

Policy Gradient for Ratio Optimization Problems: A Case Study

56th Conference on Information Sciences and Systems, 2022

Wesley A. Suttle, Alec Koppel, Ji Liu

Motivation

Ratio optimization problems have applications in a variety of sequential decision-making settings:

- risk-return trade-offs in financial portfolio management
- area sweeping tasks in robotic motion planning
- price-performance ratios in engineering and economics
- bandwidth in computer network design and management

Reinforcement learning (RL) has seen immense growth, yet RL for ratio optimization problems remains largely unexplored

Motivation

Why is RL for ratio optimization problems theoretically interesting?

- After all, naïve policy gradient methods can be applied when gradients are tractable [SZY⁺21]
- But this shallow view neglects a **rich underlying structure** inherent in many ratio optimization problems

Such problems enjoy a powerful hidden quasiconcavity property ensuring policy gradient algorithms converge to global optima

Contributions

In this work, we illuminate this structure through a specific example: maximizing the Omega ratio of a financial portfolio

More concretely, we:

- 1 propose the **Markov ratio optimization process (MROP)** framework for decision-making problems with a ratio objective
- 2 discuss the useful structure and powerful **hidden quasiconcavity** property of MROP problems
- 3 formulate the Omega ratio problem as an MROP and develop **Omega ratio actor-critic** for solving it
- 4 employ hidden quasiconcavity of the Omega ratio MROP to establish a **global optimality guarantee** for our algorithm
- 5 **experimentally evaluate** our method on a toy portfolio problem

The Omega ratio problem is an illustrative example, but MROP structure applies to a much wider range of ratio problems

Related Work

We develop the concept of hidden quasiconcavity by generalizing the notion of hidden concavity presented in

- Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvári, and Mengdi Wang. [Variational policy gradient method for reinforcement learning with general utilities](#).

Advances in Neural Information Processing Systems, 33:4572–4583, 2020

Our MROP formulation and results generalize the cost-aware MDP formulation and strengthen the corresponding results discussed in

- Wesley Suttle, Kaiqing Zhang, Zhuoran Yang, David Kraemer, and Ji Liu. [Reinforcement learning for cost-aware Markov decision processes](#).

In *International Conference on Machine Learning*, pages 9989–9999. PMLR, 2021

Additional recent works exploring global optimality results for policy gradient methods for general utility functions include

- Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, and Alec Koppel. [Beyond cumulative returns via reinforcement learning over state-action occupancy measures](#).

In *2021 American Control Conference*, pages 894–901, 2021

- Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvári, and Mengdi Wang. [On the convergence and sample efficiency of variance-reduced policy gradient method](#).

arXiv preprint arXiv:2102.08607, 2021

Markov Decision Processes

- Average-reward MDP $(\mathcal{S}, \mathcal{A}, p, r)$
 - ▶ agent starts in state $s \in \mathcal{S}$, chooses action $a \in \mathcal{A}$ using policy π
 - ▶ receives reward $r(s, a)$, state transitions to $s' \sim p(\cdot|s, a)$, repeats
- **State occupancy measure** $d_\pi \in \mathcal{D}(\mathcal{S})$ induced by π on \mathcal{S} :

$$d_\pi(s) = \lim_{t \rightarrow \infty} P(s_t = s \mid \pi)$$

- **State-action occupancy measure** $\lambda_\pi \in \mathcal{D}(\mathcal{S} \times \mathcal{A})$ induced by π :

$$\lambda_\pi(s, a) = \lim_{t \rightarrow \infty} P(s_t = s, a_t = a \mid \pi) = d_\pi(s)\pi(a|s)$$

- Expected **average reward** under policy π :

$$J(\pi) = \lim_{n \rightarrow \infty} \frac{1}{n} E_\pi \left[\sum_{i=1}^n r(s_i, a_i) \right] = \int_{\mathcal{S}} \int_{\mathcal{A}} r(s, a) \pi(a|s) d_\pi(s) da ds$$

Markov Ratio Optimization Processes

MROPs enjoy rich underlying structure unlocking stronger convergence results for policy gradient algorithms

- Fix a controlled Markov chain $(\mathcal{S}, \mathcal{A}, p)$ obtained from $(\mathcal{S}, \mathcal{A}, p, r)$
- Given $f, g : D(\mathcal{S} \times \mathcal{A}) \rightarrow \mathbb{R}$, consider the objective function:

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

Definition

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 (1) over $(\mathcal{S}, \mathcal{A}, p)$.

