\lambda)$. Furthermore, as with (D) above, any λ^* optimal to (Q) allows us to recover a policy π_{λ^*} minimizing $\rho(\lambda)$. Unlike (D), however, the objective function in (Q) is the ratio of an affine function and a concave function and is thus nonconvex, so the problem may be difficult to solve directly. Fortunately, due to the *quasiconvexity* of $\rho(\lambda)$ (cf. Definition 3.3), the problem (Q) can be transformed via the substitution $y = \lambda / c^T \lambda, t = 1/c^T \lambda$ and an application of the perspective transform (see Definition 3.7 and [2, Chap. 7]) to the concave program (Q'):

$$
(Q') \qquad \begin{aligned}
\min_{y \geq 0, t} \quad & \kappa t - \sum_{s,a} y_{sa} \log \left(\frac{\sum_a y_{sa}}{t} \right) \\
\text{s.t.} \quad & \sum_{s,a} y_{sa} = t, \quad \sum_{s,a} c_{sa} y_{sa} = 1 \\
& \sum_a y_{sa} = \sum_{s',a} p(s|s',a) y_{s'a} \quad \forall s \in \mathcal{S}.
\end{aligned}
$$

This problem can be efficiently solved using well-known methods for convex optimization [7] to obtain the optimal state-occupancy measure and corresponding optimal policy. We formalize this as the following theorem.

THEOREM 3.2. *Problem (Q') is a concave program, and any optimal solution to it is optimal for the OIR problem (Q).*

In addition to enabling efficient solution when the MDP model is known, this reformulation implies the existence of *hidden quasiconvexity* underlying any policy gradient methods developed for the OIR minimization problem, as shown in section 5.1.

Proof of Theorem 3.2.

3.1.1. Quasiconvexity of (Q). Let us first formally define quasiconvexity/-concavity. Given a scalar α and function $f : \mathbb{R}^n \to \mathbb{R}$ defined on a convex set $C \subset \mathbb{R}^n$, define the α-superlevel set of f on C to be $U(f, \alpha) = \{x \in C \mid f(x) \geq \alpha\}$ and the α-sublevel set of f on C to be $L(f, \alpha) = \{x \in C \mid f(x) \leq \alpha\}$.

DEFINITION 3.3. *Given $f : \mathbb{R}^n \to \mathbb{R}$ defined on a convex set $C \subset \mathbb{R}^n$, f is quasiconvex (resp., quasiconcave) if $L(f, \alpha)$ (resp., $U(f, \alpha)$) is convex for each $\alpha \in \mathbb{R}$.*

Now let $\Delta(\mathbb{R}^{|\mathcal{S}| \cdot |\mathcal{A}|})$ denote the unit simplex in $\mathbb{R}^{|\mathcal{S}| \cdot |\mathcal{A}|}$, and let F denote the feasible region of (Q). Clearly F is a convex subset of $\Delta(\mathbb{R}^{|\mathcal{S}| \cdot |\mathcal{A}|})$, since it is defined by linear equality and nonnegativity constraints. Note that the numerator of $\rho(\lambda)$, the objective function in (Q), is convex (linear, in fact). Also notice that $\widehat{H}(\lambda) = H(d^\lambda)$ is concave on F, which follows from the facts that the entropy $H(d)$ is concave in d, d^λ is a linear function of λ, and the composition of a concave function with an affine function is itself concave. This implies that, for any fixed $\kappa \geq 0$, the denominator of $\rho(\lambda)$ is concave and also positive by Assumption 3.1 over all its sublevel subsets of the feasible region. These facts guarantee that (Q) is a quasiconvex program with a well-behaved objective function, as formalized in the following lemma.

LEMMA 3.4. (Q) *is feasible and has an optimal solution with finite objective function value, and the objective* ρ *of* (Q) *is strictly quasiconvex on* F.

Finally, (Q) enjoys the following key property, which guarantees that any solution to the concave program described in the next section provides a globally optimal solution to the OIR minimization problem (2.1).

LEMMA 3.5. *Every local optimum of* (Q) *is a global optimum.*

Proof. The assertion follows directly from Proposition 3.3 in [2]. ∎

3.1.2. Transformation to a concave program. Now that we are assured (Q) is quasiconvex and has no spurious stationary points, we exploit the quasiconvex structure of the OIR to transform (Q) into an equivalent concave program, leveraging results from classic results from the literature on quasiconcave programming [31, 2] along the way. Define $q(\lambda) := 1/\rho(\lambda) = (\kappa + \widehat{H}(\lambda))/J(\lambda)$ and consider the problem

$$(Q'') \qquad \begin{aligned} \max_{\lambda} \quad & q(\lambda) \\ \text{s.t.} \quad & \sum_s \sum_a \lambda_{sa} = 1, \\ & \sum_a \lambda_{sa} = \sum_{s'} \sum_a p(s|s',a)\lambda_{s'a} \quad \forall s \in \mathcal{S}, \\ & \lambda \geq 0. \end{aligned}$$

Note that the feasible region F of (Q'') is identical to that of (Q). We have the following.

LEMMA 3.6. *Problem* (Q) *is equivalent to* (Q'').

Proof. Assume λ^* is optimal for (Q), i.e., $\lambda^* \in F$ and $\rho(\lambda^*) \leq \rho(\lambda)$ for all $\lambda \in F$. By Lemma 3.4, there exists $M > 0$ such that $0 < \rho(\lambda^*) \leq \rho(\lambda) < M < \infty$ for all $\lambda \in F$. Clearly $0 < 1/M < q(\lambda) \leq q(\lambda^*) < \infty$ for all $\lambda \in F$, so λ^* is optimal to (Q''). By an analogous argument, any optimal solution to (Q'') is optimal to (Q). ∎

The foregoing lemma proves that solving (Q'') also solves (Q). Crucially, as shown in Theorem 3.8 below, we can in fact transform (Q'') into a *concave* optimization problem, which will allow us to indirectly solve (Q). Before presenting the theorem, we provide an important definition.

DEFINITION 3.7. *Given* $f : \mathbb{R}^n \to \mathbb{R}$*, the perspective of* f *is the function* $P_f : \mathbb{R}^{n+1} \to \mathbb{R}$ *given by* $P_f(x,t) = tf(x/t)$ *with domain* $\textbf{dom}(P_f) = \{(x,t) \mid x/t \in \textbf{dom}(f), t > 0\}$.

We now proceed with the theorem, whose proof follows that of [2, Prop. 7.2].

THEOREM 3.8. *The quasiconcave program* (Q'') *can be converted via the variable transformation* $y = \frac{\lambda}{c^T\lambda}$, $t = \frac{1}{c^T\lambda}$ *into the following concave program:*

$$(Q''') \qquad \begin{aligned} \max_{y,t} \quad & \kappa t - \sum_s \sum_a y_{sa} \log\left(\frac{\sum_a y_{sa}}{t}\right), \\ \text{s.t.} \quad & \sum_s \sum_a y_{sa} = t, \\ & \sum_a y_{sa} = \sum_{s'} \sum_a p(s|s',a) y_{s'a} \quad \forall s \in \mathcal{S}, \\ & \sum_s \sum_a c_{sa} y_{sa} = 1, \\ & y \geq 0. \end{aligned}$$

Proof. First, the transformation $y = \frac{\lambda}{c^T\lambda}$, $t = \frac{1}{c^T\lambda}$ clearly provides a bijection between the feasible regions of (Q'') and (Q'''). Next, let $f(\lambda) = \kappa + \widehat{H}(\lambda)$ denote the numerator of $q(\lambda) = \big(\kappa + \widehat{H}(\lambda)\big)/c^T\lambda$. It is immediate that $q(\lambda) = tf(y/t)$, and recalling the definition $\widehat{H}(\lambda) = -\sum_{s,a} \lambda_{sa} \log\left(\sum_a \lambda_{sa}\right)$ allows us to see that $tf(y/t) = \kappa t - \sum_{sa} y_{sa} \log\left(\sum_a y_{sa}/t\right)$. The objectives of (Q'') and (Q''') thus share the same value for corresponding points in their feasible regions. Since $\widehat{H}(\lambda)$ is concave in λ, and since the perspective transform of a concave function is itself concave by [7, sect. 3.6.2], the objective $tf(y/t)$ of (Q''') is concave. Finally, since the feasible region of (Q''') is determined by linear equalities and positivity constraints, its feasible region is convex. Problem (Q''') is thus a concave program. ∎

Given an optimal solution (y^*, t^*) to (Q'''), we can recover an optimal solution $\lambda^* = y^*/t^*$ to (Q''), which by Lemma 3.6 is also optimal for (Q). As noted above, the policy $\pi^*(a|s) = \lambda^*_{sa}/\sum_{a'} \lambda^*_{sa'}$ thus minimizes the OIR over the MDP $(\mathcal{S}, \mathcal{A}, p, c)$. This proves Theorem 3.2.

3.2. Policy gradients. Sampling the gradient of (2.1) is not straightforward using existing tools, as obtaining stochastic estimates of $\nabla\rho(\theta)$ involves estimating

$$\nabla\rho(\theta) = \frac{\nabla J(\theta)(\kappa + H(d_\theta)) - J(\theta)\nabla H(d_\theta)}{[\kappa + H(d_\theta)]^2}. \tag{3.1}$$

Though we can use the classical policy gradient theorem (cf. eq. (3.2)) 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 $\nabla\rho(\theta)$.

3.2.1. 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 [36] that, under certain conditions,

$$\nabla J(\theta) = \sum_s d_\theta(s) \sum_a Q_\theta(s,a) \nabla\pi_\theta(a|s) = \mathbb{E}_{\pi_\theta}\Big[(c(s,a) - J(\theta))\nabla \log \pi_\theta(a|s)\Big]. \tag{3.2}$$

By following policy π_θ, we can sample from the right-hand side of (3.2) to estimate $\nabla J(\theta)$, then use this to perform stochastic gradient descent.

3.2.2. 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', the cross-entropy between d_θ and $d_{\theta'}$ is given by $CE(d_\theta, d_{\theta'}) = -\sum_s d_\theta(s)\log d_{\theta'}(s)$ and their Kullback–Leibler (KL) divergence by $D_{KL}(d_\theta \,||\, d_{\theta'}) = \sum_s \log\big(\frac{d_\theta(s)}{d_{\theta'}(s)}\big) d_\theta(s)$. Recall that $CE(d_\theta, d_{\theta'}) = H(d_\theta) + D_{KL}(d_\theta \,||\, d_{\theta'})$. We have the following.

LEMMA 3.9.

$$\text{For any } \theta' \in \Theta, \qquad \nabla H(d_\theta)\big|_{\theta=\theta'} = \nabla CE(d_\theta, d_{\theta'})\big|_{\theta=\theta'}. \tag{3.3}$$

For the proof, see the full version in [33]. This establishes an important fact: *we can estimate* $\nabla H(d_\theta)|_{\theta=\theta_t}$ *by instead estimating* $\nabla CE(d_\theta, d_{\theta_t})|_{\theta=\theta_t}$. At first glance, this simply substitutes one problem for another. However, given a fixed θ_t, for any θ, we can use the policy gradient theorem (3.2) to obtain a tractable expression for $\nabla CE(d_\theta, d_{\theta_t})|_{\theta=\theta_t}$, as described next.

