MEAN-VARIANCE EFFICIENT REINFORCEMENT LEARNING BY EXPECTED QUADRATIC UTILITY MAXIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In reinforcement learning (RL) for sequential decision making under uncertainty, existing methods proposed for considering *mean-variance* (MV) trade-off suffer from computational difficulties in computation of the gradient of the variance term. In this paper, we aim to obtain MV-efficient policies that achieve Pareto efficiency regarding MV trade-off. To achieve this purpose, we train an agent to maximize the expected quadratic utility function, in which the maximizer corresponds to the Pareto efficient policy. Our approach does not suffer from the computational difficulties because it does not include gradient estimation of the variance. In experiments, we confirm the effectiveness of our proposed methods.

1 INTRODUCTION

Reinforcement learning (RL) trains intelligent agents to solve sequential decision-making problems (Puterman, 1994; Sutton & Barto, 1998). While a typical objective is to maximize the expected cumulative reward, *risk-aware* RL has recently attracted much attention in real-world applications, such as finance and robotics (Geibel & Wysotzki, 2005; Garcıa & Fernández, 2015). Various criteria have been proposed to capture a risk, such as Value at Risk (Chow & Ghavamzadeh, 2014; Chow et al., 2017) and variance (Markowitz, 1952; Luenberger et al., 1997; Tamar et al., 2012; Prashanth & Ghavamzadeh, 2013). Among them, we consider the *mean-variance RL* (MVRL) methods that attempt to train an agent while controlling the mean-variance (MV) trade-off (Tamar et al., 2012; Prashanth & Ghavamzadeh, 2013; 2016; Xie et al., 2018; Bisi et al., 2020; Zhang et al., 2021b).

Existing MVRL methods (Tamar et al., 2012; Prashanth & Ghavamzadeh, 2013; 2016; Xie et al., 2018), typically maximize the expected cumulative reward while keeping the variance of the cumulative reward at a certain level or, equivalently, minimize the variance while keeping the expected cumulative reward at a certain level. These MVRL methods simultaneously estimate the expected reward or variance while training an agent and solve the constrained optimization problem relaxed by penalized methods. These studies have reported that RL-based methods suffer from high computational difficulty owing to the *double sampling issue* (Section 3) when approximating the gradient of the variance term (Tamar et al., 2012; Prashanth & Ghavamzadeh, 2013; 2016). To avoid this, Tamar et al. (2012) and Prashanth & Ghavamzadeh (2013) proposed multi-time-scale stochastic optimization. Further, Xie et al. (2018) proposed a method based on the Legendre-Fenchel duality (Boyd & Vandenberghe, 2004). Although these methods avoid the double sampling issue, as we experimentally report in Figure 2 of Section 6, there still remains the difficulty in training a policy.

To avoid these difficulties, this paper considers another approach for MVRL by focusing on obtaining a policy that is located on the Pareto efficient frontier in the sense of MV trade-off; that is, we cannot increase the expected reward without increasing the variance and decrease the variance without decreasing the expected reward (Section 2.2). To achieve this purpose, we propose training a RL agent by direct *expected quadratic utility maximization* (EQUMRL) with a policy gradient method (Williams, 1988; 1992; Sutton et al., 2000; Baxter & Bartlett, 2001). We show that the maximizer of the objective in EQUMRL is Pareto efficient in the sense of MV trade-off.

As an important property, EQUMRL does not suffer from the double-sampling problem because it does not include the variance estimation, which is a cause of the problem. In conventional methods, we can also obtain an MV-efficient policy when we succeed in solving the constraint problem. However,

as discussed in the work and shown in our experiments, such as Figure 2, those methods do not perform well due to the difficulty in computation. Compared with them, EQUMRL is computationally friendly and experimentally returns more MV-efficient policies. In addition, the EQUMRL is well suited to financial applications, such as portfolio management because in economic theory, using MV portfolio is justified as a maximizer of the expected quadratic utility function (Markowitz, 1952; Luenberger et al., 1997). We introduce brief survey on expected quadratic utility maximization and finance in Appendix A.We list the following three advantages to the EQUMRL approach:

(i) our proposed EQUMRL is able to learn Pareto efficient policies and has plenty of interpretations from various perspectives (Section 4.3);
(ii) our proposed EQUMRL does not suffer from the double sampling issue by avoiding explicit approximation of the variance (Section 4);
(iii) we experimentally show that our proposed EQUMRL returns more Pareto efficient policies than existing methods (Section 6).

In the following sections, we first formulate the problem setting in Sections 2–3. Then, we propose the main algorithms in Section 4. Finally, we investigate the empirical effectiveness in Section 6.

2 Problem setting

We consider a standard setting of RL, where an agent interacts with an unfamiliar, dynamic, and stochastic environment modeled by a Markov decision process (MDP) in discrete time. We define an MDP as a tuple $(\mathcal{S}, \mathcal{A}, r, P, P_0)$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a stochastic reward function with finite mean and variance, $P : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a transition kernel, and $P_0 : \mathcal{S} \to [0, 1]$ is an initial state distribution. The initial state $\mathcal{S}_1$ is sampled from P_0. Let $\pi_\theta : \mathcal{A} \times \mathcal{S} \to [0, 1]$ be a parameterized stochastic policy mapping states to actions, where θ is a tunable parameter, and we denote the parameter space by Θ. At time step t, an agent chooses an action A_t following a policy $\pi_\theta(\cdot \mid S_t)$. We assume that the policy π_θ is differentiable with respect to θ; that is, $\frac{\partial \pi_\theta(s|a)}{\partial \theta}$ exists.

Let us define the expected cumulative reward from time step t to u as $\mathbb{E}_{\pi_\theta}[G_{t:u}]$, where $G_{t:u} = \sum_{i=0}^{u-t} \gamma^i r(S_{t+i}, A_{t+i})$, $\gamma \in (0, 1]$ is a discount factor and $\mathbb{E}_{\pi_\theta}$ denotes the expectation operator over a policy π_θ, and S_1 is generated from P_0. When $\gamma = 1$, to ensure that the cumulative reward is well-defined, we usually assume that all policies are proper (Bertsekas & Tsitsiklis, 1995); that is, for any policy π_θ, an agent goes to a recurrent state S^* with probability 1, and obtains 0 reward after passing the recurrent state S^* at a stopping time τ. This finite horizon setting is called *episodic* MDPs (Puterman, 1994). For brevity, we denote $G_{t:u}$ as G when there is no ambiguity.

In this paper, we consider the trade-off between the mean and variance of the reward. Let us denote the variance of a random variable W under π_θ by $\mathbb{V}_{\pi_\theta}(W)$. Note that $\mathbb{E}_{\pi_\theta}[G]$ and $\mathbb{V}_{\pi_\theta}(G)$ are finite because $r(S_t, A_t)$ has finite mean and variance.