MROPs subsume average-reward MDPs and cost-aware MDPs [SZY⁺ 21] as special cases

Omega Ratio Definition

Performance measures for decision-making problems in finance are frequently formulated as ratios of reward to risk

- Best-known example is the Sharpe ratio [Sha66]:

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

where R is rate of return and τ is the target rate of return.

- Omega ratio is a useful alternative [KS02], representing ratio of excess return to shortfall:

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

where F_R is the cumulative distribution function of R

Unlike the Sharpe ratio, the Omega ratio captures information about all higher moments of the returns distribution

Omega Ratio Problem

We can formulate Omega ratio maximization as an MROP

First, start with the following average-reward MDP $(\mathcal{S}, \mathcal{A}, p, r)$:

- k available assets: ν^1, \dots, ν^k
- \mathcal{S} , indexes the set of all possible asset value combinations:
 $s = (\nu_s^1, \dots, \nu_s^k) \in \mathcal{S} = S_1 \times \dots \times S_k$ means asset ν^j takes value ν_s^j
- \mathcal{A} , allocations to the k assets: $\mathcal{A} = \{a \in \mathbb{R}^k \mid \sum_{i=1}^k a^i = 1, a^i \geq 0\}$
- Transition kernel $p : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{D}(\mathcal{S})$ represents the market dynamics
- Reward $r(s, a) = \int_{\mathcal{S}} \left[\sum_{j=1}^k a^j (\nu_{s'}^j - \nu_s^j) / \nu_s^j \right] p(s' | s, a) ds'$

Now define f and g for the Omega ratio MROP:

- Notice $\Omega(\pi; \tau) = \frac{\mathbb{E}_{\pi}[\max(0, R - \tau)]}{\mathbb{E}_{\pi}[\max(0, \tau - R)]}$, so take $f(\lambda_{\pi}) = \mathbb{E}_{\pi}[\max(0, R - \tau)]$,
 $g(\lambda_{\pi}) = \mathbb{E}_{\pi}[\max(0, \tau - R)]$
- It turns out f and g are both linear in λ_{π} (we will use this later on)

Why should we care about the MROP formulation?

Concave Reformulation of General MROPs

Quasiconcave Program

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

Concave Program

$$\begin{aligned} \max_{y,t \geq 0} \quad & tf(y/t) && (R_1) \\ \text{s.t.} \quad & \sum_{s,a} y_{sa} = t, \quad \sum_{s,a} c_{sa} y_{sa} = 1, \\ & \sum_a y_{sa} = \sum_{s',a} p(s|s',a)y_{s'a}, \forall s \end{aligned}$$

- With model knowledge, in the tabular case the MROP $(\mathcal{S}, \mathcal{A}, p, f, g)$ **can be solved using (R_0)**
- When f is concave in λ and $g(\lambda) = c^T \lambda > 0$ is linear and strictly positive in λ , (R_0) is a quasiconcave program
- Using substitution $y = \lambda/g(\lambda)$, $t = 1/g(\lambda)$, [ADSZ10, Prop. 7.2] ensures **solving concave program (R_1) also solves (R_0)**

Hidden Quasiconcavity for General MROPs

Our concave reformulation unlocks stronger convergence results for policy gradient algorithms for solving MROPs

- Consider differentiable policy class $\{\pi_\theta\}_{\theta \in \Theta}$
- Let $\lambda : \Theta \rightarrow D(S \times A)$ map policy parameters to state-action occupancy measures $\lambda(\theta)$
- We want to solve the potentially **highly non-concave** problem

$$\underset{\theta \in \Theta}{\text{maximize}} \quad \frac{f(\lambda_\theta)}{g(\lambda_\theta)} \quad (2)$$