3.2.3. Entropy and OIR policy gradients. Our next results enable policy gradient algorithms for maximizing $H(d_\theta)$ and minimizing (2.1). See [33] for proofs.

THEOREM 3.10. *Let an MDP* $(\mathcal{S}, \mathcal{A}, p, c)$ *and a differentiable parameterized 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.* (3.3)*) with respect to the policy parameters* θ *of the state occupancy measure entropy* $H(d_\theta)$*, evaluated at* $\theta = \theta_t$*, satisfies*

$$\nabla H(d_\theta)\big|_{\theta=\theta_t} = \mathbb{E}_{\pi_{\theta_t}}\Big[\left(-\log d_{\theta_t}(s) - H(d_{\theta_t})\right)\nabla \log \pi_{\theta_t}(a|s)\Big]. \tag{3.4}$$

With Theorem 3.10 in hand, we have the following OIR policy gradient theorem.

THEOREM 3.11. *Let MDP* $(\mathcal{S}, \mathcal{A}, p, c)$*, differentiable policy class* $\{\pi_\theta\}_{\theta\in\Theta}$*, and 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.* (3.1)*) with respect to the policy parameters* θ *of the OIR* $\rho(\theta)$ *(cf.* (2.1)*), evaluated at* $\theta = \theta_t$*, satisfies*

$$\nabla\rho(\theta_t) = \mathbb{E}_{\pi_{\theta_t}}\left[\frac{\delta_t^J\big(\kappa + H(d_{\theta_t})\big) - J(\theta_t)\delta_t^H}{\left[\kappa + H(d_{\theta_t})\right]^2}\psi_t\right], \tag{3.5}$$

where $\delta_t^J = c(s,a) - J(\theta_t)$*,* $\delta_t^H = -\log d_{\theta_t}(s) - H(d_{\theta_t})$*, and* $\psi_t = \nabla \log \pi_{\theta_t}(a|s)$*.*

The claim follows by combining (3.1) and (3.2) with Theorem 3.10.

4. Algorithms. In this section we derive two policy search schemes for minimizing (2.1). The first is based on the classic REINFORCE algorithm, while the second is an actor-critic scheme. We assume throughout that an average-cost MDP $(\mathcal{S}, \mathcal{A}, p, c)$ is fixed. The reward setting can be accommodated by Remark 2.1.

4.1. Information-Directed REINFORCE. The classic REINFORCE algorithm [37] generates a single, finite trajectory using a fixed policy, estimates the gradient of $J(\theta)$ based on the trajectory, and performs a corresponding stochastic gradient descent step. We present a related algorithm, Information-Directed REINFORCE (ID-REINFORCE), that proceeds along similar lines to minimize (2.1). Note that, in order to estimate the $H(d_{\theta_t})$ term in $\nabla\rho(\theta_t)$ (see (3.5)), it is necessary to first estimate d_{θ_t}. This task is addressed both implicitly and explicitly in previous works [10, 19, 39]. As in [10], 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 by computing the empirical visitation probabilities for each of the states $s \in \mathcal{S}$ based on sample trajectories. We focus on this setting in this paper. When $\mathcal{S}$ is large or continuous, on the other hand, various parametric and nonparametric density estimation techniques can be used to implement DensityEstimator. Pseudocode for ID-REINFORCE is given in Algorithm 4.1.

4.2. Information-Directed Actor-Critic. We next present the Information-Directed Actor-Critic (IDAC) algorithm, a variant of the classic actor-critic algorithm [16, 5] 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 the shadow reward discussed in the proof of Theorem 3.10. We assume access to the DensityEstimator oracle throughout. The classic actor-critic algorithm for minimizing $J(\theta)$ alternates between *critic* and *actor* updates. At each time step, it first computes the temporal difference (TD) error, which is a bootstrapped estimate of the amount by which the current state value function approximator, known as the *critic,* over- or underestimates the true value of the current state (see [35] for details). This TD error is then used to update the critic, which is in turn used to update the policy, or *actor.* For IDAC, we modify the classic scheme by (i) introducing an entropy critic to estimate the entropy gradient (lines 10 and 14), and (ii) altering the policy update to take a gradient descent step in the direction $-\nabla \rho(\theta_t)$ instead of $-\nabla J(\theta_t)$ (line 15). Pseudocode is provided in Algorithm 4.2.

Algorithm 4.1. ID-REINFORCE.

Initialization: Rollout length K, stepsizes $\eta > 0$ and $\tau \in (0,1]$, policy class $\{\pi_\theta\}_{\theta\in\Theta}$, entropy constant $\kappa \geq 0$. Sample s_0 and θ_0, select μ_{-1}^H, $\mu_{-1}^J > 0$. $t \leftarrow 0$.
repeat
  Generate $\{(s_i, a_i)\}_{i=1,\dots,K} \sim \pi_{\theta_t}$
  $\widehat{J(\theta_t)} = \frac{1}{K}\sum_{i=1}^K c(s_i, a_i)$
  $\mu_t^J = (1-\tau)\mu_{t-1}^J + \tau\widehat{J(\theta_t)}$
  $d_{\theta_t} = \text{DensityEstimator}(\theta_t)$
  $\widehat{H(d_{\theta_t})} = \frac{1}{K}\sum_{i=1}^K(-\log d_{\theta_t}(s_i))$
  $\mu_t^H = (1-\tau)\mu_{t-1}^H + \tau\widehat{H(d_{\theta_t})}$
  for $i = 1, \dots, k$ **do**
    $\delta_i^J = c(s_i, a_i) - \mu_t^J$
    $\delta_i^H = -\log d_{\theta_t}(s_i) - \mu_t^H$
    $\psi_i = \nabla \log \pi_{\theta_t}(a_i|s_i)$
    $\Delta_i = \delta_i^J\left(\kappa + \mu_t^H\right) - \mu_t^J \delta_i^H$
  end for
  $\theta_{t+1} = \theta_t - \eta \frac{1}{K\left[\kappa + \mu_t^H\right]^2}\sum_{i=1}^K \Delta_i \psi_i$
  $t \leftarrow t+1$
until convergence

Algorithm 4.2. IDAC.

1: **Initialization:** Rollout length K, stepsizes $\{\alpha_t\}, \{\beta_t\}, \{\tau_t\}$, policy class $\{\pi_\theta\}_{\theta\in\Theta}$, critic class $\{v_\omega\}_{\omega\in\Omega}$, entropy constant $\kappa \geq 0$. Sample $s_0, \theta_0, \omega_0^J, \omega_0^H$, select μ_{-1}^H, $\mu_{-1}^J > 0$. $t \leftarrow 0$.
2: **repeat**
3:   Generate $\{(s_i, a_i)\}_{i=1,\dots,K} \sim \pi_{\theta_t}$
4:   $\mu_t^J = (1-\tau)\mu_{t-1}^J + \tau\frac{1}{K}\sum_{i=1}^K c(s_i, a_i)$
5:   $d_{\theta_t} = \text{DensityEstimator}(\theta_t)$
6:   $\mu_t^H = (1-\tau)\mu_{t-1}^H - \tau\frac{1}{K}\sum_{i=1}^K \log d_{\theta_t}(s_i)$
7:   **for** $i = 1, \dots, K$ **do**
8:     Set $v_{\omega_t^J}(s_{K+1}) = v_{\omega_t^H}(s_{K+1}) = 0$
9:     $\delta_i^J = c(s_i, a_i) - \mu_t^J + v_{\omega_t^J}(s_{i+1}) - v_{\omega_t^J}(s_i)$
10:     $\delta_i^H = -\log d_{\theta_t}(s_i) - \mu_t^H + v_{\omega_t^H}(s_{i+1}) - v_{\omega_t^H}(s_i)$
11:     $\psi_i = \nabla \log \pi_{\theta_t}(a_i|s_i)$
12:     $\Delta_i = \delta_i^J\left(\kappa + \mu_t^H\right) - \mu_t^J \delta_i^H$
13:   **end for**
14:   $\omega_{t+1}^J = \omega_t^J + \alpha\frac{1}{K}\sum_{i=1}^K \delta_i^J \nabla v_{\omega_t^J}(s_i)$
15:   $\omega_{t+1}^H = \omega_t^H + \alpha\frac{1}{K}\sum_{i=1}^K \delta_i^H \nabla v_{\omega_t^H}(s_i)$
16:   $\theta_{t+1} = \theta_t - \beta\frac{1}{K\left[\kappa + \mu_t^H\right]^2}\sum_{i=1}^K \Delta_i \psi_i$
17:   $t \leftarrow t+1$
18: **until** convergence

4.3. Density estimation. These algorithms are similar to algorithms in [10] in that they need to estimate d_θ, which can be inefficient in continuous, high-dimensional spaces. There are two promising approaches for overcoming this. First, a variety of density estimation techniques have been successfully employed in RL and imitation learning in continuous settings, including kernel density estimation, variational autoencoders, energy-based models, and autoregressive models [10, 19, 13]. Second, particle-based methods have recently been successfully used to avoid density estimation altogether by directly estimating occupancy measure entropy [24, 21]. The density estimation issue can likely be mitigated, providing an important direction for future work.

5. Theoretical results. In this section we provide key results underpinning policy search for the OIR problem. In section 5.1, we show that all stationary points of $\rho(\theta)$ are in fact global minimizers. In section 5.2, we prove that the stochastic gradient descent scheme underlying ID-REINFORCE enjoys a nonasymptotic convergence rate depending on κ, the policy class, and ergodicity properties of the underlying MDP. Finally, section 5.3 establishes that IDAC enjoys asymptotic, almost sure convergence to a neighborhood of a stationary point. Taken together, these results prove that both algorithms converge to globally optimal solutions under suitable conditions.

5.1. Stationarity implies global optimality. As we will see, the OIR optimization problem enjoys a powerful *hidden quasiconvexity* 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

$$\min_{\theta \in \Theta} \quad \rho(\theta) = \frac{J(\theta)}{\kappa + H(d_\theta)}. \tag{5.1}$$

This result is surprising, as the objective function $\rho(\theta)$ is typically highly nonconvex. Let $\Theta \subset \mathbb{R}^k$ be convex, and let a parameterized 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 5.1. 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 5.2. The following statements hold:
1. $\lambda(\cdot)$ is a bijection between Θ and $\lambda(\Theta)$, 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.

Assumption 5.1 is standard in the policy gradient literature, and it implies that $\nabla\rho(\theta)$ exists for all $\theta \in \Theta$. Assumption 5.2 holds for reasonable examples and can likely be proven to hold in the tabular setting under suitable ergodicity conditions on the underlying MDP. See [33] for a detailed example for which Assumption 5.1 holds.

We now have the following theorem. The key idea behind the proof is to show that the stationary point θ^* corresponds to an optimal solution to the concave program (Q''') and thus also provides an optimal solution to the quasiconvex OIR minimization problem (Q). The proof builds on that of Theorem 4.2 of [39], with key modifications to accommodate the fact that the underlying OIR optimization problem is not convex, but *quasiconvex*, in the state-action occupancy measure. In particular, the result in [39] holds for convex (or concave) functionals of the state-action occupancy measure only, not ones involving *quasiconvex* (or quasiconcave) functionals. The critical innovation in the proof of Theorem 5.3 below is to leverage properties of the perspective transform combined with the smoothness conditions of Assumption 5.2 to extend the hidden convexity analysis of [39] to the quasiconvex setting.

