Policy Gradient for Ratio Optimization: A Case Study*

Wesley A. Suttle
Applied Mathematics and Statistics
Stony Brook University
wesley.suttle@stonybrook.edu

Alec Koppel
Optimal Sourcing Systems
Supply Chain Optimization Technologies, Amazon
aekoppel@amazon.com

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

***Abstract*—We consider policy gradient methods for ratio optimization problems by way of an illustrative case study: maximizing the Omega ratio of a financial portfolio. We propose a general framework for ratio optimization in sequential decision-making problems, explore the notion of hidden quasiconcavity in such problems, and propose an actor-critic algorithm for the Omega ratio problem. Our central contribution is to show that the algorithm converges almost surely to (a neighborhood of) a global optimum and to demonstrate its performance in practice.**

***Index Terms*—reinforcement learning, portfolio optimization, quasiconcave programming**

I. Introduction

Ratio optimization problems arise in a variety of sequential decision-making settings: optimizing risk-return trade-offs in financial portfolio management, minimizing price-performance ratios in engineering and economics, and maximizing network bandwidth in computer network design. Reinforcement learning (RL) has seen immense growth in recent years, with applications to a wide array of decision-making problems, yet RL for ratio optimization problems remains largely unexplored. While naïve policy gradient methods can be applied to ratio optimization problems when the corresponding gradient expressions are tractable, this shallow view neglects a rich underlying structure: many of these problems enjoy a certain *hidden quasiconcavity* property that links them to the powerful linear programming theory for Markov decision processes and enables us to prove convergence of corresponding policy gradient algorithms to *global optima*.

In this work, we illuminate this rich structure through a specific example: maximizing the Omega ratio of a financial portfolio. First, we propose a general framework for sequential ratio optimization problems, the *Markov ratio optimization process* (MROP). Next, we formulate the problem of maximizing the Omega ratio of a financial portfolio as an MROP and develop an actor-critic algorithm for solving this problem. Our main theoretical contribution is to employ hidden quasiconcavity of the Omega ratio MROP to establish a global convergence guarantee for our algorithm. We then illustrate this methodology on a portfolio optimization problem.

*All proofs of the assertions in this paper are omitted due to space limitations, but are available upon request and will appear in a forthcoming, expanded version of this paper. The research of W. Suttle was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-22-2-0003.

A. Related Work

RL is a branch of machine learning devoted to solving sequential decision-making problems; see [1] for an overview. Particularly important for the present work are the policy gradient and actor-critic methods [2], [3], which optimize a performance criterion over a parameterized policy class. Asymptotic convergence of actor-critic methods was established early on in [3], [4], yet their non-asymptotic convergence has been studied only recently [5], [6]. Global optimality has been established for objectives defined as concave forms of the long-term occupancy measure of a policy [7]–[9]; these works leverage a notion called *hidden concavity* to guarantee convergence to global optima. We generalize this notion of hidden concavity to *hidden quasiconcavity* and combine it with a two-timescale analysis to provide an analogous result for our ratio optimization algorithm.

The RL literature concerning ratio optimization problems is sparse. Unlike the standard RL setting, where the objective function belongs to an underlying Markov decision process (MDP), there is no single, underlying MDP for such problems. Since many RL techniques are based on the rich, well-understood structure of MDPs, these connections get severed when we move to the ratio setting. The state- and action-value functions that are critical to Q-learning and actor-critic methods have no clear analogues, and the ratio structure can interfere with the Lipschitz properties that are essential to convergence analyses of policy gradient methods. Nonetheless, examples of RL techniques for solving ratio optimization problems based on connections to *generalizations* of MDPs exist in the literature. MDPs with fractional costs, also known as cost-aware MDPs (CAMDPs), and associated RL algorithms were first considered in [10]. More recently, [11] proposed two new RL algorithms with convergence guarantees for solving CAMDPs. The MROP framework that we propose in this paper contains CAMDPs as a special case, and our convergence results can be applied to the actor-critic algorithm considered in [11].

We develop these approaches in the context of portfolio optimization, which studies optimal rebalancing strategies for managing financial portfolios [12], [13]. Many of the most widely-used objective functions in portfolio optimization are ratios, including the Sharpe, Calmar, Sortino, and Omega ratios [14]–[17]. Unlike other performance ratios, which typically only consider lower-order moments of the returns distri-

bution, the Omega ratio incorporates all moments and may be a more appropriate objective for complex problems. Though RL-based approaches to portfolio optimization are not new [18]–[21], to our knowledge, RL techniques for maximizing the Omega ratio remain unexplored.

II. Problem Formulation

In this section we propose a general framework for modeling sequential decision making problems where the objective is to optimize a ratio of two performance measures. This framework, the *Markov ratio optimization process* (MROP), is a generalization of the familiar MDP. We then formulate the Omega ratio portfolio optimization problem as an MROP, laying the groundwork for the Omega ratio actor-critic algorithm of the following section.