Despite its non-concavity, under suitable conditions, every stationary point of problem (2) is a global optimum

Hidden Quasiconcavity for General MROPs

Despite its non-concavity, under suitable conditions, every stationary point of problem $\max_{\theta \in \Theta} f(\lambda_\theta)/g(\lambda_\theta)$ is a global optimum:

Assumption. For a finite MROP $(\mathcal{S}, \mathcal{A}, p, f, g)$, where f is concave in λ and g is linear and strictly positive in λ , suppose the following statements hold: (i) the map $\lambda(\cdot)$ is a bijection and its image $\lambda(\Theta)$ is compact and convex; (ii) the inverse $\lambda^{-1}(\cdot)$ of $\lambda(\cdot)$ is Lipschitz continuous; (iii) the Jacobian $\nabla_\theta \lambda(\theta)$ is Lipschitz on Θ .

Theorem

If $\nabla f(\lambda_{\theta^})/g(\lambda_{\theta^*}) = 0$, then θ^* is globally optimal for $\max_{\theta \in \Theta} f(\lambda_\theta)/g(\lambda_\theta)$.*

Proof relies on concavity of the perspective transform, properties of the change of variables in our concave transformation, and extending [ZKB⁺20, Thm. 4.2].

This is the hidden quasiconcavity property: any policy gradient method that finds a stationary point finds a global optimum

Back to the Omega Ratio

Hidden quasiconcavity applies to the Omega ratio problem

- Recall the Omega ratio maximization problem we're interested in:

$$\underset{\theta \in \Theta}{\text{maximize}} \quad \Omega(\theta; \tau) = \frac{f(\lambda_\theta)}{g(\lambda_\theta)} = \frac{\mathbb{E}_{\pi_\theta} [\max(0, R - \tau)]}{\mathbb{E}_{\pi_\theta} [\max(0, \tau - R)]}. \quad (3)$$

- Remember that f and g are both linear in λ , and, so long as there is always downside risk, g will always be strictly positive
- The Omega ratio problem enjoys hidden quasiconcavity:**

Theorem

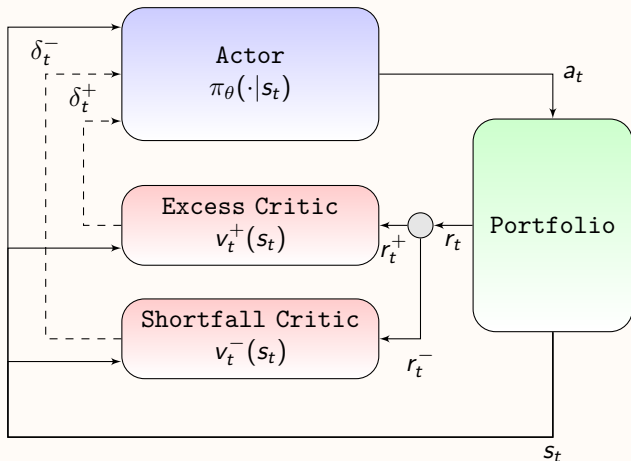
If $\nabla \Omega(\theta^*; \tau) = 0$, then θ^* is globally optimal for (3).

Policy gradient methods that find a stationary point of the Omega ratio problem are guaranteed to find a global optimum

Omega Ratio Actor-Critic

Algorithm uses two critics, one for excess return and one for shortfall.
Actor uses TD errors δ_t^+ , δ_t^- , and average excess and shortfall to estimate Omega ratio gradient.

- asset values s_t
- allocation a_t
- actual return $r_t = r(s_t, a_t)$
- excess return $r_t^+ = \max(0, r_t - \tau)$
- shortfall $r_t^- = \max(0, \tau - r_t)$
- excess TD error δ_t^+
- shortfall TD error δ_t^-



Convergence to (a neighborhood of) Global Optimum

Hidden quasiconcavity combines with two-timescale, asymptotic actor-critic analysis to produce a stronger result

Theorem

Under standard assumptions on decreasing stepsizes, linear critics, and actor differentiability conditions, the Omega ratio actor-critic algorithm given above converges almost surely to a neighborhood of a global maximum.