THEOREM 5.3. *Let Assumptions* 5.1 *and* 5.2 *hold. Let* θ^* *be a stationary point of* (5.1)*, i.e.,* $\nabla\rho(\theta^*) = 0$*. Then* θ^* *is globally optimal for* (5.1)*.*

Proof. We reformulate (5.1) as a maximization problem. Let $q(\theta) = 1/\rho(\theta) = (\kappa + H(d_\theta))/J(\theta)$. Let $\widehat{H}(\lambda_\theta) = H(d_\theta)$, where $\widehat{H}(\lambda) = H(d^\lambda)$ is the entropy of the state occupancy measure $d^\lambda \in \mathcal{D}(\mathcal{S})$ given by $d^\lambda_s = \sum_a \lambda_{sa}$, and recall that $J(\theta) = c^T\lambda_\theta$ for some vector $c > 0$ of costs. This means that $q(\theta) = (\kappa + \widehat{H}(\lambda_\theta))/J(\lambda_\theta)$. In what follows we prove that θ^* is globally optimal for $\max_{\theta\in\Theta} q(\theta)$. By Lemma 3.6, this will imply that θ^* is globally optimal for $\min_{\theta\in\Theta}\rho(\theta)$. Also note that since $\rho(\theta)$ is strictly positive on Θ, we know $q(\theta)$ is differentiable in θ and $\nabla q(\theta) = -\nabla\rho(\theta)/[\rho(\theta)]^2$ for all $\theta \in \Theta$. Since $\nabla\rho(\theta^*) = 0$ by assumption, this means $\nabla q(\theta^*) = 0$, so θ^* is a stationary point of the optimization problem $\max_{\theta\in\Theta} q(\theta)$.

We now transform the problem $\max_{\theta\in\Theta} q(\theta)$ to a concave program. For $z \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|+1}$, let $y \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ denote all but the last entry in z, and let the scalar t denote the last entry of z. We will write $z = (y, t)$ for brevity. Let $\zeta : \mathcal{D}(\mathcal{S}\times\mathcal{A}) \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|+1}$ be the mapping given by $\zeta(\lambda) = (\lambda/J(\lambda), 1/J(\lambda))$. Consider the optimization problems

$$\max_{\lambda\in\lambda(\Theta)} \quad \frac{\kappa + \widehat{H}(\lambda)}{J(\lambda)}, \tag{5.2}$$

$$\max_{z\in(\zeta\circ\lambda)(\Theta)} \quad P_{\kappa,\widehat{H}}(z), \tag{5.3}$$

where $P_{\kappa,\widehat{H}} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|+1} \to \mathbb{R}$ denotes the perspective transformation of $\kappa + \widehat{H}(\lambda)$, given by $P_{\kappa+\widehat{H}}(z) = P_{\kappa,\widehat{H}}((y,t)) = t(\kappa + \widehat{H}(y/t))$. For notational convenience we henceforth drop the dependency on κ and simply write $P_{\widehat{H}}$ and $\widehat{H}$ instead of $P_{\kappa,\widehat{H}}$ and $\kappa + \widehat{H}$. Recall that since $\widehat{H}$ is concave over the region $\mathcal{D}(\mathcal{S}\times\mathcal{A})$, its perspective transform $P_{\widehat{H}}$ is concave over the region $\zeta(\mathcal{D}(\mathcal{S}\times\mathcal{A}))$. $P_{\widehat{H}}$ is thus concave over the convex, compact region $(\zeta\circ\lambda)(\Theta) \subseteq \zeta(\mathcal{D}(\mathcal{S}\times\mathcal{A}))$.

The remainder of the proof provides the technical details demonstrating that $z^* = (\zeta\circ\lambda)(\theta^*)$ is a stationary point of (5.3). Since (5.3) is a concave program, this will imply that z^* and thus θ^* are globally optimal for their respective problems. We first show Assumption 5.2 can be extended to the mapping $\zeta\circ\lambda$. To do so, we prove the following:

(i) $\zeta\circ\lambda$ gives a bijection between Θ and $(\zeta\circ\lambda)(\Theta)$;
(ii) $\zeta\circ\lambda$ has a Lipschitz continuous inverse; and
(iii) the Jacobian $\nabla_\theta(\zeta\circ\lambda)(\theta)$ is Lipschitz.

To prove (i), recall $\zeta(\lambda) = (\lambda/J(\lambda), 1/J(\lambda))$. We know ζ is a surjection onto $(\zeta\circ\lambda)(\Theta)$ by definition, so we just need to show it is injective. Fix $\lambda \neq \lambda'$. If $J(\lambda) = J(\lambda')$, then $\lambda/J(\lambda) \neq \lambda'/J(\lambda')$, so $\zeta(\lambda) \neq \zeta(\lambda')$. If $J(\lambda) \neq J(\lambda')$, on the other hand, then $1/J(\lambda) \neq 1/J(\lambda')$, so again $\zeta(\lambda) \neq \zeta(\lambda')$. Therefore ζ is injective and thus gives a bijection. Combined with Assumption 5.2, the foregoing implies that $\zeta\circ\lambda$ gives a bijection between Θ and $(\zeta\circ\lambda)(\Theta)$, proving (i).

For (ii), the inverse of ζ is clearly $\zeta^{-1}(z) = \zeta^{-1}((y,t)) = y/t$. Since $0 < \min_i c_i \leq t \leq \max_i c_i < \infty$, ζ^{-1} has continuous, bounded partial derivatives and is thus Lipschitz continuous on $(\zeta\circ\lambda)(\Theta)$. Since the composition of Lipschitz functions is Lipschitz, $k = (\zeta\circ\lambda)^{-1} = \lambda^{-1}\circ\zeta^{-1}$ is Lipschitz continuous, proving (ii).

For (iii), an application of the chain rule gives $\nabla_\theta(\zeta\circ\lambda)(\theta) = [\nabla_\lambda \zeta(\lambda(\theta))]^T \nabla_\theta \lambda(\theta)$. Clearly $\nabla_\lambda \zeta(\lambda)$ is Lipschitz continuous and bounded over the compact set Θ. Since $\nabla_\theta \lambda(\theta)$ is (Lipschitz) continuous and bounded over Θ, we know $\lambda(\theta)$ is Lipschitz, implying that $\nabla_\lambda \zeta(\lambda(\theta))$ is Lipschitz and bounded on $\theta \in \Theta$. Furthermore, $\nabla_\theta \lambda(\theta)$ is Lipschitz by assumption and bounded over Θ, so all entries in the matrix product $[\nabla_\lambda \zeta(\lambda(\theta))]^T \nabla_\theta \lambda(\theta)$ are sums and products of Lipschitz, bounded functions over Θ. This implies that $\nabla_\theta(\zeta\circ\lambda)(\theta)$ is Lipschitz on Θ, proving (iii).

We now move on to the bounding arguments that will ultimately prove that $z^* = (\zeta\circ\lambda)(\theta^*)$ is a stationary point of (5.3). First, notice that

$$(P_{\widehat{H}} \circ \zeta \circ \lambda)(\theta) = P_{\widehat{H}}(\zeta(\lambda(\theta))) = \frac{\kappa + \widehat{H}(\lambda(\theta))}{J(\lambda(\theta))} = \frac{\kappa + \widehat{H}(\lambda_\theta)}{J(\lambda_\theta)} = q(\theta),$$

so $\nabla_\theta(P_{\widehat{H}} \circ \zeta \circ \lambda)(\theta^*) = \nabla_\theta \frac{\kappa+\widehat{H}(\lambda_{\theta^*})}{J(\lambda_{\theta^*})} = \nabla q(\theta^*) = 0$. Since $P_{\widehat{H}}$ is concave and locally Lipschitz on $(\zeta\circ\lambda)(\Theta)$, by the chain rule we have $\nabla_\theta(P_{\widehat{H}} \circ \zeta \circ \lambda)(\theta^*) = [\nabla_\theta(\zeta\circ\lambda)(\theta^*)]^T \nabla_z P_{\widehat{H}}(z^*) = 0$. This implies, for all $\theta \in \Theta$,

$$\langle \nabla_z P_{\widehat{H}}(z^*), \nabla_\theta(\zeta\circ\lambda)(\theta^*)(\theta-\theta^*)\rangle = \langle [\nabla_\theta(\zeta\circ\lambda)(\theta^*)]^T \nabla_z P_{\widehat{H}}(z^*), \theta-\theta^*\rangle = 0. \tag{5.4}$$

Equation (5.4) is important to the bounding arguments presented next.

In the following equations, let $\theta = k(z)$ and $\theta^* = k(z^*)$. Adding and subtracting $\langle \nabla_z P_{\widehat{H}}(z^*), \nabla_\theta(\zeta\circ\lambda)(\theta^*)(\theta-\theta^*)\rangle$, using (5.4), and applying the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
\langle \nabla_z P_{\widehat{H}}(z^*), z-z^*\rangle &= \langle \nabla_z P_{\widehat{H}}(z^*), (\zeta\circ\lambda)(\theta) - (\zeta\circ\lambda)(\theta^*)\rangle \\
&= \langle \nabla_z P_{\widehat{H}}(z^*), \nabla_\theta(\zeta\circ\lambda)(\theta^*)(\theta-\theta^*)\rangle \\
&\quad + \langle \nabla_z P_{\widehat{H}}(z^*), (\zeta\circ\lambda)(\theta) - (\zeta\circ\lambda)(\theta^*) - \nabla_\theta(\zeta\circ\lambda)(\theta^*)(\theta-\theta^*)\rangle \\
&= \langle \nabla_z P_{\widehat{H}}(z^*), (\zeta\circ\lambda)(\theta) - (\zeta\circ\lambda)(\theta^*) - \nabla_\theta(\zeta\circ\lambda)(\theta^*)(\theta-\theta^*)\rangle \\
&\le \left\|\nabla_z P_{\widehat{H}}(z^*)\right\| \left\|(\zeta\circ\lambda)(\theta) - (\zeta\circ\lambda)(\theta^*) - \nabla_\theta(\zeta\circ\lambda)(\theta^*)(\theta-\theta^*)\right\|.
\end{aligned}$$

Since $\nabla_\theta(\zeta\circ\lambda)(\theta)$ is Lipschitz, there exists $K_0 > 0$ such that

$$\left\|(\zeta\circ\lambda)(\theta) - (\zeta\circ\lambda)(\theta^*) - \nabla_\theta(\zeta\circ\lambda)(\theta^*)(\theta-\theta^*)\right\| \le \frac{K_0}{2}\left\|\theta-\theta^*\right\|^2.$$

In addition, $k = (\zeta\circ\lambda)^{-1}$ is Lipschitz, so there exists $K_1 > 0$ such that

$$\left\|\theta-\theta^*\right\|^2 = \left\|k(z)-k(z^*)\right\|^2 \le K_1^2 \left\|z-z^*\right\|^2.$$

Combining these inequalities yields that

$$\langle \nabla_z P_{\widehat{H}}(z^*), z-z^*\rangle \le \frac{K_0 K_1^2}{2}\left\|\nabla_z P_{\widehat{H}}(z^*)\right\| \left\|z-z^*\right\|^2$$

for all $z \in (\zeta\circ\lambda)(\Theta)$. Since $(\zeta\circ\lambda)(\Theta)$ is convex, we can replace z above with $(1-\alpha)z^* + \alpha z$ for any $\alpha \in [0,1]$, which gives

$$\alpha\langle \nabla_z P_{\widehat{H}}(z^*), z-z^*\rangle \le \frac{K_0 K_1^2 \alpha^2}{2}\left\|\nabla_z P_{\widehat{H}}(z^*)\right\| \left\|z-z^*\right\|^2$$

for all $z \in (\zeta\circ\lambda)(\Theta)$ and $\alpha \in [0,1]$. Dividing both sides by α and taking the limit as α approaches 0 from above, we obtain

$$\langle \nabla_z P_{\widehat{H}}(z^*), z-z^*\rangle \le 0 \quad \forall z \in (\zeta\circ\lambda)(\Theta).$$

Since problem (5.3) is concave in z, this implies that $z^* = (\zeta \circ \lambda)(\theta^*)$ is a stationary point of that problem. The solution z^* is therefore a global optimal solution to (5.3), implying that θ^* is globally optimal for (5.1). ∎

This powerful hidden quasiconvexity 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 [4, 1, 23, 39, 3] using persistent exploration conditions, our result instead builds on hidden convexity arguments from [39], which apply to parameterized policies. However, Theorem 5.3 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 *quasiconvexity* for ratio objectives, not hidden *convexity,* which requires reformulation via an application of the perspective transform (cf. section 3.1). In these ways, Theorem 5.3 is a strict generalization of existing results.