A. Markov Decision Processes

Consider an average-reward MDP $(\mathcal{S}, \mathcal{A}, p, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p : \mathcal{S} \times \mathcal{A} \to \mathcal{D}(\mathcal{S})$ is the transition probability kernel mapping state-action pairs to distributions over the state space, and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. At each timestep t, the agent is in state s_t, chooses an action a_t according to some policy $\pi : \mathcal{S} \to \mathcal{D}(\mathcal{A})$ mapping states to distributions over the action set, and incurs a corresponding reward $r(s_t, a_t)$. The system then transitions into a new state $s_{t+1} \sim p(\cdot|s_t, a_t)$. Note that, if we ignore the reward function r, we are left with a *controlled Markov chain* $(\mathcal{S}, \mathcal{A}, p)$. This is important in the definition of an MROP in the following section.

Given a policy π, let $d_\pi \in \mathcal{D}(\mathcal{S})$ denote the steady-state occupancy measure over $\mathcal{S}$ induced by π. In addition, let $\lambda_\pi \in \mathcal{D}(\mathcal{S} \times \mathcal{A})$ denote the state-action occupancy measure induced by π over $\mathcal{S} \times \mathcal{A}$. Note that $\lambda_\pi(s, a) = d_\pi(s)\pi(a|s)$. Finally, let $J(\pi) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_\pi \left[\sum_{i=1}^n r(s_i, a_i)\right] = \int_{\mathcal{S}} \int_{\mathcal{A}} r(s, a)\pi(a|s)d_\pi(s)\ da\ ds$ denote the long-run average reward of using policy π.

B. Markov Ratio Optimization Processes

Fix a controlled Markov chain $(\mathcal{S}, \mathcal{A}, p)$. Given two functions $f, g : \mathcal{D}(\mathcal{S} \times \mathcal{A}) \to \mathbb{R}$, consider the problem of finding a policy π maximizing the objective function

$$\frac{f(\lambda_\pi)}{g(\lambda_\pi)} \tag{1}$$

over the controlled Markov chain $(\mathcal{S}, \mathcal{A}, p)$. This leads us to the following generalization of the classic MDP:

Definition 1. A *Markov ratio optimization process* (MROP), given by $(\mathcal{S}, \mathcal{A}, p, f, g)$, is a discrete-time stochastic decision-making process where the goal is to find a policy π maximizing the ratio $f(\lambda_\pi)/g(\lambda_\pi)$ over the controlled Markov chain $(\mathcal{S}, \mathcal{A}, p)$.

This MROP definition subsumes the average-reward MDP (see, e.g., [22]) and the cost-aware MDP defined in [11] as special cases. We now turn our attention to a particular case.

C. Omega Ratio

A pervasive trend in the financial literature is to formulate measures of total performance for sequential decision-making problems as ratios of some combination of risk and reward functions. Let $\tau \in \mathbb{R}$ be a given real number. One useful such example is the *Sharpe ratio*, defined by

$$\mathrm{Sh}(R; \tau) = \frac{\mathbb{E}[R - \tau]}{\sqrt{\mathrm{Var}(R - \tau)}}, \tag{2}$$

which is a classical case of a measure of risk-adjusted returns. Here, τ is a target return, and the ratio rewards returns that exceed the target while punishing high-volatility returns. A potential weakness of the Sharpe ratio is its emphasis on the first and second moments of the portfolio returns distribution, which works best when the distribution is roughly normal but suffers against skewed or multi-modal distributions.

Letting the cumulative distribution function for the portfolio return R be denoted F_R, the *Omega ratio* [17] is given by

$$\Omega(R; \tau) = \frac{\int_\tau^\infty [1 - F_R(r)]\ dr}{\int_{-\infty}^\tau F_R(r)\ dr}. \tag{3}$$

The Omega ratio can be interpreted as the ratio of the expected excess (above threshold τ) returns to the expected shortfall (below threshold τ) returns of the portfolio. A distinct advantage of the Omega ratio over financial measures such as the Sharpe ratio is that the Omega ratio incorporates information about all moments of R.