- Proof uses similar arguments to those in the actor-critic analysis of [SZY⁺21] combined with the global optimality results above

Portfolio Allocation Problem

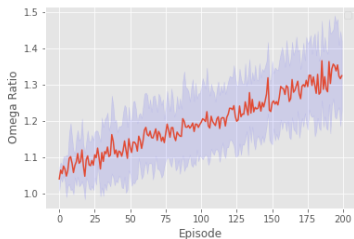


Figure: Learning curve for Omega ratio actor-critic with mean and 95% confidence intervals over five replications. Each episode is 360 months and $\tau = 0$.

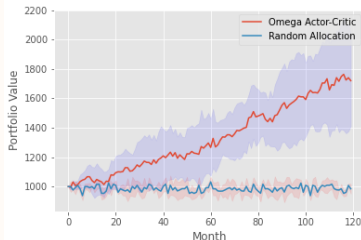


Figure: Portfolio performance over 10 years with mean and 95% confidence intervals over 100 replications. Portfolio value is measured in dollars.

- Compared tabular Omega actor-critic with random allocation strategy
- Three assets with average monthly returns 2.5%, 2%, and 0.1% and volatilities 3.5%, 3.3%, and 3.4%
- Agents allowed to rebalance the portfolio monthly

Omega actor-critic agent learns a policy striking a balance between maximizing return and limiting downside risk

Ongoing and Future Work

Our recent work tackles a more challenging MROP, the occupancy information ratio (OIR) problem, which provides a new framework for addressing the exploration/exploitation trade-off in RL:

- Wesley A Suttle, Alec Koppel, and Ji Liu. **Occupancy information ratio: Infinite-horizon, information-directed, parameterized policy search.**
arXiv preprint arXiv:2201.08832, 2022

Ongoing work includes application of OIR techniques to data-generation for offline RL, as well as a more complete and thorough study of the theory and applications of the MROP problem; these are the subject of forthcoming works.


Thanks for listening! Questions? Comments?


wesley.suttle@stonybrook.edu


The research of W. Suttle was sponsored by the U.S. Army Research Laboratory (ARL) and was accomplished under Cooperative Agreement Number W911NF-22-2-0003. The research of J. Liu was supported in part by ARL Cooperative Agreement W911NF-21-2-0098. A. Koppel contributed to this work while at CISD, ARL, in Adelphi, MD 20783.


References


 Mordecai Avriel, Walter E Diewert, Siegfried Schaible, and Israel Zang.
Generalized Concavity.
SIAM, 2010.


 Con Keating and William F Shadwick.
A universal performance measure.
Journal of Performance Measurement, 6(3):59–84, 2002.


 William F Sharpe.
Mutual fund performance.
The Journal of Business, 39(1):119–138, 1966.

 Wesley A Suttle, Alec Koppel, and Ji Liu.
Occupancy information ratio: Infinite-horizon, information-directed, parameterized policy search.
arXiv preprint arXiv:2201.08832, 2022.

 Wesley Suttle, Kaiqing Zhang, Zhuoran Yang, David Kraemer, and Ji Liu.
Reinforcement learning for cost-aware Markov decision processes.
In International Conference on Machine Learning, pages 9989–9999. PMLR, 2021.

 Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, and Alec Koppel.
Beyond cumulative returns via reinforcement learning over state-action occupancy measures.
In 2021 American Control Conference, pages 894–901, 2021.

 Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvári, and Mengdi Wang.
Variational policy gradient method for reinforcement learning with general utilities.
Advances in Neural Information Processing Systems, 33:4572–4583, 2020.

 Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvári, and Mengdi Wang.
On the convergence and sample efficiency of variance-reduced policy gradient method.
arXiv preprint arXiv:2102.08607, 2021.