5.2. Nonasymptotic convergence rate. We next establish a convergence rate for the following projected gradient descent scheme for solving (5.1):

$$\theta_{t+1} = \mathrm{Proj}_\Theta\left(\theta_t - \eta\nabla\rho(\theta_t)\right) = \arg\min_\theta \left[\rho(\theta_t) + \langle\nabla\rho(\theta_t), \theta-\theta_t\rangle + \frac{1}{2\eta}\left\|\theta-\theta_t\right\|^2\right] \tag{5.5}$$

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 (5.5) is a reformulation of Algorithm 4.1 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. Recall the mapping $\zeta : \mathcal{D}(\mathcal{S}\times\mathcal{A}) \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|+1}$ from the proof of Theorem 5.3, which was 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 5.1 and properties of entropy, $\min_\theta \rho(\theta) > 0$ and $\max_\theta \rho(\theta) < \infty$. In addition to Assumptions 5.1 and 5.2, we will need the following.

Assumption 5.4. $\nabla\rho(\theta)$ is Lipschitz and $L > 0$ is the smallest number such that $\|\nabla\rho(\theta) - \nabla\rho(\theta')\| \le L\|\theta-\theta'\|$ for all $\theta,\theta'\in\Theta$.

We have the following convergence rate result for the projected gradient descent scheme (5.5). The key idea behind the proof is to link the objective function $\rho(\theta)$ that the updates (5.5) are minimizing to the concave structure of the transformed problem (Q'''). This allows us to derive the bound (5.6) by studying related bounds for the concave objective function of (Q'''). Similar to Theorem 5.3 above, our key innovation is that we leverage properties of the perspective transform combined with the Lipschitz condition of Assumption 5.4 to extend the rate analysis of Theorem 4.4 in [39]—which holds only for convex (or concave) functionals of the state-action occupancy measure—to the quasiconvex setting.

THEOREM 5.5. *Let Assumptions* 5.1, 5.2, *and* 5.4 *hold. Denote the diameter of the convex, compact set* $(\zeta\circ\lambda)(\Theta)$ *by* $D_\zeta = \max_{z,z'\in(\zeta\circ\lambda)(\Theta)}\|z-z'\|$. *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* $t\ge 0$,

$$\rho(\theta_t) - \rho(\theta^*) \le \frac{4M^2 L_1 \ell^2 D_\zeta^2}{t+1}, \tag{5.6}$$

where ℓ is the minimal Lipschitz constant of the inverse mapping $(\zeta \circ \lambda)^{-1}$.

Proof. To draw the connection with (Q'''), we first transform (5.5) into an equivalent projected gradient *ascent* scheme. Define $q(\theta) = 1/\rho(\theta) = (\kappa + H(d_\theta))/J(\theta)$, and notice that $\nabla q(\theta) = -\nabla\rho(\theta)/[\rho(\theta)]^2$. The projected gradient ascent scheme becomes

$$\begin{aligned}\theta_{t+1} = \operatorname{Proj}_\Theta(\theta_t - \eta\nabla\rho(\theta_t)) &\overset{(a)}{=} \operatorname{Proj}_\Theta\left(\theta_t + \eta\,[\rho(\theta_t)]^2\,\nabla q(\theta_t)\right)\\ &\overset{(b)}{=} \arg\max_\theta \left(q(\theta_t) + \langle \nabla q(\theta_t), \theta - \theta_t\rangle - \frac{[\rho(\theta_t)]^2}{2\eta}\|\theta - \theta_t\|^2\right)\\ &\overset{(c)}{=} \arg\max_\theta \left(q(\theta_t) + \langle \nabla q(\theta_t), \theta - \theta_t\rangle - \frac{1}{2\eta_t}\|\theta - \theta_t\|^2\right),\end{aligned} \tag{5.7}$$

where (a) follows by noticing that $\nabla\rho(\theta) = -[\rho(\theta)]^2\nabla q(\theta)$ and making the appropriate substitution, (b) holds by definition of the Proj operator, and (c) results from defining $\eta_t = \eta[\rho(\theta_t)]^2 = [\rho(\theta_t)]^2/K_0$.

We next identify a family $\{K_c \mid c \in [m, M]\}$ of Lipschitz constants of the gradient $\nabla q(\theta)$. By Assumption 5.4, $\nabla q(\theta)$ is Lipschitz and $\|\nabla q(\theta) - \nabla q(\theta')\| \le L_0\|\theta - \theta'\|$, for all $\theta, \theta' \in \Theta$, where $L_0 = m^2 L$. Let $K = \max\{L_0, M^2 L_0\} = \max\{m^2 L, M^2 m^2 L\}$. Then, for all scalars $c \in [m, M]$, the gradient $\nabla q(\theta)$ satisfies $\|\nabla q(\theta) - \nabla q(\theta')\| \le K_c\|\theta - \theta'\|$, where $K_c = K/c^2$ is the desired Lipschitz constant. This family of Lipschitz constants will be critical in the analysis to follow.

We now study the sequence $\alpha_t = [q(\theta^*) - q(\theta_t)]/2K_m\ell^2 D_\zeta^2$ for $t \ge 0$, ultimately using properties of the sequence to show that $q(\theta^*) - q(\theta_t) \le 4K_m\ell^2 D_\zeta^2/(1+t)$ for all $t \ge 0$. The remainder of the proof proceeds along lines similar to that of Theorem 4.4 in [39], with critical modifications to accommodate the use of the perspective transform, the variable transformation ζ, and the nonconstant stepsizes $\eta_t = \eta[\rho(\theta_t)]^2$.

Define $\widehat{H}(\lambda) = H(d^\lambda)$, where $H(d^\lambda)$ is the entropy of the state occupancy measure $d^\lambda(s) = \sum_a \lambda(s,a)$ corresponding to the state-action occupancy measure $\lambda \in \mathcal{D}(\mathcal{S}\times\mathcal{A})$. For a given $\kappa \ge 0$, let $P_{\kappa,\widehat{H}} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|+1} \to \mathbb{R}$ denote the perspective transformation of $\kappa + \widehat{H}(\lambda)$, given by $P_{\kappa+\widehat{H}}(z) = P_{\kappa,\widehat{H}}((y,t)) = t(\kappa + \widehat{H}(y/t))$. For notational convenience we henceforth drop the dependency on κ and simply write $P_{\widehat{H}}$.

Our next step is to make use of the concave structure of $P_{\widehat{H}}$, $\zeta \circ \lambda$, and (Q''') to analyze $\{\alpha_t\}_{t\in\mathbb{N}}$. We first derive several useful inequalities regarding $P_{\widehat{H}}$ and $\zeta \circ \lambda$. Notice that $q(\theta) = (\kappa + \widehat{H}(\lambda_\theta))/J(\lambda_\theta) = P_{\widehat{H}}(\zeta(\lambda_\theta)) = P_{\widehat{H}}((\zeta\circ\lambda)(\theta))$ for all $\theta \in \Theta$. This means that we can rewrite α_t as

$$\alpha_t = \left[P_{\widehat{H}}((\zeta\circ\lambda)(\theta^*)) - P_{\widehat{H}}((\zeta\circ\lambda)(\theta_t))\right]/2K_m\ell^2 D_\zeta^2. \tag{5.8}$$

Notice $P_{\widehat{H}}$ is concave over $\zeta(\mathcal{D}(\mathcal{S}\times\mathcal{A}))$, since $\widehat{H}$ is concave over $\mathcal{D}(\mathcal{S}\times\mathcal{A})$ and the perspective transform preserves concavity. $P_{\widehat{H}}$ is thus concave over the convex, compact region $(\zeta\circ\lambda)(\Theta) \subseteq \zeta(\mathcal{D}(\mathcal{S}\times\mathcal{A}))$. Furthermore, since $P_{\widehat{H}}((\zeta\circ\lambda)(\theta)) = q(\theta)$, we have $\nabla P_{\widehat{H}}((\zeta\circ\lambda)(\theta)) = \nabla q(\theta)$, so $\nabla P_{\widehat{H}}((\zeta\circ\lambda)(\theta))$ is K_c-Lipschitz for any $K_c = K/c^2, c \in [m, M]$. This implies (see Lemma 1.2.3 in [25]), for any $c \in [m, M]$, that

$$\left|P_{\widehat{H}}((\zeta\circ\lambda)(\theta)) - P_{\widehat{H}}((\zeta\circ\lambda)(\theta_t)) - \left\langle \nabla P_{\widehat{H}}((\zeta\circ\lambda)(\theta_t)), \theta - \theta_t\right\rangle\right| \le \frac{K_c}{2}\left\|\theta-\theta_t\right\|^2,$$

whence, for any $\theta \in \Theta$,

$$\begin{aligned} P_{\widehat{H}}((\zeta\circ\lambda)(\theta)) &\ge P_{\widehat{H}}((\zeta\circ\lambda)(\theta_t)) + \left\langle \nabla P_{\widehat{H}}((\zeta\circ\lambda)(\theta_t)), \theta - \theta_t\right\rangle - \frac{K_c}{2}\left\|\theta-\theta_t\right\|^2 \\ &\ge P_{\widehat{H}}((\zeta\circ\lambda)(\theta)) - K_c\left\|\theta-\theta_t\right\|^2. \end{aligned} \tag{5.9}$$

In light of these inequalities, and using the fact that $\eta_t = [\rho(\theta_t)]^2/K = 1/K_{\rho(\theta_t)}$ by setting $c = \rho(\theta_t)$ in the definition of K_c, we have