D. Omega Ratio Maximization as an MROP

We now formulate optimization of the Omega ratio as an MROP. We begin with an average-reward MDP $(\mathcal{S}, \mathcal{A}, p, r)$. Let $\nu^1, \ldots, \nu^k$ denote k available assets. Assume each asset ν^i can take on only positive values inside some set $S^i \subset \mathbb{R}^+$. Let $\mathcal{S} = S_1 \times \ldots \times S_k$ denote the set of all possible combinations of values that the assets can take. Let $\mathcal{A} := \{a \in \mathbb{R}^k \mid \sum_{i=1}^k a^i = 1, a^i \geq 0\}$ denote the k-dimensional probability simplex. We interpret a given element $a = (a^1, \ldots, a^k) \in \mathcal{A}$ as an allocation of principal to each of the assets $\nu^1, \ldots, \nu^k$: the proportion a^j of our total principal is allocated to asset ν^j, for each $j = 1, \ldots, k$. Let a transition kernel $p : \mathcal{S} \times \mathcal{A} \to \mathcal{D}(\mathcal{S})$ be given that represents the market dynamics.[1] The expected return of choosing allocation a in state s is given by $r(s, a) = \int_{\mathcal{S}} \left[\sum_j a^j(\nu^j_{s'} - \nu^j_s)/\nu^j_s\right] p(s'|s, a)\ ds'$, where ν^j_s denotes the value of asset ν^j in state s.[2] Notice that, at timestep t, we can estimate $r(s_t, a_t)$ using $\widehat{r}(s_t, a_t) = \sum_j a^j(\nu^j_{s_{t+1}} - \nu^j_{s_t})/\nu^j_{s_t}$. Bearing this in mind, we henceforth assume we have an estimate of $r(s_t, a_t)$ at each timestep.

Based on the above average-reward MDP $(\mathcal{S}, \mathcal{A}, p, r)$, we formulate the Omega ratio optimization problem as an MROP

[1] In most cases the transition dynamics, p, will be independent of the action chosen, since the act of rebalancing the portfolio will have a negligible effect on the overall market. In certain situations, however, such as when the total principal of the portfolio is very large, actions taken by the portfolio manager may influence the market. Since our results go through for the more general case where p depends on the action chosen, we give our results for this case.

[2] This reward can be easily modified to take transaction costs into account.

as follows. Given a policy π, the returns distribution R induced by π, and a threshold τ, we rewrite the Omega ratio defined in (3) in a more tractable form. We can perform a simple integration by parts to obtain $\int_\tau^\infty [1-F_R(x)]dx = \mathbb{E}\left[\max(0, R-\tau)\right]$ and $\int_{-\infty}^\tau F_R(x)dx = \mathbb{E}\left[\max(0, \tau - R)\right]$. This implies that we can rewrite the Omega ratio (3) as follows:

$$\begin{aligned}\Omega(\pi;\tau) &= \frac{\mathbb{E}_\pi\left[\max(0, R-\tau)\right]}{\mathbb{E}_\pi\left[\max(0, \tau - R)\right]} \\ &= \frac{\int_{\mathcal{S}}\int_{\mathcal{A}} \max(0, r(s,a)-\tau)\lambda_\pi(s,a)\ da\ ds}{\int_{\mathcal{S}}\int_{\mathcal{A}} \max(0, \tau - r(s,a))\lambda_\pi(s,a)\ da\ ds}\end{aligned} \tag{4}$$

Taking $f(\lambda_\pi) = \mathbb{E}_\pi\left[\max(0, R-\tau)\right]$ and $g(\lambda_\pi) = \mathbb{E}_\pi\left[\max(0, \tau - R)\right]$ completes our formulation of the Omega ratio maximization problem as an MROP, where our goal is to find a policy π maximizing objective (4).

III. Algorithm

A. Omega Ratio Policy Gradient

Consider an MDP $(\mathcal{S}, \mathcal{A}, p, r)$ and parametrized policy class $\{\pi_\theta\}_{\theta\in\Theta}$. Given policy parameter θ, define the relative state value function $V_\theta(s) = \sum_{t=0}^\infty \mathbb{E}_{\pi_\theta}\left[r(s,a) - J(\theta) \mid s_0 = s\right]$, relative state-action value function $Q_\theta(s,a) = \sum_{t=0}^\infty \mathbb{E}_{\pi_\theta}\left[r(s,a) - J(\theta) \mid s_0 = s, a_0 = a\right]$, and advantage function $A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$. Under the assumption that $\pi_\theta(a|s)$ is differentiable in θ, for all $s \in \mathcal{S}, a \in \mathcal{A}$, by the classic policy gradient theorem [2] we have

$$\nabla J(\theta) = \mathbb{E}_{\pi_\theta}\Big[A_\theta(s,a)\nabla \log \pi_\theta(a|s)\Big]. \tag{5}$$

We are interested in applying a variation of the policy gradient method to maximizing the Omega ratio (4). Since this objective is a ratio of two expectations, however, the policy gradient theorem (5) does not directly apply. Instead, we apply standard calculus in conjunction with two distinct applications of (5) to obtain a tractable gradient expression as follows. For a fixed risk-free rate of return τ, let $r^+(s,a) = \max(0, r(s,a) - \tau)$ and $r^-(s,a) = \max(0, \tau - r(s,a))$, with corresponding MDPs $(\mathcal{S}, \mathcal{A}, p, r^+)$ and $(\mathcal{S}, \mathcal{A}, p, r^-)$. Next, define $J^+(\theta) = \mathbb{E}_{\pi_\theta}\left[r^+(s,a)\right]$ and $J^-(\theta) = \mathbb{E}_{\pi_\theta}\left[r^-(s,a)\right]$, as well as the corresponding value and advantage functions $V_\theta^+, Q_\theta^+, A_\theta^+$, and $V_\theta^-, Q_\theta^-, A_\theta^-$. Clearly, $\Omega(\theta;\tau) := \Omega(\pi_\theta;\tau) = \frac{J^+(\theta)}{J^-(\theta)}$. Furthermore, by the quotient rule we have

$$\nabla_\theta \Omega(\theta;\tau) = \frac{J^-(\theta)\nabla J^+(\theta) - J^+(\theta)\nabla J^-(\theta)}{\left[J^-(\theta)\right]^2}. \tag{6}$$