2.1 Trajectory variance perspective.

A direction of MVRL is to consider the trajectory mean $\mathbb{E}_{\pi_\theta}[G_{t:u}]$ and trajectory variance $\mathbb{V}_{\pi_\theta}(G_{t:u})$:

$$\mathbb{E}_{\pi_\theta}[G] \overset{\text{trade-off}}{\Longleftrightarrow} \mathbb{V}_{\pi_\theta}(G).$$

A typical method for considering the MV trade-off is to train a policy under some constrains. Tamar et al. (2012), Prashanth & Ghavamzadeh (2013), and Xie et al. (2018) formulated MVRL as

$$\max_{\theta \in \Theta} \mathbb{E}_{\pi_\theta}[G] \quad \text{s.t. } \mathbb{V}_{\pi_\theta}(G) = \eta.$$

We call an algorithm based on this formulation a trajectory MV-controlled RL. In the trajectory MV-controlled RL, the goal is to maximize the expected cumulative reward, while controlling the trajectory-variance at a certain level. To be more precise, their actual constraint condition is $\mathbb{V}_{\pi_\theta}(G) \leq \eta$. However, if $\mathbb{V}_{\pi_\theta}(G) = \eta$ is feasible, the optimizer satisfies the equality in applications, where we need to consider MV trade-off, such as a financial portfolio management. Therefore, we only consider an equality constraint. To solve this problem, Tamar et al. (2012), Prashanth & Ghavamzadeh (2013), and Xie et al. (2018) considered a penalized method defined as

$\max_{\theta\in\Theta} \mathbb{E}_{\pi_\theta}[G] - \delta g\big(\mathbb{V}_{\pi_\theta}(G) - \eta\big)$, where $\delta > 0$ is a constant and $g : \mathbb{R} \to \mathbb{R}$ is a penalty function, such as $g(x) = x$ or $g(x) = x^2$. See Remark 1 for more details.

2.2 MV-EFFICIENT POLICY.

To consider the MV trade-off avoiding the computational difficulty, this paper aims to train a Pareto efficient policy in the sense of the MV trade-off, where we cannot increase the mean (resp. decrease the variance) without increasing the variance (resp. decreasing the mean). Following the existing literature mainly in economics, finance, and operations research (Luenberger et al., 1997), we define a trajectory MV-efficient policy as a policy π_θ such that there is no other policy with $\theta' \in \Theta$, where $\mathbb{E}_{\pi_\theta}[G] \leq \mathbb{E}_{\pi_{\theta'}}[G]$ and $\mathbb{V}_{\pi_\theta}[G] > \mathbb{V}_{\pi_{\theta'}}[G]$, or $\mathbb{V}_{\pi_\theta}[G] \geq \mathbb{V}_{\pi_{\theta'}}[G]$ and $\mathbb{E}_{\pi_\theta}[G] < \mathbb{E}_{\pi_{\theta'}}[G]$. Efficient frontier is defined as a set of the MV-efficient policies. From the definition of the MV-controlled policies, the trained policies belong to the efficient frontier.

3 POLICY GRADIENT AND DOUBLE SAMPLING ISSUE

In this paper, we consider training a policy by policy gradient methods. In MVRL, we usually require the gradients of $\mathbb{E}_{\pi_\theta}[G]$ and $\mathbb{V}_{\pi_\theta}(G) = \mathbb{E}\left[(G - \mathbb{E}_{\pi_\theta}[G])^2\right]$. Tamar et al. (2012) and Prashanth & Ghavamzadeh (2013) show that the gradients of $\mathbb{E}_{\pi_\theta}[G]$ and $\mathbb{E}_{\pi_\theta}\left[G^2\right]$ are given as

$$\nabla_\theta \mathbb{E}_{\pi_\theta}[G] = \mathbb{E}_{\pi_\theta}\left[G \sum_{t=1}^{\tau} \nabla_\theta \log \pi_\theta(S_t, A_t)\right], \quad \nabla_\theta \mathbb{E}_{\pi_\theta}\left[G^2\right] = \mathbb{E}\left[G^2 \sum_{t=1}^{\tau} \nabla_\theta \log \pi_\theta(S_t, A_t)\right].$$

Because optimizing the policy π_θ directly using the gradients is computationally intractable, we replace them with their unbiased estimators. Suppose that there is a simulator generating a trajectory k with $\{(S_t^k, A_t^k, r(S_t^k, A_t^k))\}_{t=1}^{\tau^k}$, where τ^k is the stopping time of the trajectory. Then, we can construct unbiased estimators of $\mathbb{E}_{\pi_\theta}[G]$ and $\mathbb{E}_{\pi_\theta}\left[G^2\right]$ as follows (Tamar et al., 2012):

$$\widehat{\nabla}_\theta \mathbb{E}_{\pi_\theta}[G] = \widehat{G}^k \sum_{t=1}^{\tau^k} \nabla_\theta \log \pi_\theta(S_t^k, A_t^k) \text{ and } \widehat{\nabla}_\theta \mathbb{E}_{\pi_\theta}\left[G^2\right] = \left(\widehat{G}^k\right)^2 \sum_{t=1}^{\tau^k} \nabla_\theta \log \pi_\theta(S_t^k, A_t^k), \quad (1)$$

where $\widehat{G}^k$ is a sample approximation of $\mathbb{E}_{\pi_\theta}[G]$ at the episode k. Besides, because $\nabla_\theta\big(\mathbb{E}_{\pi_\theta}[G]\big)^2 = 2\mathbb{E}_{\pi_\theta}[G]\nabla_\theta \mathbb{E}_{\pi_\theta}[G]$, the gradient of the variance is given as $\nabla_\theta \mathbb{V}_{\pi_\theta}(G) = \nabla_\theta\Big(\mathbb{E}_{\pi_\theta}[G^2] - \big(\mathbb{E}_{\pi_\theta}[G]\big)^2\Big) = \mathbb{E}_{\pi_\theta}\left[G^2 \sum_{t=1}^{\tau} \nabla_\theta \log \pi_\theta(S_t, A_t)\right] - 2\mathbb{E}_{\pi_\theta}[G]\nabla_\theta \mathbb{E}_{\pi_\theta}[G]$.

However, obtaining an unbiased estimator of $\nabla_\theta\big(\mathbb{E}_{\pi_\theta}[G]\big)^2 = 2\mathbb{E}_{\pi_\theta}[G]\nabla_\theta \mathbb{E}_{\pi_\theta}[G]$ is difficult because it requires sampling from two different trajectories for approximating $\mathbb{E}_{\pi_\theta}[G]$ and $\nabla_\theta \mathbb{E}_{\pi_\theta}[G]$. This issue is called double sampling issue and makes the optimization problem difficult when we include the variance $\mathbb{V}_{\pi_\theta}(G)$ into the objective function directly. Tamar et al. (2012) and Prashanth & Ghavamzadeh (2013) reported this double sampling issue caused from the gradient estimation of the variance and proposed solutions based on multi-time-scale stochastic optimization. Recall that they consider the penalized objective function $\max_{\theta\in\Theta} \mathbb{E}_{\pi_\theta}[G] - \delta g\big(\mathbb{V}_{\pi_\theta}(G) - \eta\big)$ to obtain MV-controlled policies, where the double sampling issue caused by the gradient of $\delta g\big(\mathbb{V}_{\pi_\theta}(G) - \eta\big)$.