(5.10)

$$\begin{aligned} &P_{\widehat{H}}((\zeta\circ\lambda)(\theta_{t+1})) \ge P_{\widehat{H}}((\zeta\circ\lambda)(\theta_t)) + \left\langle \nabla P_{\widehat{H}}((\zeta\circ\lambda)(\theta_t)), \theta_{t+1} - \theta_t\right\rangle \\ &\quad - \frac{K_{\rho(\theta_t)}}{2}\left\|\theta_{t+1}-\theta_t\right\|^2 \\ &\overset{(a)}{=} \max_{\theta\in\Theta}\left(P_{\widehat{H}}((\zeta\circ\lambda)(\theta_t)) + \left\langle \nabla P_{\widehat{H}}((\zeta\circ\lambda)(\theta_t)), \theta - \theta_t\right\rangle - \frac{K_{\rho(\theta_t)}}{2}\left\|\theta-\theta_t\right\|^2\right) \\ &\overset{(b)}{\ge} \max_{\theta\in\Theta}\left(P_{\widehat{H}}((\zeta\circ\lambda)(\theta)) - K_{\rho(\theta_t)}\left\|\theta-\theta_t\right\|^2\right) \overset{(c)}{\ge} \max_{\theta\in\Theta}\left(P_{\widehat{H}}((\zeta\circ\lambda)(\theta)) - K_m\left\|\theta-\theta_t\right\|^2\right) \\ &\overset{(d)}{\ge} \max_{\alpha\in[0,1]}\{P_{\widehat{H}}((\zeta\circ\lambda)(\theta_\alpha)) - K_m\left\|\theta_\alpha-\theta_t\right\|^2 \;\big|\; \theta_\alpha = k\big(\alpha(\zeta\circ\lambda)(\theta^*) \\ &\quad + (1-\alpha)(\zeta\circ\lambda)(\theta_t)\big)\}, \end{aligned}$$

where (a) follows from the optimality of the update (5.7), (b) holds by (5.9), (c) follows from the fact that $K_{\rho(\theta_t)} \le K_m$, and (d) follows by the convexity of $(\zeta\circ\lambda)(\Theta)$.

Let $k(\cdot) = (\zeta\circ\lambda)^{-1}(\cdot)$ as in Theorem 5.3. By Assumption 5.2 and the proof of Theorem 5.3, we know that $k(\cdot)$ is ℓ-Lipschitz. Now notice that

$$\begin{aligned} P_{\widehat{H}}((\zeta\circ\lambda)(\theta_\alpha)) &= P_{\widehat{H}}\Big((\zeta\circ\lambda)\big(k(\alpha(\zeta\circ\lambda)(\theta^*) + (1-\alpha)(\zeta\circ\lambda)(\theta_t))\big)\Big) \\ &= P_{\widehat{H}}(\alpha(\zeta\circ\lambda)(\theta^*) + (1-\alpha)(\zeta\circ\lambda)(\theta_t)) \\ &\ge \alpha P_{\widehat{H}}((\zeta\circ\lambda)(\theta^*)) + (1-\alpha)P_{\widehat{H}}((\zeta\circ\lambda)(\theta_t)), \end{aligned} \tag{5.11}$$

where the first equality holds by the definition of θ_α given in (5.10), the second follows from the fact that $k((\zeta\circ\lambda)(\theta)) = \theta$ for any $\theta\in\Theta$, and the final inequality is yielded by the concavity of $P_{\widehat{H}}$ over $(\zeta\circ\lambda)(\Theta)$. Furthermore,

$$\begin{aligned} \left\|\theta_\alpha-\theta_t\right\|^2 &\overset{(a)}{=} \left\|k(\alpha(\zeta\circ\lambda)(\theta^*) + (1-\alpha)(\zeta\circ\lambda)(\theta_t)) - k((\zeta\circ\lambda)(\theta_t))\right\|^2 \\ &\overset{(b)}{\le} \ell^2\left\|\alpha(\zeta\circ\lambda)(\theta^*) + (1-\alpha)(\zeta\circ\lambda)(\theta_t) - (\zeta\circ\lambda)(\theta_t)\right\|^2 \\ &\le \alpha^2\ell^2\left\|(\zeta\circ\lambda)(\theta^*) - (\zeta\circ\lambda)(\theta_t)\right\|^2 \overset{(c)}{\le} \alpha^2\ell^2 D_\zeta^2, \end{aligned} \tag{5.12}$$

where (a) holds by the definition of θ_α and the fact that $k((\zeta\circ\lambda)(\theta)) = \theta$, (b) follows since k is ℓ-Lipschitz, and (c) results from the definition of D_ζ given in the statement of the theorem. Now, the inequalities (5.10), (5.11), and (5.12) combine to yield

$$
\begin{aligned}
&P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_{t+1})) \\
&\quad \overset{(a)}{\leq} \min_{\alpha \in [0,1]} \Big\{ P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_\alpha)) + K_m \left\| \theta_\alpha - \theta_t \right\|^2 \\
&\qquad\qquad \Big| \ \theta_\alpha = k\big(\alpha(\zeta \circ \lambda)(\theta^*) + (1-\alpha)(\zeta \circ \lambda)(\theta_t)\big) \Big\} \\
&\quad \overset{(b)}{\leq} \min_{\alpha \in [0,1]} \Big(P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - \alpha P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) \\
&\qquad\qquad - (1-\alpha) P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t)) + K_m \alpha^2 \ell^2 D_\zeta^2 \Big) \\
&\quad = \min_{\alpha \in [0,1]} \Big((1-\alpha)\big(P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))\big) + K_m \alpha^2 \ell^2 D_\zeta^2 \Big), \qquad (5.13)
\end{aligned}
$$

where inequality (a) results from multiplying both sides of (5.10) by -1 and adding $P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*))$, and (5.11) and (5.12) together yield (b).

Using (5.10), (5.11), (5.12), and (5.13), we now analyze the sequence $\{\alpha_t\}_{t \in \mathbb{N}}$ defined in (5.8). We first use (5.13) to derive a useful recursive inequality for $\{\alpha_t\}_{t \in \mathbb{N}}$. Notice that $\alpha_t \geq 0$ for all $t \geq 0$. Now, assume that $\alpha_0 \geq 1$. This implies that $P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_0)) \geq 2K_m \ell^2 D_\zeta^2$, so the minimum in (5.13) is attained when $\alpha = 1$. But then $\alpha_1 \leq 1/2$. Since this argument is independent of the choice of t, we can assume without loss of generality that $\alpha_t \leq 1$, for all $t \geq 0$, by simply discarding α_0 if it is greater than 1.

We next show $\alpha_{t+1} \leq \alpha_t$ for all $t \geq 0$. Since $\alpha_t \leq 1$, α_t is always the minimizer of the right-hand side of (5.13), which can be seen by setting the derivative with respect to α equal to 0 and solving for α. Substituting α_t into (5.13), we see that

(5.14)

$$
\begin{aligned}
&P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_{t+1})) \\
&\quad \leq \left(1 - \frac{P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))}{2K_m \ell^2 D_\zeta^2}\right) \big(P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))\big) \\
&\qquad + \Big(\frac{P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))}{2K_m \ell^2 D_\zeta^2}\Big)^2 K_m \ell^2 D_\zeta^2 \\
&\quad \overset{(a)}{=} \left(1 - \frac{P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))}{2K_m \ell^2 D_\zeta^2}\right) \big(P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))\big) \\
&\qquad + \frac{P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))}{4K_m \ell^2 D_\zeta^2} \big(P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))\big) \\
&\quad \overset{(b)}{=} \left(1 - \frac{P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))}{4K_m \ell^2 D_\zeta^2}\right) \big(P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))\big),
\end{aligned}
$$

where (a) results by noticing that one of the $K_m \ell^2 D_\zeta^2$ terms cancels and (a) can be obtained by factoring out the $P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t))$ term and simplifying. Dividing both sides of (5.14) by $2K_m \ell^2 D_\zeta^2$ shows that $\alpha_{t+1} \leq \alpha_t$.

Dividing both sides of (5.14) by $4K_m \ell^2 D_\zeta^2$ yields the recursive inequality $\frac{\alpha_{t+1}}{2} \leq \big(1 - \frac{\alpha_t}{2}\big)\frac{\alpha_t}{2}$, which implies

$$
\frac{2}{\alpha_{t+1}} \geq \frac{1}{\big(1 - \frac{\alpha_t}{2}\big)\frac{\alpha_t}{2}} = \frac{2\big(1 - \frac{\alpha_t}{2}\big) + \alpha_t}{\big(1 - \frac{\alpha_t}{2}\big)\alpha_t} = \frac{2}{\alpha_t} + \frac{1}{1 - \frac{\alpha_t}{2}} \geq \frac{2}{\alpha_t} + 1 \geq \frac{2}{\alpha_0} + t.
$$

Since $\alpha_0 \leq 1$, this gives us that

$$\frac{\alpha_{t+1}}{2} \leq \frac{1}{t+\frac{2}{\alpha_0}} \leq \frac{1}{t+2}. \tag{5.15}$$

Multiplying both sides by $4K_m\ell^2 D_\zeta^2$ finally yields, for all $t \geq 0$,

$$q(\theta^*) - q(\theta_t) = P_{\widehat{H}}((\zeta \circ \lambda)(\theta^*)) - P_{\widehat{H}}((\zeta \circ \lambda)(\theta_t)) \leq \frac{4K_m\ell^2 D_\zeta^2}{t+1}. \tag{5.16}$$

This establishes the convergence rate result for (5.7). We finish the proof by using this result to derive the corresponding rate for the OIR projected gradient descent scheme (5.5). Since $q(\theta) = 1/\rho(\theta)$ and $K_m = K/m^2$, (5.16) implies that

$$\frac{\rho(\theta_t)-\rho(\theta^*)}{M^2} \leq \frac{\rho(\theta_t)-\rho(\theta^*)}{\rho(\theta^*)\rho(\theta_t)} = \frac{1}{\rho(\theta^*)} - \frac{1}{\rho(\theta_t)} = q(\theta^*) - q(\theta_t) \leq \frac{4K_m\ell^2 D_\zeta^2}{t+1} = \frac{4K\ell^2 D_\zeta^2}{m^2(t+1)}.$$

Since $K = \max\{m^2 L, M^2 m^2 L\}$ and letting $L_1 = \max\{L, M^2 L\}$, we have

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

Coupled with Theorem 5.3, this result provides a nonasymptotic convergence rate to *global optimality* for algorithms solving the OIR minimization problem (5.1).

Remark 5.6. While the dependence on the hyperparameter κ does not appear explicitly in the convergence rate, it does implicitly influence the rate. In particular, as $\kappa \to \infty$, the objective $\rho(\theta)$ becomes arbitrarily close to the constant function with value 0. Therefore, the suboptimality gap converges more quickly as $\kappa \to \infty$, since the possible variation of $\rho(\theta)$ about 0 goes to null as $\rho(\theta)$ gets closer to the constant function with value 0. Therefore, if one multiplies the OIR objective by κ, one obtains the scaled OIR objective $\kappa\rho(\theta) = J(\theta)/\big(1 + \frac{H(d_\theta)}{\kappa}\big) \to J(\theta)$ as $\kappa \to \infty$. However, altering the objective in this way also changes the behavior of the right-hand side of the rate given in inequality (5.6). To see this, notice that we can analyze this situation by applying Theorem 5.5 with $J(\theta)$ in the definition of $\rho(\theta)$ replaced by $\kappa J(\theta)$—i.e., we simply scale our costs by κ. If we recall the definition $M = \max_\theta \rho(\theta)$ from Theorem 5.5, however, then as $\kappa \to \infty$ we have that $M = \max_\theta \kappa J(\theta)/(\kappa + H(d_\theta)) \to \max_\theta J(\theta)$. This means that as $\kappa \to \infty$, its effect on the convergence rate disappears from the right-hand side of inequality (5.6), leaving us with a standard $\mathcal{O}(1/t)$ rate.