Combined with the classic policy gradient theorem (5), this gives $\nabla\Omega(\theta;\tau) =$

$$\mathbb{E}_{\pi_\theta}\left[\frac{J^-(\theta)A_\theta^+(s,a) - J^+(\theta)A_\theta^-(s,a)}{\left[J^-(\theta)\right]^2}\nabla\log\pi_\theta(a|s)\right], \tag{7}$$

As long as we can estimate $J^+(\theta), J^-(\theta), A_\theta^+$, and A_θ^-, we can therefore also estimate the gradient of the Omega ratio.

B. Actor-Critic Algorithm

Based on the policy gradient expression above, we now develop a variant of the classic actor-critic algorithm [3], [23] for maximizing the Omega ratio. In order to estimate $\nabla\Omega(\theta;\tau)$, we use two critics, one for each of the state value functions associated with the MDPs $(\mathcal{S}, \mathcal{A}, p, r^+)$ and $(\mathcal{S}, \mathcal{A}, p, r^-)$ defined above. We assume for ease of exposition that we use the same class $\{v_\omega\}_{\omega\in\Omega}$ of state value function approximators for both critics, but this assumption can be removed both in practice and in the convergence analysis. Let a parametric policy class $\{\pi_\theta\}_{\theta\in\Theta}$ and a class $\{v_\omega\}_{\omega\in\Omega}$ of state value function approximators be given. The algorithm is as follows.

Initialize stepsize sequences $\{\alpha_t\}, \{\beta_t\}$ and initial policy parameter θ_0. Initialize the V^+ and V^- critic parameters ω_0^+ and ω_0^-, respectively, as well as constants $\mu_{-1}^+, \mu_{-1}^- > 0$. At timestep t, update the estimates μ_t^+, μ_t^- of $J^+(\theta_t), J^-(\theta_t)$ via

$$\mu_t^+ = (1-\alpha_t)\mu_{t-1}^+ + \alpha_t r^+(s_t, a_t), \tag{8}$$

$$\mu_t^- = (1-\alpha_t)\mu_{t-1}^- + \alpha_t r^-(s_t, a_t). \tag{9}$$

Next, update the TD errors[3]

$$\delta_t^+ = r^+(s_t,a_t) - \mu_t^+ + v_{\omega_t^+}(s_{t+1}) - v_{\omega_t^+}(s_t), \tag{10}$$

$$\delta_t^- = r^-(s_t,a_t) - \mu_t^- + v_{\omega_t^-}(s_{t+1}) - v_{\omega_t^-}(s_t). \tag{11}$$

Next, perform the critic updates

$$\omega_{t+1}^+ = \omega_t^+ + \alpha_t\delta_t^+\nabla v_{\omega_t^+}(s_t), \tag{12}$$

$$\omega_{t+1}^- = \omega_t^- + \alpha_t\delta_t^-\nabla v_{\omega_t^-}(s_t). \tag{13}$$

Finally, carry out the actor update

$$\theta_{t+1} = \theta_t - \beta_t\frac{\mu_t^-\delta_t^+ - \mu_t^+\delta_t^-}{\left[\mu_t^-\right]^2}\nabla\log\pi_{\theta_t}(a_t|s_t). \tag{14}$$

The iterative algorithm just given is what we will study in our convergence analysis. In practice, however, it is often advantageous to perform rollouts of some length K at each timestep, then use data from the entire rollout to obtain lower-variance gradient estimates for the critic and actor updates.