4 EQUMRL WITH TRAJECTORY VARIANCE PERSPECTIVE

In social sciences, we often assume that an agent maximizes the expected quadratic utility for considering MV trade-off. Based on this, we propose the EQUMRL for obtaining an MV-efficient policy. For a cumulative reward G, we define the quadratic utility function as $u^{\text{trajectory}}(G; \alpha, \beta) = \alpha G - \frac{1}{2}\beta G^2$, where $\alpha > 0$ and $\beta \geq 0$. In EQUMRL, we train a policy by maximizing the expected value of the quadratic utility function,

$$\mathbb{E}_{\pi_\theta}\left[u^{\text{trajectory}}(G; \alpha, \beta)\right] = \alpha \mathbb{E}_{\pi_\theta}[G] - \frac{1}{2}\beta \mathbb{E}_{\pi_\theta}[G^2]. \quad (2)$$

The EQUMRL is agnostic to the learning method, that is, we can implement it with various existing algorithms such as REINFORCE and actor-crtic.

4.1 EQUMRL AND MV EFFICIENCY

The expected quadratic utility function $\mathbb{E}_{\pi_\theta}\left[u^{\text{trajectory}}(G;\alpha,\beta)\right]$ is known to be Pareto efficient in the sense of mean and variance when its optimal value satisfies $\mathbb{E}_{\pi_\theta}[G] \leq \frac{\alpha}{\beta}$. In order to confirm this, we can decompose the expected quadratic utility as

$$\mathbb{E}_{\pi_\theta}\left[u^{\text{trajectory}}(G;\alpha,\beta)\right] = -\frac{1}{2}\beta\left(\mathbb{E}_{\pi_\theta}[G] - \frac{\alpha}{\beta}\right)^2 + \frac{\alpha^2}{2\beta} - \frac{1}{2}\beta\mathbb{V}_{\pi_\theta}(G). \tag{3}$$

When a policy $\pi \in \Pi$ is the maximizer of the expected quadratic utility, it is equivalent to an MV-efficient policy (Borch, 1969; Baron, 1977; Luenberger et al., 1997). Following (Luenberger et al., 1997, p.237–239), we explain this as follows:

1. Among policies π_θ with a fixed mean $\mathbb{E}_{\pi_\theta}[G] = \mu$, the policy with the lowest variance maximizes the expected quadratic utility function because $\mathbb{E}_{\pi_\theta}\left[u^{\text{trajectory}}(G;\alpha,\beta)\right] = \alpha\mu - \frac{1}{2}\beta\mu^2 - \frac{1}{2}\beta\mathbb{V}_{\pi_\theta}(G)$ is a monotonous decreasing function on $\mathbb{V}_{\pi_\theta}(G)$;
2. Among policies π_θ with a fixed variance $\mathbb{V}_{\pi_\theta}(G) = \sigma^2$ and mean $\mathbb{E}_{\pi_\theta}[G] \leq \frac{\alpha}{\beta}$, the policy with the highest mean maximizes the expected quadratic utility function because $\mathbb{E}_{\pi_\theta}\left[u^{\text{trajectory}}(G;\alpha,\beta)\right] = -\frac{1}{2}\beta\left(\mathbb{E}_{\pi_\theta}[G] - \frac{\alpha}{\beta}\right)^2 + \frac{\alpha^2}{2\beta} - \frac{1}{2}\beta\sigma^2$ is a monotonous increasing function on $\mathbb{E}_{\pi_\theta}[G] \leq \frac{\alpha}{\beta}$.

Based on the above property, we propose maximizing the expected quadratic utility function in RL; that is, training an agent to directly maximize the expected quadratic utility function for MV control instead of solving a constrained optimization. We call the framework that makes the RL objective function an expected quadratic utility EQUMRL.

Unlike the expected cumulative reward maximization in the standard RL setting, at time t, it is desirable to include the past cumulative reward to the state S_t because our objective function depends on it even given S_t. Let us consider the objective at time t with the infinite horizon setting:

$$\begin{aligned}&\alpha\mathbb{E}_{\pi_\theta}\left[G_{0:\infty} - \beta G_{0:\infty}^2 \middle| S_0, A_0, r(S_0,A_0), \ldots, S_{t-1}, A_{t-1}, r(S_{t-1},A_{t-1}), S_t\right]\\ &= C + \alpha\gamma^t\left\{(1 - 2\beta\gamma^t G_{0:t-1})\mathbb{E}_{\pi_\theta}[G_{t:\infty}|S_t] - \beta\gamma^t\mathbb{E}_{\pi_\theta}[G_{t:\infty}^2|S_t]\right\},\end{aligned}$$

where C is a constant and recall that $G_{0:t-1} = \sum_{i=0}^{t-1}\gamma^i r(S_i, A_0)$.Thus, for a better decision-making, we include the past cumulative reward into the state space.

4.2 IMPLEMENTATION OF EQUMRL

In this section, we introduce how to train a policy with the EQURL. We defined the objective function of the EQUMRL, and EQUMRL is an agnostic in learning method. As examples, we show an implementation based on REINFORCE (Williams, 1992; Brockman et al., 2016) and Actor-Critic (AC) methods (Williams & Peng, 1991; Mnih et al., 2016). We use unbiased estimators of the gradients defined in (1).

REINFORCE-based trajectory EQUMRL. For an episode k with the length n, the proposed algorithm replaces $\mathbb{E}_{\pi_\theta}\left[G\right]$ and $\mathbb{E}_{\pi_\theta}\left[G^2\right]$ with the sample approximations $\sum_{t=1}^{n}\gamma^{t-1}r(S_t,A_t)$ and $\left(\sum_{t=1}^{n}\gamma^{t-1}r(S_t,A_t)\right)^2$, respectively (Tamar et al., 2012); that is, the unbiased gradients are given as (1). Therefore, for a sample approximation $\widehat{G}^k$ of $\mathbb{E}_{\pi_\theta}\left[G^2\right]$ at the episode k, we optimize the policy with ascending the unbiased gradient

$$\widehat{\nabla}\mathbb{E}_{\pi_\theta}\left[u^{\text{trajectory}}(G;\alpha,\beta)\right] = \left(\alpha\widehat{G}^k - \frac{1}{2}\beta\left(\widehat{G}^k\right)^2\right)\sum_{t=1}^{n}\nabla_\theta\log\pi_\theta(S_t^k, A_t^k).$$

We summarize the algorithm as the pseudo-code in Algorithm 1.

Here, we present three advantages of EQUMRL.The first advantage concerns computation. In EQUMRL, we do not suffer from the double sampling issue because the term $\left(\mathbb{E}_{\pi_\theta}[G]\right)^2$, which causes the problem, is absent, and we must only estimate the gradients of $\mathbb{E}_{\pi_\theta}\left[G\right]$ and $\mathbb{E}_{\pi_\theta}\left[G^2\right]$. The second advantage is that it provides a variety of interpretations, as listed in Section 4.3.