Remark 5.7. When compared with the corresponding result in [39], to which it is closely related, the bound (5.6) of Theorem 5.5 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 (5.6) 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

$$M \leq \max_{\theta\in\Theta}\frac{C}{\kappa + H(d_\theta)} = \frac{C}{\kappa + \min_{\theta\in\Theta} H(d_\theta)}. \tag{5.17}$$

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 (5.6). 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 (5.6). This highlights the practical usefulness of the constant 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. Note this assumption is limited to Theorem 5.5, whereas Theorem 5.12 below allows stochastic approximation error and Theorem 5.3 above is independent of estimation issues. Though this assumption is a drawback for Theorem 5.5, we highlight that it allows us to succinctly focus on a core insight of this work: hidden quasiconvexity unlocks an information-dependent convergence rate to global optimality. We also note that, for REINFORCE-like algorithms like those considered in Theorem 5.5, 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 for future work, and we expect it to involve extending the analysis in [38] to the OIR problem.

5.3. Actor-critic convergence. We conclude this section by proving almost sure convergence of IDAC to a neighborhood of a stationary point of (5.1). By Theorem 5.3, this implies IDAC converges almost surely (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 [5, 40, 1]. We analyze the algorithm as given in Algorithm 4.2 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,

$$\theta_{t+1} = \Gamma\left[\theta_t - \beta_t \frac{\delta_t^J(\kappa + \mu_t^H) - \mu_t^J \delta_t^H}{\left(\kappa + \mu_t^H\right)^2} \nabla \log \pi_{\theta_t}(a_t|s_t)\right], \tag{5.18}$$

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., [18, 6, 5]), is for purposes of theoretical analysis and is typically not needed in practice. In addition to Assumption 5.1, we impose the following.

Assumption 5.8. 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 5.9. 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 5.1, 5.8, and 5.9 are standard in two-timescale convergence analyses for actor-critic algorithms [5].

To prepare for the proof of Theorem 5.12, the main result of this section, we first prove Lemmas 5.10 and 5.11. Our analysis leverages the average-reward actor-critic results in [5] as well as the results for ratio optimization actor-critic in [34]. For a given policy parameter θ, let $D_\theta = \mathrm{diag}(d_\theta) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ denote the matrix with the elements of d_θ along the diagonal and zeros everywhere else. Define the state cost vector for the average-cost MDP $(\mathcal{S}, \mathcal{A}, p, c)$ to be $c_\theta = [c_\theta(s)]_{s\in\mathcal{S}}^\top \in \mathbb{R}^{|\mathcal{S}|}$, where $c_\theta(s) = \sum_{a\in\mathcal{A}} \pi_\theta(a|s) c(s,a)$. Similarly, let $r_\theta = [-\log d_\theta(s)]_{s\in\mathcal{S}}^\top \in \mathbb{R}^{|\mathcal{S}|}$ denote the state reward vector for the shadow MDP $(\mathcal{S},\mathcal{A},p,r)$, where $r(s,a)=-\log d_\theta(s)$. Note that the ergodicity condition of Assumption 5.1 implies that $d_\theta(s)>0$ for all $s\in\mathcal{S},\theta\in\Theta$, so $r(s,a)$ is always defined and finite. Finally, let $P_\theta\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote the state transition probability matrix under policy π_θ, i.e., $P_\theta(s'|s)=\sum_{a\in\mathcal{A}}\pi_\theta(a|s)p(s'|s,a)$ for any $s,s'\in\mathcal{S}$. We first show convergence of the critics.

LEMMA 5.10. *Under Assumption* 5.9*, given a fixed policy parameter* $\theta\in\Theta$*, the critic updates in lines* 4, 6, 13, 14 *of Algorithm* 4.2 *converge as follows:* $\lim_{t\to\infty}\mu_t^J=J(\theta)$ *a.s.,* $\lim_{t\to\infty}\mu_t^H=H(d_\theta)$ *a.s.,* $\lim_{t\to\infty}\omega_t^J=\omega_\theta^J$ *a.s., and* $\lim_{t\to\infty}\omega_t^H=\omega_\theta^H$ *a.s., where* ω_θ^J *and* ω_θ^H *are, respectively, the unique solutions to*

$$\Phi^\top D_\theta\big[c_\theta-J(\theta)\cdot\mathbf{1}+P_\theta(\Phi\omega^J)-\Phi\omega^J\big]=\mathbf{0},$$
$$\Phi^\top D_\theta\big[r_\theta-H(d_\theta)\cdot\mathbf{1}+P_\theta(\Phi\omega^H)-\Phi\omega^H\big]=\mathbf{0}.$$

Proof. Since the policy π_θ is held fixed and the shadow MDP reward $-\log d_\theta(s)$ can be exactly evaluated for any $s\in\mathcal{S}$, the proof of Lemma 4 in [5] can be applied separately to the average-cost recursions in lines 4, 9, and 13 and the shadow MDP recursions in lines 6, 10, and 14 of Algorithm 4.2 to obtain the result. ∎

As in Lemma 5 of [5], this result shows that the sequences $\{\omega_t^J\}$ and $\{\omega_t^H\}$ converge a.s. to the limit points ω_θ^J and ω_θ^H of the TD(0) algorithm with linear function approximation for their respective MDPs. Due to the use of linear function approximation, when used in the policy update step the value function estimates $v_\theta^J=\Phi\omega_\theta^J$ and $v_\theta^H=\Phi\omega_\theta^H$ may result in biased gradient estimates. Similar to the bias characterization given in Lemma 4 in [5], this bias can be characterized as follows.

LEMMA 5.11. *Fix* $\theta\in\Theta$*. Let* $\delta_t^{\theta,J}=c(s_t,a_t)-J(\theta)+\phi(s_{t+1})^\top\omega_\theta^J-\phi(s_t)^\top\omega_\theta^J$ *and* $\delta_t^{\theta,H}=-\log d_\theta(s_t)-H(d_\theta)+\phi(s_{t+1})^\top\omega_\theta^H-\phi(s_t)^\top\omega_\theta^H$ *denote the stationary estimates of the TD errors at time t. Let* $\overline{v}_\theta^J=\mathbb{E}_{\pi_\theta}[c(s,a)-J(\theta)+\phi(s')^\top\omega_\theta^J]$ *and* $\overline{v}_\theta^H=\mathbb{E}_{\pi_\theta}[-\log d_\theta(s)-H(d_\theta)+\phi(s')^\top\omega_\theta^H]$*. Now let* $\epsilon_\theta^J=\sum_{s\in\mathcal{S}}d_\theta(s)\left[\nabla_\theta\overline{v}_\theta^J(s)-\nabla_\theta\phi(s)^\top\omega_\theta^J\right]$ *and* $\epsilon_\theta^H=\sum_{s\in\mathcal{S}}d_\theta(s)\left[\nabla_\theta\overline{v}_\theta^H(s)-\nabla_\theta\phi(s)^\top\omega_\theta^H\right]$*. We then have that*

$$\mathbb{E}_{\pi_\theta}\left[\frac{\delta_t^{\theta,J}\left[\kappa+H(d_\theta)\right]-J(\theta)\delta_t^{\theta,H}}{\left[\kappa+H(d_\theta)\right]^2}\nabla\log\pi_\theta(a_t|s_t)\right]=\nabla\rho(\theta)+\frac{\epsilon_\theta^J\left[\kappa+H(d_\theta)\right]-J(\theta)\epsilon_\theta^H}{\left[\kappa+H(d_\theta)\right]^2}.$$

Proof. By [5, Lemma 4] and Theorem 3.10, $\mathbb{E}_{\pi_\theta}\left[\delta_t^{\theta,J}\nabla\log\pi_\theta(a_t|s_t)\right]=\nabla J(\theta)+\epsilon_\theta^J$ and $\mathbb{E}_{\pi_\theta}\left[\delta_t^{\theta,H}\nabla\log\pi_\theta(a_t|s_t)\right]=\nabla H(d_\theta)+\epsilon_\theta^H$. This implies that

$$\begin{aligned}&\mathbb{E}_{\pi_\theta}\left[\frac{\delta_t^{\theta,J}\left[\kappa+H(d_\theta)\right]-J(\theta)\delta_t^{\theta,H}}{\left[\kappa+H(d_\theta)\right]^2}\nabla\log\pi_\theta(a_t|s_t)\right]\\&\quad=\frac{\left[\kappa+H(d_\theta)\right]\mathbb{E}_{\pi_\theta}\left[\delta_t^{\theta,J}\nabla\log\pi_\theta(a_t|s_t)\right]-J(\theta)\mathbb{E}_{\pi_\theta}\left[\delta_t^{\theta,H}\nabla\log\pi_\theta(a_t|s_t)\right]}{\left[\kappa+H(d_\theta)\right]^2}\\&\quad=\frac{\left[\kappa+H(d_\theta)\right]\left(\nabla J(\theta)+\epsilon_\theta^J\right)-J(\theta)\left(\nabla H(d_\theta)+\epsilon_\theta^H\right)}{\left[\kappa+H(d_\theta)\right]^2}\\&\quad=\nabla\rho(\theta)+\frac{\epsilon_\theta^J\left[\kappa+H(d_\theta)\right]-J(\theta)\epsilon_\theta^H}{\left[\kappa+H(d_\theta)\right]^2},\end{aligned}$$

which completes the proof. ∎

We now establish convergence of the actor step, and thus the actor-critic algorithm. Given any continuous function $f:\Theta\to\mathbb{R}^d$, define the function $\hat{\Gamma}(\cdot)$ using the projection operator Γ to be $\hat{\Gamma}(f(\theta))=\lim_{\eta\to 0^+}\left[\Gamma(\theta+\eta\cdot f(\theta))-\theta\right]/\eta$. Define

$$\epsilon_\theta=\frac{\epsilon_\theta^J\left[\kappa+H(d_\theta)\right]-J(\theta)\epsilon_\theta^H}{\left[\kappa+H(d_\theta)\right]^2}. \tag{5.19}$$

Consider the ODEs

$$\dot{\theta}=\hat{\Gamma}(\nabla\rho(\theta)), \tag{5.20}$$

$$\dot{\theta}=\hat{\Gamma}(\nabla\rho(\theta)+\epsilon_\theta). \tag{5.21}$$

Notice that, by the definition of $\hat{\Gamma}$, the right-hand side of (5.20) is simply $\Gamma(\nabla\rho(\theta))$ when there exists $\eta_0>0$ such that $\theta+\eta\nabla\rho(\theta)\in\Theta$ for all $\eta<\eta_0$. When such an η_0 does not exist, $\hat{\Gamma}(\nabla\rho(\theta))$ 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. Similar statements hold for (5.21). For further discussion of the definition of $\hat{\Gamma}$ and related results, see [17, p. 191]. For the projected ODE interpretation, see section 4.3 of [18].