IV. Theoretical Results

A. Concave Reformulation

Given access to the transition probability function p and functions f, g of an MROP with finite state and action spaces, the problem of finding the optimal state-action occupancy measure can be formulated as the following problem:

$$\begin{aligned}\max_\lambda\quad & \frac{f(\lambda)}{g(\lambda)} \\ \text{s.t.}\quad & \sum_s\sum_a \lambda_{sa} = 1, \\ & \sum_a \lambda_{sa} = \sum_{s'}\sum_a p(s|s',a)\lambda_{s'a}, \quad \forall s \in \mathcal{S}, \\ & \lambda \geq 0.\end{aligned} \tag{R_0}$$

See [22] for details on why the constraints ensure λ is a valid occupancy measure given the transition probability function p. When f is concave and g is affine and positive over the feasible region, problem (R_0) is a *quasiconcave* optimization

[3]It is well-known that the TD errors δ_t^+, δ_t^- approximate the advantage functions $A_{\theta_t}^+(s_t,a_t), A_{\theta_t}^-(s_t,a_t)$, respectively.

problem. A function $h : \mathbb{R}^n \to \mathbb{R}$ is quasiconcave if $h(\gamma x + (1-\gamma)y) \geq \min(h(x), h(y))$, for all $x, y \in \mathbb{R}^n, \gamma \in [0,1]$. Note that when $c > 0$ is a vector of positive costs and $g(\lambda) = c^T\lambda$, as in the classic MDP setting, then g is linear (and thus affine) and positive over the feasible region. In this case, by [24, Prop. 7.2] we know that (R_0) can be transformed via the variable transformation $y = \frac{\lambda}{g(\lambda)}, t = \frac{1}{g(\lambda)}$ to the equivalent concave program

$$\begin{aligned} \max_{y,t} \quad & tf\left(\frac{y}{t}\right) \\ \text{s.t.} \quad & \sum_s \sum_a y_{sa} - t = 0, \qquad (R_1) \\ & \sum_a y_{sa} - \sum_{s'} \sum_a p(s|s',a) y_{s'a} = 0, \quad \forall s \in \mathcal{S}, \\ & \sum_s \sum_a c_{sa} y_{sa} + bt = 1, \\ & y \geq 0. \end{aligned}$$

Solving (R_1) to obtain the optimal (y^*, t^*), we can immediately recover the optimal state-action occupancy measure $\lambda^* = y^*/t^*$ for (R_0). An optimal policy π^* for the MROP $(\mathcal{S}, \mathcal{A}, p, f, g)$ is then recovered via $\pi^*(a|s) = \lambda^*_{sa} / \sum_{a'} \lambda^*_{sa'}$.

Returning to our MROP formulation of the Omega ratio optimization problem, in the case where $\mathcal{S}$ and $\mathcal{A}$ are finite[4] we can express the Omega ratio as

$$\Omega(\pi; \tau) = \frac{\sum_s \sum_a \max(0, r(s,a) - \tau)\lambda_\pi(s,a)}{\sum_s \sum_a \max(0, \tau - r(s,a))\lambda_\pi(s,a)}. \qquad (15)$$

The numerator and denominator of this objective are both linear in the state-action occupancy measure and the denominator is strictly positive so long as there is always non-negligible downside risk (i.e., positive expected shortfall), so the concave transformation described above for the more general MROP setting applies to our Omega ratio MROP. We use this fact below to prove that the algorithm described in the previous section converges to the global optimum of the Omega ratio.

B. Hidden Quasiconcavity

In this section we show that the problem of maximizing the Omega ratio enjoys a powerful *hidden quasiconcavity* property. Importantly, this property implies that policy gradient algorithms like the actor-critic algorithm described above find the global optimum of the Omega ratio problem under certain conditions, even when the policy parametrization results in a highly non-concave objective.

Consider the Omega ratio MROP $(\mathcal{S}, \mathcal{A}, p, f, g)$ given above, where $f(\lambda_\pi) = \mathbb{E}_\pi[\max(0, R - \tau)]$, $g(\lambda_\pi) = \mathbb{E}_\pi[\max(0, \tau - R)]$, τ is the risk-free rate of return, and $\mathcal{S}$ and $\mathcal{A}$ are finite. Assume that there is always non-negligible downside risk, so $g(\lambda_\pi)$ is always strictly positive, and recall that f and g are both linear in λ. Let a convex, compact set $\Theta \subset \mathbb{R}^d$, where $d < |\mathcal{S}||\mathcal{A}|$, and a parametrized policy class $\{\pi_\theta\}_{\theta \in \Theta}$ be given. Let $\lambda : \Theta \to \mathcal{D}(\mathcal{S} \times \mathcal{A})$ be a function mapping each parameter vector $\theta \in \Theta$ to the state-action occupancy measure $\lambda(\theta) := \lambda_\theta := \lambda_{\pi_\theta}$ induced by the policy π_θ over $\mathcal{S} \times \mathcal{A}$. Consider the optimization problem

$$\max_{\theta \in \Theta} \quad \Omega(\theta; \tau) = \frac{f(\lambda_\theta)}{g(\lambda_\theta)}. \qquad (16)$$

Note that the objective in (16) is potentially highly non-concave as a function of θ. It turns out that, under the following conditions, every stationary point of this problem nonetheless corresponds to a global optimum.

Assumption 1. The following statements hold:

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

We have the following characterization of the hidden quasiconcavity property.

Theorem 1. Let Assumption 1 hold, and let θ^* be a stationary point of (16), i.e., $\nabla\Omega(\theta^*; \tau) = 0$. Then θ^* is globally optimal for (16).