The third advantage is the ease of theoretical analysis. For example, referring to the results of Bertsekas & Tsitsiklis (1996), we derive the following result on the convergence of the gradient, which can be applied to simple policy gradient algorithms, not only our proposed REINFORCE-based algorithm.

Algorithm 1 REINFORCE-based EQUMRL

Initialize the policy parameter θ_0;
for $k = 1, 2, \ldots$ **do**
  Generate $\{(S_t^k, A_t^k, r(S_t^k, A_t^k))\}_{t=1}^n$ on π_θ;
  Update policy parameters $\theta_{k+1} \leftarrow \theta_k + \eta_i \widehat{\nabla} \mathbb{E}_{\pi_{\theta_k}} \left[u^{\mathrm{trajectory}}(G; \alpha, \beta) \right]$.
end for

Theorem 1. *Consider an update rule such that* $\theta_{k+1} \leftarrow \theta_k + \eta_i \widehat{\nabla} \mathbb{E}_{\pi_{\theta_k}} \left[u^{\mathrm{trajectory}}(G; \alpha, \beta) \right]$*, where the learning rates* η_k *are non-negative and satisfy* $\sum_{k=0}^{\infty} \eta_k = \infty$ *and* $\sum_{k=0}^{\infty} \eta_k^2 < \infty$*. Suppose that (a) episode always finishes in finite horizon* n*; (b) the policy* π_θ *has always bounded first and second partial derivatives. Then,* $\lim_{i \to \infty} \nabla_\theta \mathbb{E}_{\pi_{\theta_i}} [\alpha G - \frac{1}{2} \beta G^2] = 0$ *almost surely.*

It is expected that non-asymptotic results can be derived by restricting the policy class and the optimization algorithm as Agarwal et al. (2020) and Zhang et al. (2021a), but this is not the scope of this paper, which aims to provide a general framework.

AC-based trajectory EQUMRL. Another implementation of the EQUMRL is to apply the AC algorithm (Williams & Peng, 1991; Mnih et al., 2016). For an episode k with the length n, following Prashanth & Ghavamzadeh (2013; 2016), we train the policy by a gradient defined as

$$\nabla_\theta \log \pi_\theta(S_t^k, A_t^k) \left\{ \left(\alpha \widetilde{G}_{t:t+n-1}^k - \frac{1}{2} \beta \widetilde{G}_{t:t+n-1}^{k,2} \right) - \left(\alpha M_{\hat{\omega}_k^{(1)}}^{(1)}(S_t^k) - \frac{1}{2} \beta M_{\hat{\omega}_k^{(2)}}^{(2)}(S_t^k) \right) \right\},$$

where $\widetilde{G}_{t:t+n-1}^k = \widehat{G}_{t:t+n-1}^k + \gamma^n M_{\hat{\omega}_k^{(1)}}^{(1)}(S_{t+n}^k)$, $\widetilde{G}_{t:t+n-1}^{k,2} = \left(\widehat{G}_{t:t+n-1}^k \right)^2 + 2\gamma^n \widehat{G}_{t:t+n-1}^k M_{\hat{\omega}_k^{(1)}}^{(1)}(S_{t+n}^k) + \gamma^{2n} M_{\hat{\omega}_k^{(2)}}^{(2)}(S_{t+n}^k)$, and $M_{\hat{\omega}_k^{(1)}}^{(1)}(S_t^k)$ and $M_{\hat{\omega}_k^{(2)}}^{(2)}(S_t^k)$ are models of $\mathbb{E}[G_{t+1:\infty}]$ and $\mathbb{E}[G_{t+1:\infty}^2]$ with parameters $\hat{\omega}_k^{(1)}$ and $\hat{\omega}_k^{(2)}$. For more details, see Prashanth & Ghavamzadeh (2013; 2016).

Remark 1 (Existing approaches)**.** For the double sampling issue, Tamar et al. (2012) and Prashanth & Ghavamzadeh (2013; 2016) proposed multi-time-scale stochastic optimization. Their approaches are known to be sensitive to the choice of step-size schedules, which are not easy to control (Xie et al., 2018). Xie et al. (2018) proposed using the Legendre-Fenchel dual transformation with coordinate descent algorithm. First, based on Lagrangian relaxation, Xie et al. (2018) set an objective function as $\max_{\theta \in \Theta} \mathbb{E}_{\pi_\theta}[G] - \delta \left(\mathbb{V}_{\pi_\theta}(G) - \eta \right)$. Then, Xie et al. (2018) transformed the objective function as $\max_{\theta \in \Theta, y \in \mathbb{R}} 2y \left(\mathbb{E}_{\pi_\theta}[G] + \frac{1}{2\delta} \right) - y^2 - \mathbb{E}_{\pi_\theta} \left[G^2 \right]$ and trained an agent by solving the optimization problem via a coordinate descent algorithm. However, this approach does not reflect the constraint η because the constraint condition η vanishes from the objective function. This problem is caused by their objective function based on the penalty function $g(x) = x$: $\mathbb{E}_{\pi_\theta}[G] - \delta \left(\mathbb{V}_{\pi_\theta}(G) - \eta \right)$, where the first derivative does not include η. To avoid this problem, we need an iterative algorithm to decide an optimal δ or change $g(x)$ from x but it is not obvious how to incorporate them into the approach.

Remark 2 (Difference from existing MV approaches)**.** Readers may assert that EQUMRL simply omits $\left(\mathbb{E}_{\pi_\theta}[G] \right)^2$ from existing MV-controlled RL methods, which usually includes the explicit variance term in the objective function, and is the essentially the same. However, there are significant differences; one of the main findings of this paper is our formulation of a simpler RL problem to obtain an MV-efficient policy. Existing MV-controlled RL methods suffer from computational difficulties caused by the double sampling issue. However, we can obtain MV-efficient policy without going through the difficult problem. In addition, EQUMRL shows better performance in experiments even from the viewpoint of the constrained problem because it is difficult to choose parameters to avoid the double sampling issue in existing approaches. Thus, the EQUMRL has advantage in avoiding solving more difficult constrained problems for considering MV trade-off.

4.3 Interpretations of EQUMRL with Gradeint Estimation

We can interpret EQUMRL as an approach for (i) a targeting optimization problem to achieve an expected cumulative reward ζ, (ii) an expected cumulative reward maximization with regularization, and (iii) expected utility maximization via Taylor approximation.