We now present the main result of this subsection, which establishes convergence of the actor-critic algorithm. Its proof follows that of Theorem 1 in [5], with key modifications to accommodate complications arising from the fact that the objective to be minimized is a ratio; specifically, we ensure that (i) the resulting noise terms are indeed asymptotically negligible, and (ii) the Lipschitz properties of the gradient $\nabla\rho(\theta)$ necessary for the ODE analysis are satisfied.

THEOREM 5.12. *Let $\mathcal{Z}$ denote the set of asymptotically stable equilibria of the ODE* (5.20). *Given any $\varepsilon>0$, define $\mathcal{Z}^\varepsilon=\{z\mid\inf_{z'\in\mathcal{Z}}\|z-z'\|\le\varepsilon\}$. For any $\theta\in\Theta$, let ε_θ be defined as in* (5.19). *Under Assumptions* 5.1, 5.8, *and* 5.9, *given any $\varepsilon>0$, there exists $\delta>0$ such that for $\{\theta_t\}$ obtained from Algorithm* 4.2 *with projection* (5.18), *if $\sup_t\|\epsilon_{\theta_t}\|<\delta$, then $\theta_t\to\mathcal{Z}^\varepsilon$ a.s. as $t\to\infty$.*

Proof. Let $\mathcal{F}_t=\sigma(\theta_k,k\le t)$ denote the σ-algebra generated by the θ-iterates up to time t. Define $\delta_t=\frac{\delta_t^J\left[\kappa+\mu_t^H\right]-\mu_t^J\delta_t^H}{\left[\kappa+\mu_t^H\right]^2}$ and $\delta_t^\theta=\frac{\delta_t^{J,\theta}[\kappa+H(d_\theta)]-J(\theta)\delta_t^{H,\theta}}{[\kappa+H(d_\theta)]^2}$. In addition, define the noise terms $M_t^{(1)}=\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)-\mathbb{E}\left[\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)\mid\mathcal{F}_t\right]$ and $M_t^{(2)}=\mathbb{E}[(\delta_t-\delta_t^{\theta_t})\nabla\log\pi_{\theta_t}(a_t|s_t)\mid\mathcal{F}_t]$. Finally, define the function $h(\theta_t)=\mathbb{E}_{\pi_{\theta_t}}[\delta_t^{\theta_t}\nabla\log\pi_{\theta_t}(a_t|s_t)\mid\mathcal{F}_t]=\mathbb{E}_{\pi_{\theta_t}}[\delta_t^{\theta_t}\nabla\log\pi_{\theta_t}(a_t|s_t)]$, which is the gradient expression from Lemma 5.11. Note that simultaneously taking an expectation with respect to π_{θ_t} and conditioning on $\mathcal{F}_t$ is redundant, so we can suppress one or the other in our notation without altering the meaning.

We can now rewrite the projected actor update (5.18) as

$$\theta_{t+1}=\Gamma\Big(\theta_t-\beta_t\delta_t\nabla\log\pi_{\theta_t}(a_t|s_t)\Big)=\Gamma\left(\theta_t-\beta_t\left[h(\theta_t)+M_t^{(1)}+M_t^{(2)}\right]\right).$$

We show that this update scheme asymptotically tracks the ODE (5.21) a.s. by demonstrating that the noise terms $\{M_t^{(1)}\}$ form an almost surely bounded martingale difference sequence, that the terms $\{M_t^{(2)}\}$ are asymptotically negligible, and that h is Lipschitz and thus the ODE is well-posed.

Since $\delta_t\to\delta_t^{\theta_t}$ a.s. by Lemma 5.10, we have that $M_t^{(2)}\to 0$ a.s., so the noise terms $\{M_t^{(2)}\}$ are indeed asymptotically negligible. Next, recall the tower property of conditional expectations: for any $\mathcal{F}$-measurable random variable X and any sub-σ-algebras $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$, we have $\mathbb{E}\left[\mathbb{E}\left[X|\mathcal{G}\right]|\mathcal{H}\right] = \mathbb{E}\left[\mathbb{E}\left[X|\mathcal{H}\right]|\mathcal{G}\right] = \mathbb{E}\left[X|\mathcal{G}\right]$. Since $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ for all $t \geq 0$, this implies that for all $t \geq 0$,

$$\begin{aligned}
\mathbb{E}&\left[M_{t+1}^{(1)}|\mathcal{F}_t\right] \\
&= \mathbb{E}\left[\delta_{t+1}\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}) - \mathbb{E}\left[\delta_{t+1}\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}) \mid \mathcal{F}_{t+1}\right] \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\delta_{t+1}\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}) \mid \mathcal{F}_t\right] \\
&\quad - \mathbb{E}\left[\mathbb{E}\left[\delta_{t+1}\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}) \mid \mathcal{F}_{t+1}\right] \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\delta_{t+1}\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}) \mid \mathcal{F}_t\right] - \mathbb{E}\left[\delta_{t+1}\nabla \log \pi_{\theta_{t+1}}(a_{t+1}|s_{t+1}) \mid \mathcal{F}_t\right] = 0,
\end{aligned}$$

so $\{M_t^{(1)}\}$ is an $\mathcal{F}$-martingale difference sequence, where $\mathcal{F}$ is the filtration $\mathcal{F} = \{\mathcal{F}_t\}$.

To see that $M_t^{(1)}$ is a.s. bounded, first notice that $\mu_0^H > 0$ and $0 < \inf_t d_{\theta_t}(s_t) \leq \sup_t d_{\theta_t}(s_t) \leq 1$, so $\{\mu_t^H\}$ is uniformly bounded both above and below away from zero. A similar argument applies to $\{\mu_t^J\}$. Coupled with almost sure boundedness of $\{\omega_t^J\}$ and $\{\omega_t^H\}$, this implies that $\{\delta_t\}$ and thus $\{M_t^{(1)}\}$ are a.s. bounded. As discussed in sections 2.1–2.2 of [6], the facts that $\{M_t^{(1)}\}$ is a.s. bounded martingale difference noise and $\{M_t^{(2)}\}$ is asymptotically negligible ensure that as long as the right-hand side of the ODE (5.21) is Lipschitz, the iterates generated by (5.18) will asymptotically track it.

To see that h is Lipschitz in θ, first rewrite

$$\begin{aligned}
h(\theta_t) = \frac{1}{[\kappa + H(d_{\theta_t})]^2}\Big(&[\kappa + H(d_{\theta_t})]\, \mathbb{E}_{\pi_{\theta_t}}\left[\delta_t^{J,\theta_t}\nabla \log \pi_{\theta_t}(a_t|s_t)\right] \\
&- J(\theta_t)\mathbb{E}_{\pi_{\theta_t}}\left[\delta_t^{H,\theta_t}\nabla \log \pi_{\theta_t}(a_t|s_t)\right]\Big).
\end{aligned}$$

We verify that each of the component terms in this expression is Lipschitz and bounded. Recall that a function is Lipschitz if it is continuously differentiable with bounded derivatives. First, as discussed in the proof of [5, Lem. 5], $J(\theta)$, $d_\theta(s)$, $\nabla\pi_\theta(a|s)$, and $\Phi\omega_\theta^J$ are all Lipschitz and bounded for all $s \in \mathcal{S}, a \in \mathcal{A}$. Thus $J(\theta)$ and $\mathbb{E}_{\pi_\theta}\left[\delta_t^{J,\theta}\nabla \log \pi_\theta(a_t|s_t)\right]$ are Lipschitz and bounded on Θ. The remaining terms we need to inspect are $\kappa + H(d_\theta)$, $1/\left[\kappa + H(d_\theta)\right]^2$, and $\mathbb{E}_{\pi_\theta}\left[\delta_t^{H,\theta}\nabla \log \pi_\theta(a_t|s_t)\right]$.

Theorem 3.10 implies $\nabla H(d_\theta)$ is continuous and bounded. To see this, notice

$$\begin{aligned}
\nabla H(d_\theta) &= \mathbb{E}_{\pi_\theta}\Big[\left(-\log d_\theta(s) - H(d_\theta)\right)\nabla \log \pi_\theta(a|s)\Big] \\
&= \sum_s d_\theta(s)\sum_a \pi_\theta(a|s)\Big[\left(-\log d_\theta(s) - H(d_\theta)\right)\nabla \log \pi_\theta(a|s)\Big] \\
&= \sum_s d_\theta(s)\sum_a \nabla\pi_\theta(a|s)\Big[\left(-\log d_\theta(s) - H(d_\theta)\right)\Big].
\end{aligned}$$

By the ergodicity condition of Assumption 5.1, we have that $d_\theta(s) > 0$ for all $s \in \mathcal{S}$, which means that the $-\log d_\theta(s)$ is always defined. Since d_θ is continuous, we furthermore have that $-\log d_\theta(s)$ and $H(d_\theta)$ are both continuous. The gradient $\nabla\pi_\theta(a|s)$ is continuous by Assumption 5.1. Finally, since Θ is a compact set, we know that $d_\theta(s), \nabla\pi_\theta(a|s), -\log d_\theta(s)$, and $H(d_\theta)$ remain bounded, implying that $\nabla H(d_\theta)$ is continuous and bounded, since it is formed by taking products and sums of continuous, bounded functions. $H(d_\theta)$ is thus Lipschitz and bounded, as is the term $\kappa + H(d_\theta)$, for any constant $\kappa \geq 0$. Furthermore, since $d_\theta(s) > 0$ for all $s \in \mathcal{S}$, and since Θ is compact, there exists some constant B such that $\inf_{\theta\in\Theta} H(d_\theta) = B > 0$. This means that $1/\left[\kappa + H(d_\theta)\right]^2 \leq 1/\left[\kappa + B\right]^2$ for all $\theta \in \Theta$. The term $1/\left[\kappa + H(d_\theta)\right]^2$ is therefore Lipschitz and bounded, as well. Finally, notice that $\mathbb{E}_{\pi_\theta}\left[\delta_t^{H,\theta}\nabla\log\pi_\theta(a_t|s_t)\right] =$

(5.22)

$$\begin{aligned}&\mathbb{E}_{\pi_\theta}\left[\left(-\log d_\theta(s_t) - H(d_\theta) + \phi(s_{t+1})^\top\omega_\theta^H - \phi(s_t)^\top\omega_\theta^H\right)\nabla\log\pi_\theta(a_t|s_t)\right]\\ &= \sum_s d_\theta(s)\sum_a \nabla\pi_\theta(a|s)\left[-\log d_\theta(s) - H(d_\theta) - \phi(s)^\top\omega_\theta^H + \sum_{s'} p(s'|s,a)\phi(s')^\top\omega_\theta^H\right].\end{aligned}$$

As discussed above, $d_\theta(s), \pi_\theta(a|s), -\log d_\theta(s)$, and $H(d_\theta)$ are all continuously differentiable with bounded derivatives on Θ. Furthermore, given that $H(d_\theta)$ is Lipschitz and bounded both above and away from zero, $\Phi\omega_\theta^H$ is Lipschitz and bounded for reasons analogous to those for $\Phi\omega_\theta^J$. Expression (5.22) is thus Lipschitz and bounded.