The proof is modeled after that of [7, Thm. 4.2], with key modifications to accommodate the fact that the underlying ratio optimization problem is not concave, but *quasiconcave* in the state-action occupancy measure. Theorem 1 implies that any algorithm that finds a stationary point of the Omega ratio optimization problem (16) in fact finds its global optimum. This result greatly strengthens the asymptotic convergence analysis provided in the following section.

C. Convergence

In this section we show almost sure (a.s.) convergence of the Omega ratio actor-critic algorithm to a neighborhood of a stationary point of the optimization problem $\max_{\theta \in \Theta} \Omega(\theta; \tau)$. As discussed in Section IV-B, this implies that, thanks to the hidden quasiconcavity of the Omega ratio optimization problem, the algorithm converges a.s. to a neighborhood of a *global* optimum. This result is much stronger than existing asymptotic results for actor-critic schemes, which typically only guarantee convergence to a neighborhood of a local optimum or saddle point.

We analyze the algorithm as given in equations (8)-(14), with the addition of a projection operation to equation (14):

$$\theta_{t+1} = \Gamma\left(\theta_t - \beta_t \frac{\mu_t^- \delta_t^+ - \mu_t^+ \delta_t^-}{\left[\mu_t^-\right]^2} \nabla \log \pi_{\theta_t}(a_t|s_t)\right), \qquad (17)$$

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 literature (see, e.g., [3], [25], [26]) is for purposes of theoretical analysis, and

[4]For practical purposes, the range of possible values the assets can attain is bounded, i.e., for each asset e^i we have $e^i \in S^i = [0, M^i]$, for some upper bound M^i. This means that the state space $\mathcal{S} = S^1 \times \ldots \times S^k$ is also bounded. Since $\mathcal{A}$ is already bounded, we can discretize both $\mathcal{S}$ and $\mathcal{A}$ to obtain a finite state- and action-space approximation to the original MDP.

is typically not needed in practice. We make the following additional assumptions, which are standard in the literature.

Assumption 2. The stepsize sequences $\{\alpha_n\}$ and $\{\beta_n\}$ satisfy $\sum_n \alpha_n = \sum_n \beta_n = \infty,\ \sum_n \alpha_n^2 + \beta_n^2 < \infty$, and $\lim_n \frac{\beta_n}{\alpha_n} = 0$.

Assumption 3. 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.

Assumption 4. The set of permissible policy parameters $\Theta \subset \mathbb{R}^d$ is a compact set. Furthermore, 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.

The following analysis leverages the average-reward actor-critic results in [3] and borrows heavily from the analysis of cost-aware actor-critic in [11]. For a given policy parameter θ, let $D_\theta = \text{diag}(d_\theta) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote the matrix with the elements of d_θ along the diagonal and zeros everywhere else. Define the state reward vector for the MDP $(\mathcal{S}, \mathcal{A}, p, r^+)$ to be $r_\theta^+ = [r_\theta^+(s)]_{s\in\mathcal{S}}^\top \in \mathbb{R}^{|\mathcal{S}|}$, where $r_\theta^+(s) = \sum_{a\in\mathcal{A}} \pi_\theta(a|s) r^+(s,a)$. Define the state reward vector r_θ^- for the MDP $(\mathcal{S}, \mathcal{A}, p, r^-)$ similarly. 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 have convergence of the critics. The proofs of this and the subsequent results are routine; see [3], [11] for details.

Lemma 1. Under Assumption 3, given a fixed policy parameter $\theta \in \Theta$, the recursive updates (8)-(13) converge as follows: $\lim_{t\to\infty} \mu_t^+ = J^+(\theta)$ a.s., $\lim_{t\to\infty} \mu_t^- = J^-(\theta)$ a.s., $\lim_{t\to\infty} \omega_t^+ = \omega_\theta^+$ a.s., and $\lim_{t\to\infty} \omega_t^- = \omega_\theta^-$ a.s., where ω_θ^+ and ω_θ^- are, respectively, the unique solutions to

$$\Phi^\top D_\theta \left[r_\theta^+ - J^+(\theta)\cdot\mathbf{1} + P_\theta(\Phi\omega^+) - \Phi\omega^+ \right] = \mathbf{0}, \tag{18}$$

$$\Phi^\top D_\theta \left[r_\theta^- - J^-(\theta)\cdot\mathbf{1} + P_\theta(\Phi\omega^-) - \Phi\omega^- \right] = \mathbf{0}. \tag{19}$$

This result shows that the sequences $\{\omega_t^+\}$ and $\{\omega_t^-\}$ converge a.s. to the limit points ω_θ^+ and ω_θ^- 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 $v_\theta^+ = \Phi\omega_\theta^+$ and $v_\theta^- = \Phi\omega_\theta^-$ may result in biased gradient estimates. Similar to the bias characterization given in [3, Lemma 4], this bias can be characterized as follows.