First, we can also interpret EQUMRL as mean squared error (MSE) minimization between a cumulative reward R and a target return ζ; that is,

$$\arg\min_{\theta\in\Theta} J(\theta;\zeta) = \arg\min_{\theta\in\Theta} \mathbb{E}_{\pi_\theta}\left[\left(\zeta - G\right)^2\right] \tag{4}$$

We can decompose the MSE into the bias and variance as

$$\begin{aligned}\mathbb{E}_{\pi_\theta}\left[\left(\zeta - G\right)^2\right] &= \underbrace{\left(\zeta - \mathbb{E}_{\pi_\theta}[G]\right)^2}_{\text{Bias}} + \underbrace{2\mathbb{E}_{\pi_\theta}\left[\left(\zeta - \mathbb{E}_{\pi_\theta}[G]\right)\left(\mathbb{E}_{\pi_\theta}[G] - G\right)\right]}_{0} + \underbrace{\mathbb{E}_{\pi_\theta}\left[\left(\mathbb{E}_{\pi_\theta}[G] - G\right)^2\right]}_{\text{Variance}} \\ &= \zeta^2 - 2\zeta\mathbb{E}_{\pi_\theta}[G] + \left(\mathbb{E}_{\pi_\theta}[G]\right)^2 + \mathbb{E}_{\pi_\theta}\left[\left(\mathbb{E}_{\pi_\theta}[G] - G\right)^2\right].\end{aligned}$$

Thus, the MSE minimization (4) is equivalent to EQUMRL (3), where $\zeta = \frac{\alpha}{\beta}$. The above equation provides an important implication for the setting of ζ. If we know the reward is shifted by x, we only have to adjust ζ to $\zeta + x$. This is because ζ only affects the bias term in the above equation. The equation also provides another insight. If our assumption $\max_{\pi_\theta} \mathbb{E}_{\pi_\theta}[G] \leq \zeta$ is violated, $\mathbb{E}_{\pi_\theta}[G]$ will not be maximized with a fixed variance and will be biased towards ζ; that is, EQUMRL cannot find the MV-efficient policies. In applications, we can confirm whether the optimization works by checking whether average value of the empirically realized cumulative rewards is less than ζ.

Second, we can regard the quadratic utility function as an expected cumulative reward maximization with a regularization term defined as $\mathbb{E}\left[R^2\right]$; that is, minimization of the risk $\mathcal{R}(\pi_\theta)$:

$$\mathcal{R}(\theta) = \underbrace{-\mathbb{E}_{\pi_\theta}[G]}_{\text{Risk of expected cumulative reward maximization}} + \underbrace{\psi\mathbb{E}_{\pi_\theta}\left[G^2\right]}_{\text{Regularization term}}$$

where $\psi > 0$ is a regulation parameter and $\psi = \frac{\beta}{2\alpha} = \frac{1}{2\zeta}$. As $\psi \to 0$ ($\zeta \to \infty$), $\mathcal{R}(\theta) \to -\mathbb{E}_{\pi_\theta}[G]$.

Third, the quadratic utility function is the quadratic Taylor approximation of a smooth utility function $u(G)$ because for $G_0 \in \mathbb{R}$, we can expand it as $u(G) \approx u(G_0) + U'(G_0)(G - G_0) + U''(G_0)(G - G_0)^2 + \cdots$; that is, quadratic utility is an approximation of various risk-averse utility functions. This property also supports the use of the quadratic utility function in practice (Kroll et al., 1984).

It also should be noted that the EQUMRL is closely related to the fields of economics and finance, where the ultimate goal is to maximize the utility of an agent, which is also referred to as an *investor*. The quadratic utility function is a standard risk-averse utility function often assumed in financial theory Luenberger et al. (1997) to justify an MV portfolio; that is, an MV portfolio maximizes an investor's utility function if the utility function is quadratic (see Appendix A). Therefore, our approach can be interpreted as a method that directly achieves the original goal.

Remark 3 (Specification of utility function (hyper-parameter selection))**.** Next, we discuss how to decide the parameters α and β, which are equivalent to ζ, ξ, and ψ. The meanings are equivalent to constrained conditions of MVRL; that is, we predetermine these hyperparameters depending on our attitude toward risk. For instance, we propose the following three directions for the parameter choice. First, we can determine $\frac{\beta}{2\alpha} = \psi$ based on economic theory or market research (Ziemba et al., 1974; Kallberg et al., 1983) (Appendix A). Luenberger et al. (1997) proposed some questionnaires to investors for the specification. Second, we set $\zeta = \frac{1}{2\psi}$ as the targeted reward that investors aim to achieve. Third, through cross-validation, we can optimize the regularization parameter ψ to maximize some criteria, such as the Sharpe ratio (Sharpe, 1966). However, we note that in time-series related tasks, we cannot use standard cross-validation owing to dependency. Therefore, in our experiments with a real-world financial dataset, we show various results under different parameters.

Remark 4 (From MV-efficient RL to MV-controlled RL)**.** We can also apply the MV-efficient RL method as the MV-controlled RL method. The parameters of the expected quadratic utility function correspond to the variance that we want to achieve. For example, the larger ζ is in (4), the larger the variance will be. Although we do not know the explicit correspondence between the parameter of the expected quadratic utility function and the variance, it is possible to control the variance by choosing an appropriate policy from among those learned under several parameters. In Figure 2 in Section 6, we show the mean and variance of several policies trained with several parameters. These results are measured with test data, but by outputting multiple candidates with the training data (ex. cross-validation), we can choose a policy with the desired variance.

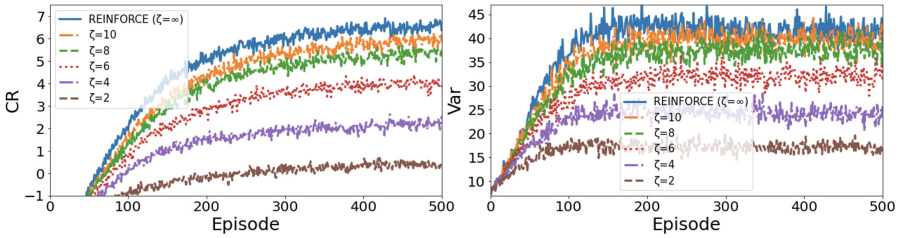

Figure 1: The CRs and Vars in the training process of the experiment using the synthetic dataset.

5 EQUMRL UNDER PER-STEP VARIANCE PERSPECTIVE

Another direction of MVRL is to consider the trade-off between the per-step mean $\mathbb{E}_{\pi_\theta}[r(S_t, A_t)]$ and per-step variance $\mathbb{V}_{\pi_\theta}(r(S_t, A_t))$, $\mathbb{E}_{\pi_\theta}[r(S_t, A_t)] \overset{\text{trade-off}}{\Longleftrightarrow} \mathbb{V}_{\pi_\theta}(r(S_t, A_t))$. As well as the trajectory perspective, Bisi et al. (2020) and Zhang et al. (2021b) proposed training a policy to maximize the penalized objective function, $\max_{\theta\in\Theta} \sum_{t=1}^{T} \gamma^{t-1}\Big(\mathbb{E}_{\pi_\theta}[r(S_t, A_t)] - \lambda \mathbb{V}_{\pi_\theta}(r(S_t, A_t))\Big)$. We call an algorithm with this formulation a per-step MV-controlled RL. See Appendix B for more details.