By the foregoing, h is Lipschitz, since it is formed by taking products and sums of Lipschitz, bounded functions. The ODE (5.21) is therefore well-posed, and its equilibrium set $\mathcal{Z}$ is well-defined. A similar argument to the one just presented can be used to show that (5.20) with equilibrium set $\mathcal{Y}$ is also well-posed. The remainder of the arguments in the proof of Lemma 5 in [5] now apply to prove that $\theta_t \to \mathcal{Y}$ a.s. as $t \to \infty$, and that as $\sup_\theta \|\epsilon_\theta\| \to 0$, the trajectories of (5.21) converge to those of (5.20). In particular, this implies that for a given $\varepsilon > 0$, there exists a $\delta > 0$ such that if $\sup_\theta \|\epsilon_\theta\| < \delta$, then $\theta_t \to \mathcal{Z}^\varepsilon$ a.s. as $t \to \infty$. □

Combined with Theorem 5.3, Theorem 5.12 establishes almost sure convergence of IDAC to a neighborhood of a *global* optimum of the OIR minimization problem (5.1). Note that if the features Φ are expressive enough, ε will be small or even zero.

6. Experiments. We conducted numerical experiments showing that when the reward signal is sparse, OIR methods lead to improved performance when compared with vanilla RL methods. We conducted 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 AC on three relatively small gridworlds. For the second set of experiments, we compared a neural network version of IDAC with the A2C, DQN, and PPO algorithms on a larger, more complex gridworld. Due to the well-known practical advantages of actor-critic methods over higher-variance REINFORCE-based approaches [35], we focused on IDAC in our experiments. All the environments emit sparse reward signals in the sense that the majority of costs convey no information about the central task of finding the goal state. On all environments, OIR policy gradient methods outperform the vanilla RL methods that we tested. Due to space limitations, we only present the neural network experiments in this section. See [33] for all experimental results.

6.1. Implementation. A 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 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 $a \in \{\text{stay, up, down, left, right}\}$. 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. 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 given in (6.1),

$$c(s,a) = \begin{cases} c_{\text{goal}} & \text{if } s = s_{\text{goal}} \text{ and } a \in \mathcal{A}(s), \\ c_{\text{allowed}} & \text{if } s \neq s_{\text{goal}} \text{ and } a \in \mathcal{A}(s), \\ c_{\text{blocked}} & \text{if } a \notin \mathcal{A}(s), \end{cases} \tag{6.1}$$

where $0 < c_{\text{goal}} < c_{\text{allowed}} < c_{\text{blocked}}$.

For the neural network experiments, we implemented IDAC with a categorical policy using two-layer, fully connected neural networks for both policy and value function approximators, and we compared against the Stable Baselines3 [28] implementations of A2C, DQN, and PPO with two-layer, fully connected neural networks for all policies and value function approximators.

6.2. Neural network experiment results. Figure 1 illustrates the performance of neural IDAC and A2C, DQN, and PPO on the gridworld environment pictured with $c_{\text{goal}} = 0.1$, $c_{\text{allowed}} = 5$, and $c_{\text{blocked}} = 10$. To generate the data for these figures, we trained 48 instances of neural IDAC and 15 instances of each of the A2C, DQN, and PPO algorithms. Average costs were computed for each episode, and the sample means and 95% confidence intervals were used to create the learning curves. As the figure illustrates, IDAC outperformed all three. Furthermore, none of A2C, DQN, and PPO found the goal state after 1.2e+5 time steps. Hyperparameters $\alpha = 0.0001$, $\beta = 0.0002$, $\tau = 0.1$, $\kappa = 0.1$, and 512 hidden units for each layer in both the policy and value functions for neural IDAC were selected through trial and error. After experimenting with a range of different parameters and detecting no noticeable difference in performance, Stable Baselines3 default parameters for A2C, DQN, and PPO were used. This included learning rates 0.0007 for A2C, 0.0003 for PPO, and 0.0001 for DQN, as well as 64-width layers for all networks.

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

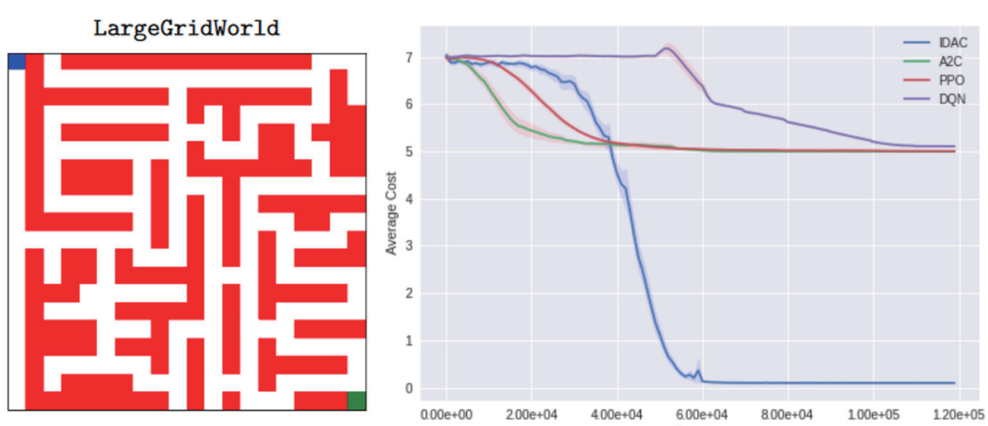

Fig. 1. *Comparison of neural network IDAC with common deep RL methods. Plot gives means and* 95% *confidence intervals. Optimal average cost is* 0.1. *Training took place over* 1e+6 *time steps; no further improvement occurred beyond time step* 1.2e+5.

cost is 0.1, they converge to and remain at these suboptimal values for the remainder of training. Sparse rewards and overconfidence in past experience likely caused this premature convergence. Meanwhile, since neural IDAC minimizes $\rho(\theta)$ instead of $J(\theta)$, it swiftly locates the goal state and 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.

7. Conclusion. In this paper we have developed policy gradient methods for a new RL objective, the OIR. En route, we have elaborated a rich theory underlying these methods, including a concave programming reformulation of the OIR optimization problem with links to the powerful linear programming theory for MDPs; policy gradient theorems for the OIR setting; and both asymptotic and nonasymptotic convergence theory with global optimality guarantees under appropriate assumptions. We have furthermore presented empirical results that indicate promising performance compared with state-of-the-art methods on sparse-reward problems. Interesting directions for future work include extensions to more general classes of ratio optimization problems, development of variants of the IDAC algorithm for continuous spaces using suitable density estimation techniques, exploration of whether the OIR enables faster-than-linear nonasymptotic rate analyses, and thorough empirical evaluation of deep RL variants of IDAC on a range of benchmark problems.

Disclaimer. This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JP Morgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

REFERENCES

[1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, *Optimality and approximation with policy gradient methods in Markov decision processes*, in Proceedings of the Thirty Third Conference on Learning Theory, 2020, pp. 64–66.
[2] M. Avriel, W. E. Diewert, S. Schaible, and I. Zang, *Generalized Concavity*, SIAM, Philadelphia, 2010, https://doi.org/10.1137/1.9780898719437.
[3] 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*, J. Mach. Learn. Res., 25 (2024), pp. 1–58.
[4] J. Bhandari and D. Russo, *Global optimality guarantees for policy gradient methods*, Oper. Res., 72 (2024), pp.1906–1927.
[5] S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee, *Natural actor-critic algorithms*, Automatica, 45 (2009), pp. 2471–2482.
[6] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*, Cambridge University Press, 2008.
[7] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
[8] R. M. Gray, *Entropy and Information Theory*, Springer Science & Business Media, 2011.
[9] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, *Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor*, in Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 1861–1870.

[10] E. Hazan, S. Kakade, K. Singh, and A. Van Soest, *Provably efficient maximum entropy exploration*, in Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 2681–2691.

[11] N. Karmarkar, *A new polynomial-time algorithm for linear programming*, in Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing, 1984, pp. 302–311.

[12] L. G. Khachiyan, *A polynomial algorithm in linear programming*, Dokl. Akad. Nauk SSSR, 244, 1979, pp. 1093–1096.

[13] K. Kim, A. Jindal, Y. Song, J. Song, Y. Sui, and S. Ermon, *Imitation with neural density models*, in Advances in Neural Information Processing Systems 34, 2021, pp. 5360–5372.

[14] J. Kirschner, T. Lattimore, and A. Krause, *Information directed sampling for linear partial monitoring*, in Proceedings of the 33rd Conference on Learning Theory, 2020, pp. 2328–2369.

[15] J. Kirschner, T. Lattimore, C. Vernade, and C. Szepesvári, *Asymptotically optimal information-directed sampling*, in Proceedings of the 34th Conference on Learning Theory, 2021, pp. 2777–2821.

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

[17] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer Science & Business Media, 1978.

[18] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Stoch. Model. Appl. Probab., Springer-Verlag, New York, 2003.

[19] L. Lee, B. Eysenbach, E. Parisotto, E. Xing, S. Levine, and R. Salakhutdinov, *Efficient Exploration via State Marginal Matching*, preprint, arXiv:1906.05274, 2019.

[20] 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*, preprint, arXiv:1509.02971, 2015.

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

[22] X. Lu, B. Van Roy, V. Dwaracherla, M. Ibrahimi, I. Osband, and Z. Wen, *Reinforcement learning, bit by bit*, Found. Trends Mach. Learn., 16 (2023), pp. 733–865.

[23] J. Mei, C. Xiao, C. Szepesvári, and D. Schuurmans, *On the global convergence rates of softmax policy gradient methods*, in Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 6820–6829.

[24] M. Mutti, L. Pratissoli, and M. Restelli, *Task-agnostic exploration via policy gradient of a non-parametric state entropy estimate*, in Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, 2021, pp. 9028–9036.

[25] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Springer Science & Business Media, 2003.

[26] N. Nikolov, J. Kirschner, F. Berkenkamp, and A. Krause, *Information-directed exploration for deep reinforcement learning*, in Proceedings of 7th International Conference on Learning Representations, 2019.

[27] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, 2014.

[28] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, and N. Dormann, *Stable Baselines*3, GitHub repository, 2019, https://stable-baselines3.readthedocs.io/en/master/.

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

[30] D. Russo and B. Van Roy, *An information-theoretic analysis of Thompson sampling*, J. Mach. Learn. Res., 17 (2016), pp. 2442–2471.

[31] S. Schaible, *Fractional programming.* I. *Duality*, Manage. Sci., 22 (1976), pp. 858–867.

[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal Policy Optimization Algorithms*, preprint, arXiv:1707.06347, 2017.

[33] W. A. Suttle, A. Koppel, and J. Liu, *Occupancy Information Ratio: Infinite-Horizon, Information-Directed, Parameterized Policy Search*, preprint, https://arxiv.org/abs/2201.08832, 2023.

[34] W. A. Suttle, K. Zhang, Z. Yang, D. Kraemer, and J. Liu, *Reinforcement learning for cost-aware Markov decision processes*, in Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 9989–9999.

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

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

[37] R. J. Williams, *Simple statistical gradient-following algorithms for connectionist reinforcement learning*, Mach. Learn., 8 (1992), pp. 229–256.

[38] J. Zhang, A. S. Bedi, M. Wang, and A. Koppel, *Beyond cumulative returns via reinforcement learning over state-action occupancy measures*, in Proceedings of the 2021 American Control Conference, 2021, pp. 894–901.

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

[40] K. Zhang, A. Koppel, H. Zhu, and T. Başar, *Global convergence of policy gradient methods to (almost) locally optimal policies*, SIAM J. Control Optim., 58 (2020), pp. 3586–3612, https://doi.org/10.1137/19M1288012.