Lemma 2. Fix $\theta \in \Theta$. Let

$$\delta_t^{\theta,+} = r^+(s_t,a_t) - J^+(\theta) + \phi(s_{t+1})^\top \omega_\theta^+ - \phi(s_t)^\top \omega_\theta^+,$$
$$\delta_t^{\theta,-} = r^-(s_t,a_t) - J^-(\theta) + \phi(s_{t+1})^\top \omega_\theta^- - \phi(s_t)^\top \omega_\theta^-,$$

denote the stationary estimates of the TD-errors at time t. Let

$$\overline{v}_\theta^+ = \mathbb{E}_{\pi_\theta}\left[r^+(s,a) - J^+(\theta) + \phi(s')^\top \omega_\theta^+ \right],$$
$$\overline{v}_\theta^- = \mathbb{E}_{\pi_\theta}\left[r^-(s,a) - J^-(\theta) + \phi(s')^\top \omega_\theta^- \right],$$

where s' denotes the state visited after s, and let

$$\epsilon_\theta^+ = \sum_{s\in\mathcal{S}} d_\theta(s) \left[\nabla_\theta \overline{v}_\theta^+(s) - \nabla_\theta \phi(s)^\top \omega_\theta^+ \right],$$
$$\epsilon_\theta^- = \sum_{s\in\mathcal{S}} d_\theta(s) \left[\nabla_\theta \overline{v}_\theta^-(s) - \nabla_\theta \phi(s)^\top \omega_\theta^- \right].$$

We then have that

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

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 $\epsilon_\theta = \left(J^-(\theta)\epsilon_\theta^+ - J^+(\theta)\epsilon_\theta^-\right)/\left[J^-(\theta)\right]^2$, and consider the ordinary differential equations (ODEs)

$$\dot{\theta} = \hat{\Gamma}(\nabla\Omega(\theta;\tau)), \tag{20}$$

$$\dot{\theta} = \hat{\Gamma}(\nabla\Omega(\theta;\tau) + \epsilon_\theta). \tag{21}$$

Let $\mathcal{Z}, \mathcal{Y}$ denote the sets of asymptotically stable equilibria of the ODEs (20), (21), respectively. In addition, given a set $\mathcal{B}$ and constant $\varepsilon > 0$, define the ε-neighborhood of $\mathcal{B}$ to be $\mathcal{B}^\varepsilon = \{x \mid \inf_{z\in\mathcal{B}} \|x - z\| \leq \varepsilon\}$. We then have the following theorem.

Theorem 2. Under Assumptions 2 and 4, given any $\varepsilon > 0$, there exists $\delta > 0$ such that, for $\{\theta_t\}$ obtained from the recursive scheme (8)-(14) , if $\sup_t \|\epsilon_{\theta_t}\| < \delta$, then $\theta_t \to \mathcal{Z}^\varepsilon$ a.s. as $t \to \infty$.

Combined with the results of Section IV-B, Theorem 2 establishes almost sure convergence of our Omega ratio actor-critic algorithm to a neighborhood of a global optimum of $\Omega(\theta;\tau)$ when linear function approximation is used for the critic. Note that, as the expressive capacity of our approximation scheme improves, the biases and resulting error terms tend to zero.

V. Experimental Results

To illustrate the theory developed above, we trained our Omega ratio maximization algorithm on a small portfolio optimization problem and compared it with a random allocation strategy. We implemented the Omega ratio actor-critic algorithm as described in Section III. We used tabular softmax policies, i.e., $\pi_\theta(a_i|s) = \exp(\theta^T\psi(s,a_i))/\sum_j \exp(\theta^T\psi(s,a_j))$, where $\theta \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ and $\psi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ maps each state-action pair to a unique standard basis vector $e_k \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$, where e_k has a 1 in its kth entry and 0 everywhere else. We similarly used tabular representation for the value functions: $v_\omega(s) = \omega^T\phi(s)$. Through trial and error we chose hyperparameters $\alpha_t = 2.0, \beta_t = 1.0, \tau = 0.1, \mu_0^+ = \mu_0^- = 1.0$, and we initialized the actor and critic parameters using zero-mean Gaussian distributions. Our portfolio environment consisted of three assets. Each asset was constrained to lie in the interval $[50, 55]$, which we discretized into five distinct values. We also

Fig. 1: Learning curve for Omega ratio actor-critic with mean and 95% confidence intervals over five replications. Each episode is 360 months and $\tau = 0$.