We define a per-step MV-efficient policy as a policy π_θ such that there is no other policy with $\theta' \in \Theta$, where, for each t, $\mathbb{E}_{\pi_\theta}[r(S_t, A_t)] \leq \mathbb{E}_{\pi_{\theta'}}[r(S_t, A_t)]$ and $\mathbb{V}_{\pi_\theta}[r(S_t, A_t)] > \mathbb{V}_{\pi_{\theta'}}[r(S_t, A_t)]$, or $\mathbb{V}_{\pi_\theta}[r(S_t, A_t)] \geq \mathbb{V}_{\pi_{\theta'}}[r(S_t, A_t)]$ and $\mathbb{E}_{\pi_\theta}[r(S_t, A_t)] < \mathbb{E}_{\pi_{\theta'}}[r(S_t, A_t)]$. For constants $\alpha, \beta > 0$, we define the quadratic utility function for per-step reward setting as $u^{\text{per-step}}(r(S_t, A_t); \alpha, \beta) = \alpha r(S_t, A_t) - \beta r^2(S_t, A_t)$. For an finite horizon case with $\gamma = 1$, we consider maximizing the expected cumulative quadratic utility function defined as $\mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} u^{\text{per-step}}(r(S_t, A_t); \alpha, \beta)\right] = \alpha \mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} r(S_t, A_t)\right] - \beta \mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} r^2(S_t, A_t)\right]$. When applying REINFORCE-based algorithm to train an agent, the sample approximation of the gradient is given as $\widehat{\nabla}\mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} u^{\text{per-step}}(r(S_t^k, A_t^k); \alpha, \beta)\right] = \sum_{t=1}^{T}\{\alpha r(S_t^k, A_t^k) - \beta\big(r(S_t^k, A_t^k)\big)^2\}\nabla_\theta \log \pi_\theta(S_t^k, A_t^k)$. Appendix B contains the performance using a synthetic dataset.

6 EXPERIMENTS

This section investigates the empirical performance of the proposed EQUMRL in trajectory variance setting using synthetic and real-world financial datasets. We train the policy by REINFORCE-based algorithm. We conduct two experiments. In the first experiments, following Tamar et al. (2012; 2014), and Xie et al. (2018), we conduct portfolio management experiments with synthetic datasets. In the third experiment, we conduct a portfolio management experiment with a dataset of Fama & French (1992), a standard benchmark in finance. In Appendix C, following Tamar et al. (2012; 2014), and Xie et al. (2018), we also conduct American-style option experiments with synthetic datasets. We implemented algorithms following the Pytorch example[1]. For algorithms using neural networks, we use a three layer perceptron, where the numbers of the units in two hidden layers are the same as that of the input node, and that of the output node is 2. We note that in all results, naively maximizing the reward or minimizing the variance do not ensure a better algorithm; we evaluate an algorithm based on how it controls the MV trade-off. We denote the hyperparameter of the EQUMRL by ζ, which has the same meaning as (α, β) and ψ. For all experiments, we adopt episodic MDPs; that is, $\gamma = 1$.

6.1 PORTFOLIO MANAGEMENT WITH A SYNTHETIC DATASET

Following Tamar et al. (2012) and Xie et al. (2018), we consider a portfolio composed of two asset types: a liquid asset with a fixed interest rate r_{l}, and a non-liquid asset with a time-dependent interest rate taking either $r_{\text{nl}}^{\text{low}}$ or $r_{\text{nl}}^{\text{high}}$, and the transition follows a switching probability p_{switch}. An investor can sell the liquid asset at every time step $t = 1, 2, \ldots, T$ but the non-liquid asset can only be sold after the maturity W periods. This means that when holding 1 liquid asset, we obtain r_{l} per period; when holding 1 non-liquid asset at the t-th period, we obtain $r_{\text{nl}}^{\text{low}}$ or $r_{\text{nl}}^{\text{high}}$ at the $t + W$-th period.

[1] https://github.com/pytorch/examples/tree/master/reinforcement_learning

Table 1: The results of the portfolio management with a synthetic data. The upper table shows the CRs and Vars over $100,000$ trials. The lower table shows the MSEs for ζ over $100,000$ trials.

	REINFORCE	EQUM			Tamar		Xie	
	$(\zeta = \infty)$	$\zeta = 10$	$\zeta = 6$	$\zeta = 4$	$\mathbb{V} = 80$	$\mathbb{V} = 50$	$\lambda = 100$	$\lambda = 10$
CR	6.729	6.394	4.106	2.189	6.709	2.851	0.316	0.333
Var	32.551	31.210	24.424	18.518	32.586	21.573	15.883	15.992

	REINFORCE	EQUM				
Target Value	$(\zeta = \infty)$	$\zeta = 10$	$\zeta = 8$	$\zeta = 6$	$\zeta = 4$	$\zeta = 2$
MSE from $\zeta = 10$	**51.669**	53.399	55.890	65.307	83.061	105.492
$\zeta = 8$	**42.586**	42.975	43.375	45.730	55.816	71.480
$\zeta = 6$	41.503	40.551	38.860	**34.154**	36.570	45.467
$\zeta = 4$	48.420	46.127	42.345	30.577	**25.324**	27.455
$\zeta = 2$	63.337	59.703	53.830	35.000	22.078	**17.442**

Besides, the non-liquid asset has a risk of not being paid with a probability p_{risk}; that is, if the non-liquid asset defaulted during the W periods, we could not obtain any rewards by having the asset. An investor can change the portfolio by investing a fixed fraction w of the total capital M in the non-liquid asset at each time step. A typical investment strategy is to construct a portfolio using both liquid and non-liquid assets for decreasing the variance. Following Xie et al. (2018), we set $r_{\text{l}} = 1.001$, $r_{\text{nl}}^{\text{low}} = 1.1$, $r_{\text{nl}}^{\text{high}} = 2$, $p_{\text{switch}} = 0.1$, $p_{\text{risk}} = 0.05$, $W = 4$, $w = 0.2$, and $M = 1$. As a performance metric, we use the average cumulative reward (CR) and its variance (Var) when investing for 50 periods. We compare the EQUMRL with the REINFORCE, MV-controlled methods proposed by Tamar et al. (2012) (Tamar), and Xie et al. (2018) (Xie). We denote the variance constraint of Tamar et al. (2012) as Var and Lagrange multiplier of Xie et al. (2018) as λ. For training Tamar, Xie, and EQUMRL, we set the Adam optimizer with learning rate 0.01 and weight decay parameter 0.1. For each algorithm, we report performances under various hyperparameters as much as possible.