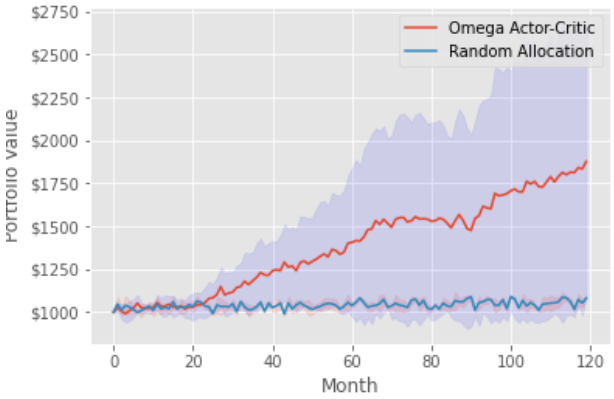

Fig. 2: Portfolio performance over 10 years of learned and random policies with mean and 95% confidence intervals over 100 replications.

discretized the action space into five possible values along each dimension. The transition probabilities were chosen so that the average monthly returns of the three assets were $\%2.5, \%5$, and $-\%2$, with volatilities $\%3.4, \%3.6$, and $\%3.2$, respectively. The agent was permitted to rebalance the portfolio monthly. We trained our Omega ratio algorithm using five different initializations, pictured in Figure 1, then evaluated the resulting policies against uniformly random policies in Figure 2. Figure 1 shows that the algorithm steadily increases the Omega ratio, while Figure 2 illustrates that the resulting policies significantly outperform the random baseline policy.

VI. Conclusion

In this paper we have investigated policy gradient algorithms for ratio optimization problems via an illustrative case study, leveraging the hidden quasiconcavity of such problems to prove convergence to (neighborhoods of) global optima. Although our algorithm and results are given for the Omega ratio in this paper, the same steps apply to a much more general class of MROPs; this is the topic of forthcoming work.

References

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

[2] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation." in *NIPS*, vol. 99. Citeseer, 1999, pp. 1057–1063.

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

[4] V. S. Borkar and V. R. Konda, "The actor-critic algorithm as multi-time-scale stochastic approximation," *Sadhana*, vol. 22, no. 4, pp. 525–543, 1997.

[5] H. Kumar, A. Koppel, and A. Ribeiro, "On the sample complexity of actor-critic method for reinforcement learning with function approximation," *arXiv preprint arXiv:1910.08412*, 2019.

[6] T. Xu, Z. Wang, and Y. Liang, "Improving sample complexity bounds for (natural) actor-critic algorithms," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 4358–4369.

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

[8] J. Zhang, A. S. Bedi, M. Wang, and A. Koppel, "Beyond cumulative returns via reinforcement learning over state-action occupancy measures," in *2021 American Control Conference (ACC)*, 2021, pp. 894–901.

[9] J. Zhang, C. Ni, Z. Yu, C. Szepesvari, and M. Wang, "On the convergence and sample efficiency of variance-reduced policy gradient method," *arXiv preprint arXiv:2102.08607*, 2021.

[10] Z. Ren and B. H. Krogh, "Markov decision processes with fractional costs," *IEEE Transactions on Automatic Control*, vol. 50, no. 5, pp. 646–650, 2005.

[11] W. Suttle, K. Zhang, Z. Yang, D. Kraemer, and J. Liu, "Reinforcement learning for cost-aware Markov decision processes," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9989–9999.

[12] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[13] B. M. Rom and K. W. Ferguson, "Post-modern portfolio theory comes of age," *Journal of Investing*, vol. 3, no. 3, pp. 11–17, 1994.

[14] W. F. Sharpe, "Mutual fund performance," *The Journal of Business*, vol. 39, no. 1, pp. 119–138, 1966.

[15] T. W. Young, "Calmar ratio: A smoother tool," *Futures*, vol. 20, no. 1, p. 40, 1991.

[16] F. A. Sortino and L. N. Price, "Performance measurement in a downside risk framework," *The Journal of Investing*, vol. 3, no. 3, pp. 59–64, 1994.

[17] C. Keating and W. F. Shadwick, "A universal performance measure," *Journal of Performance Measurement*, vol. 6, no. 3, pp. 59–84, 2002.

[18] J. E. Moody, M. Saffell, Y. Liao, and L. Wu, "Reinforcement learning for trading systems and portfolios." in *KDD*, 1998, pp. 279–283.

[19] R. Neuneier, "Enhancing Q-learning for optimal asset allocation," in *Advances in neural information processing systems*, 1998, pp. 936–942.

[20] J. Moody and M. Saffell, "Learning to trade via direct reinforcement," *IEEE Transactions on Neural Networks*, vol. 12, no. 4, pp. 875–889, 2001.

[21] Z. Jiang, D. Xu, and J. Liang, "A deep reinforcement learning framework for the financial portfolio management problem," *arXiv preprint arXiv:1706.10059*, 2017.

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

[23] V. Konda, "Actor-Critic Algorithms," Ph.D. dissertation, MIT, 2002.

[24] M. Avriel, W. E. Diewert, S. Schaible, and I. Zang, *Generalized Concavity*. SIAM, 2010.

[25] H. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, ser. Stochastic Modelling and Applied Probability. Springer New York, 2003.

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