First, we show CRs and Vars of the EQUMRL during the training process in Figure 1, where we conduct 100 trials on the test environment to compute CRs and Vars for each episode. Here, we also show the performance of REINFORCE, which corresponds to the EQUMRL with $\zeta = \alpha/\beta = \infty$. As Figure 1 shows, the EQUMRL trains MV-efficient policies well depending on the parameter ζ. Next, we compare the EQUMRL on the test environment with the REINFORCE, Tamar, and Xie. We conduct $100,000$ trials on the test environment to compute CRs and Vars. In Figure 2, we plot performances under several hyperparameters, where the horizontal axis denotes the Var, and the vertical axis denotes the CR. Trained agents with a higher CR and lower Var are Pareto efficient. As the result shows, the EQUMRL returns more efficient portfolios than the others in almost all cases. We conjecture that this is because while the EQUMRL is an end-to-end optimization for obtaining an efficient agent, the other methods consist of several steps for solving the constrained optimization, where those multiple steps can be sources of the suboptimal result. We show CRs and Vars of some of their results in the upper table of Table 1. The MSEs between ζ and CR are also shown in the lower table of Table 1, where we can confirm that the EQUMRL succeeded in minimizing the MSEs.

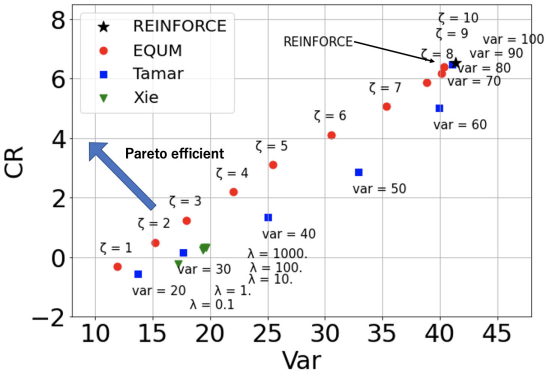

Figure 2: MV efficiency of the portfolio management experiment. Higher CRs and lower Vars methods are MV Pareto efficient.

6.2 Portfolio management with a real-world dataset

We use standard benchmarks called Fama & French (FF) datasets[2] to ensure the reproducibility (Fama & French, 1992). Among FF datasets, we use the FF25, FF48 and FF100 datasets, where the FF25 and FF 100 dataset includes 25 and 100 assets formed based on size and book-to-market ratio; the FF48 dataset contains 48 assets representing different industrial sectors. We use all datasets covering

[2] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Table 2: The performance of each portfolio model during the out-of-sample period (from July 2000 to June 2020) for FF25 dataset (upper table) , FF48 (middle table), and FF100 (lower table). For each dataset, the best performance is highlighted in **bold**.

Method	EW	MV	EGO	BLD	Tamar			Xie			EQUM		
					$\mathbb{V}=15$	$\mathbb{V}=30$	$\mathbb{V}=60$	$\lambda=10$	$\lambda=100$	$\lambda=1000$	$\zeta=0.5$	$\zeta=0.75$	$\zeta=1.5$
FF25													
CR↑	0.80	0.09	0.81	0.84	1.15	1.01	0.80	0.90	1.15	1.15	**1.53**	1.25	1.11
Var↓	28.62	53.57	30.65	22.16	18.53	15.04	25.29	25.96	25.77	25.77	24.27	15.02	**11.77**
R/R↑	0.52	0.04	0.51	0.62	0.93	0.90	0.55	0.61	0.79	0.79	1.07	1.12	**1.13**
MaxDD↓	0.54	0.75	0.58	0.52	0.35	0.35	0.56	0.54	0.51	0.51	0.36	0.31	**0.27**
FF48													
CR↑	0.81	0.11	0.97	0.75	0.68	0.38	0.82	0.50	1.08	1.01	**1.60**	1.05	0.91
Var↓	22.91	77.02	31.91	15.98	40.69	16.00	18.04	27.12	23.77	26.31	31.97	18.69	**10.34**
R/R↑	0.59	0.04	0.60	0.65	0.37	0.33	0.67	0.33	0.76	0.68	**0.98**	0.84	0.98
MaxDD↓	0.30	0.48	0.31	0.25	0.32	0.20	0.21	0.27	0.25	0.26	0.29	0.21	**0.16**
FF100													
CR↑	0.81	0.11	0.81	0.85	1.13	0.79	0.67	0.57	0.98	1.24	0.95	0.95	**1.43**
Var↓	29.36	57.97	32.35	21.83	25.50	20.34	16.50	87.92	28.78	41.06	15.50	**14.26**	33.09
R/R↑	0.52	0.05	0.49	0.63	0.77	0.61	0.57	0.21	0.63	0.67	0.83	**0.87**	0.86
MaxDD↓	0.33	0.46	0.34	0.27	0.25	0.23	0.25	0.51	0.32	0.31	**0.19**	0.26	0.31

monthly data from July 1980 to June 2020. Consider an episodic MDP. Let the action space be the set of m assets, and $y_{a,t}$ and $w_{a,t}$ be the return and the portfolio weight of an asset a at time t. The reward (portfolio return) at time $1 \leq t \leq T$ is defined as $y_t = \sum_{a=1}^{m} y_{a,t} w_{a,t} - \Lambda \sum_{a=1}^{m} |w_{a,t} - w_{a,t-1}|$, where $\Lambda = 0.001$ is the penalty of the portfolio weight turnover(change of the action). On the t-th period, the agent observes a state $S_t = ((y_{a,t-1}, ..., y_{a,t-12}, w_{a,t-1})_{a \in \mathcal{A}}, \sum_{s=0}^{t-1} y_s)$, and decides a portfolio weight $(w_{a,t})_{a \in \mathcal{A}}$ as $w_{a,t} = \pi(a, s_t)$. Between periods t and $t+1$, the agent has an asset a with the ratio $w_{a,t}$.See Section 4.1 for the reason why we include $sum_{s=0}^{t-1} y_s$ in the state.

We use the following portfolio models: an equally-weighted portfolio (EQ, DeMiguel et al., 2009); a mean-variance portfolio (MV, Markowitz, 1952); a Kelly growth optimal portfolio Shen et al. (2019); Portfolio blending via Thompson sampling (BLD, Shen & Wang, 2016). We also compare our proposed method with the methods proposed by Tamar et al. (2012) and Xie et al. (2018). Denote a method proposed by Tamar et al. (2012) by Tamar, and choose the parameter var from $\{15, 30, 60\}$. Denote a method proposed by Xie et al. (2018) by Xie, and choose the parameter λ from $\{10, 100, 1000\}$. Denote the REINFORCE-based trajectory EQUMRL by EQUMRL, and choose the parameter ζ from $\{0.5, 0.75, 1.5\}$.

We apply the following standard measures in finance for evaluation (Brandt, 2010). The cumulative reward (CR), annualized risk as the standard deviation of return (RISK) and risk-adjusted return (R/R) are defined as follows: $\mathrm{CR} = 1/T \sum_{t=1}^{T} y_t$, $\mathbb{V} = 1/T \sum_{t=1}^{T} (y_t - \mathrm{CR})^2$, and $\mathrm{R/R} = \sqrt{12} \times \mathrm{CR}/\sqrt{\mathrm{Var}}$. R/R is the most important measure for a portfolio strategy and is often referred to as the Sharpe Ratio (Sharpe, 1966). We also evaluate the maximum draw-down (MaxDD), which is another widely used risk measure (Magdon-Ismail & Atiya, 2004). In particular, MaxDD is the largest drop from a peak defined as $\mathrm{MaxDD} = \min_{t \in [1,T]} \left(0, \frac{W_t}{\max_{\tau \in [1,t]} W_\tau} - 1\right)$, where W_k is the cumulative return of the portfolio until time k; that is, $W_t = \prod_{t'=1}^{t} (1 + y_{t'})$.

Table 2 reports the performances of the portfolios. In almost all cases, the EQUMRL achieves the highest R/R and the lowest MaxDD. Therefore, we can confirm that the EQUMRL has a high R/R, and avoids a large drawdown. The real objective (minimizing variance with a penalty on return targeting) for Tamar, MVP, and EQUMRL is shown in Appendix D. Except for FF48's MVP, the objective itself is smaller than that of EQUMRL. Since the values of the objective are proportional to R/R, we can empirically confirm that the better optimization, the better performance.

7 Conclusion

In this paper, we proposed EQUMRL for MV-efficient RL. Compared with the conventional MVRL methods, EQUMRL is computationally friendly. The proposed EQUMRL also includes various interpretations, such as targeting optimization and regularization, which expands the scope of applications of the method. We investigated the effectiveness of the EQUMRL compared with the standard RL and existing MVRL methods through experiments using synthetic and real-world datasets.

REFERENCES

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 64–66, 09–12 Jul 2020.

Baron, D. P. On the utility theoretic foundations of mean-variance analysis. *The Journal of Finance*, 32, 1977.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Bertsekas, D. P. and Tsitsiklis, J. N. Neuro-dynamic programming: An overview. In *Proceedings of 1995 34th IEEE Conference on Decision and Control*, volume 1, pp. 560–564. IEEE, 1995.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-dynamic programming.* Athena Scientific, Belmont, MA, 1996.

Bisi, L., Sabbioni, L., Vittori, E., Papini, M., and Restelli, M. Risk-averse trust region optimization for reward-volatility reduction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020.

Bodnar, T., Parolya, N., and Schmid, W. On the exact solution of the multi-period portfolio choice problem for an exponential utility under return predictability. *European Journal of Operational Research*, 246(2):528–542, 2015a.

Bodnar, T., Parolya, N., and Schmid, W. A closed-form solution of the multi-period portfolio choice problem for a quadratic utility function. *Annals of Operations Research*, 229(1):121–158, 2015b.

Bodnar, T., Okhrin, Y., Vitlinskyy, V., and Zabolotskyy, T. Determination and estimation of risk aversion coefficients. *Computational Management Xcience*, 15(2):297–317, 2018.

Borch, K. A Note on Uncertainty and Indifference Curves. *The Review of Economic Studies*, 1969.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Brandt, M. W. Portfolio choice problems. In *Handbook of Financial Econometrics: Tools and Techniques*, pp. 269–336. Elsevier, 2010.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym, 2016.

Bulmuş, T. and Özekici, S. Portfolio selection with hyperexponential utility functions. *OR spectrum*, 36(1):73–93, 2014.

Chow, Y. and Ghavamzadeh, M. Algorithms for cvar optimization in mdps. *Advances in Neural Information Processing Systems*, 27:3509–3517, 2014.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

DeMiguel, V., Garlappi, L., and Uppal, R. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009.

Fama, E. F. and French, K. R. The cross-section of expected stock returns. *The Journal of Finance*, 1992.

Fishburn, P. C. and Burr Porter, R. Optimal portfolios with one safe and one risky asset: Effects of changes in rate of return and risk. *Management science*, 22(10):1064–1073, 1976.

Garcıa, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Geibel, P. and Wysotzki, F. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

Kallberg, J., Ziemba, W., et al. Comparison of alternative utility functions in portfolio selection problems. *Management Science*, 29(11):1257–1276, 1983.

Kroll, Y., Levy, H., and Markowitz, H. M. Mean-variance versus direct utility maximization. *The Journal of Finance*, 39(1):47–61, 1984.

Lintner, J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1):13–37, 1965.

Luenberger, D. G. et al. Investment science. *OUP Catalogue*, 1997.

Magdon-Ismail, M. and Atiya, A. F. Maximum drawdown. *Risk Magazine*, 2004.

Markowitz, H. Portfolio selection. *The Journal of Finance*, 1952.

Markowitz, H. *Portfolio selection: efficient diversification of investments*. Yale university press, 1959.

Markowitz, H. M. and Todd, G. P. *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, volume 66. John Wiley & Sons, 2000.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. *Microeconomic Theory*, volume 1. Oxford university press New York, 1995.

Merton, R. C. Lifetime portfolio selection under uncertainty: The continuous-time case. *The Review of Economics and Statistics*, pp. 247–257, 1969.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pp. 1928–1937, 2016.

Morgenstern, O. and Von Neumann, J. *Theory of games and economic behavior*. Princeton University Press, 1953.

Mossin, J. Equilibrium in a capital asset market. *Econometrica: Journal of the econometric society*, pp. 768–783, 1966.

Prashanth, L. and Ghavamzadeh, M. Actor-critic algorithms for risk-sensitive mdps. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*, pp. 252–260, 2013.

Prashanth, L. and Ghavamzadeh, M. Variance-constrained actor-critic algorithms for discounted and average reward mdps. *Machine Learning*, 105(3):367–417, 2016.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.

Sharpe, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442, 1964.

Sharpe, W. F. Mutual fund performance. *The Journal of Business*, 1966.

Shen, W. and Wang, J. Portfolio blending via thompson sampling. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1983–1989, 2016.

Shen, W., Wang, B., Pu, J., and Wang, J. The kelly growth optimal portfolio with ensemble learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1134–1141, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pp. 1057–1063, 2000.

Tamar, A., Di Castro, D., and Mannor, S. Policy gradients with variance related risk criteria. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1651–1658, 2012.

Tamar, A., Mannor, S., and Xu, H. Scaling up robust mdps using function approximation. In *International Conference on Machine Learning*, pp. 181–189, 2014.

Tobin, J. Liquidity preference as behavior towards risk. *The Review of Economic Studies*, 25(2): 65–86, 1958.

Williams, R. *Toward a Theory of Reinforcement-learning Connectionist Systems*. Northeastern University, 1988.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Williams, R. J. and Peng, J. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

Xie, T., Liu, B., Xu, Y., Ghavamzadeh, M., Chow, Y., Lyu, D., and Yoon, D. A block coordinate ascent algorithm for mean-variance optimization. *Advances in Neural Information Processing Systems*, 31:1065–1075, 2018.

Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. Sample efficient reinforcement learning with reinforce. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10887–10895, May 2021a.

Zhang, S., Liu, B., and Whiteson, S. Mean-variance policy iteration for risk-averse reinforcement learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021b.

Ziemba, W., Parkan, C., and Brooks-Hill, R. Calculation of investment portfolios with risk free borrowing and lending. *Management Science*, 21(2), 1974.