

# A unified algorithm framework for mean-variance optimization in discounted Markov decision processes

Shuai Ma<sup>a</sup>, Xiaoteng Ma<sup>b</sup>, Li Xia<sup>a,\*</sup>

<sup>a</sup>*School of Business, Sun Yat-sen University, Guangzhou, 510275, P. R. China*

<sup>b</sup>*Department of Automation, Tsinghua University, Beijing, 100086, P. R. China*

---

## Abstract

This paper studies the risk-averse mean-variance optimization in infinite-horizon discounted Markov decision processes (MDPs). The involved variance metric concerns reward variability during the whole process, and future deviations are discounted to their present values. This discounted mean-variance optimization yields a reward function dependent on a discounted mean, and this dependency renders traditional dynamic programming methods inapplicable since it suppresses a crucial property—time consistency. To deal with this unorthodox problem, we introduce a pseudo mean to transform the untreatable MDP to a standard one with a redefined reward function in standard form and derive a discounted mean-variance performance difference formula. With the pseudo mean, we propose a unified algorithm framework with a bilevel optimization structure for the discounted mean-variance optimization. The framework unifies a variety of algorithms for several variance-related problems including, but not limited to, risk-averse variance and mean-variance optimizations in discounted and average MDPs. Furthermore, the convergence analyses missing from the literature can be complemented with the proposed framework as well. Taking the value iteration as an example, we develop a discounted mean-variance value iteration algorithm and prove its convergence to a local optimum with the aid of a Bellman local-optimality equation. Finally, we conduct a numerical experiment on

---

\*Corresponding author. Email: xiali5@sysu.edu.cn

Email addresses: mash35@mail.sysu.edu.cn (Shuai Ma),  
ma-xt17@mails.tsinghua.edu.cn (Xiaoteng Ma)

portfolio management to validate the proposed algorithm.

*Keywords:* Dynamic programming, Markov decision process, discounted mean-variance, bilevel optimization, Bellman local-optimality equation

---

## 1. Introduction

Financial optimizations usually involve trade-offs between profit and risk, and variance is to risk what mean is to profit. It could be the reason why the mean-variance optimization theory initiated by Markowitz (1952) is one of the most prevalent financial optimization frameworks. As a cornerstone of the modern portfolio theory, the mean-variance optimization theory has been extensively applied in a variety of financial problems, such as portfolio selection (Best and Grauer 1991), hedging (Kouvelis et al. 2018), pricing (Kandel et al. 1989), etc. It appeals to both academia and industry not only for its simplicity but also for being a satisfactory proxy for other types of risk minimization rules. Levy and Levy (2003) show that when diversification between assets is allowed, the mean-variance optimization theory and the prospect theory (Tversky and Kahneman 1992) almost coincide, which justifies its robustness. One stream of works on the mean-variance optimization is from the perspective of stochastic control (Li and Ng 2000, Basak and Chabakauri 2010, Zhang et al. 2012). In this paper, we study it from the perspective of Markov decision processes (MDPs), where it is often assumed that the state and action spaces are finite and the reward is bounded in a discrete-time scenario.

In the framework of MDPs, a variety of works (Sobel 1982, Tamar et al. 2012, Xie et al. 2018) concern the variance of total discounted reward, or *return*, i.e.,  $\mathbb{V}\{\sum_{t=1}^{\infty} \alpha^{t-1} r(X_t)\}$ , given  $\alpha$  the discount factor and  $r(X_t)$  the immediate reward at time epoch  $t$ . Its counterpart in average MDPs is termed the limiting average variance, which is defined by  $\lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}\{(\sum_{t=1}^T r(X_t) - \mathbb{E}\{\sum_{t=1}^T r(X_t)\})^2\}$  (Hernández-Lerma et al. 1999). These two variance metrics focus on the fluctuation of cumulative reward at the final epoch. However, it is seemingly preferable to consider risks at every time point in many practical

problems. For example, an autonomous driving system should pay attention to every detail along the road to ensure safe driving. In finance, it is easy to manipulate a stock price at a specified time point, but it is barely possible to do it during the whole process.

Motivated by the above observations, we consider a steady-state variance metric in the mean-variance optimization. The steady-state variance is also known as the long-run variance (Filar et al. 1989), which is defined by  $\lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}\{\sum_{t=1}^T (r(X_t) - \eta_a)^2\}$ . It quantifies the dispersion of immediate rewards from the long-run average  $\eta_a = \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}\{\sum_{t=1}^T r(X_t)\}$  by *averaging* the reward deviations during the whole process. The difference between the limiting average variance and the steady-state variance is discussed by Xia (2016). Different from the classic definition, we introduce a discount factor in the steady-state variance, and optimize the discounted mean-variance objective  $(\eta - \beta\zeta)$ , where  $\eta$  is the normalized discounted mean,  $\beta$  is a risk-aversion parameter, and  $\zeta = \mathbb{E}\{\sum_{t=1}^{\infty} \alpha^{t-1} (r(X_t) - \eta)^2\}$  is the *discounted steady-state variance*. The motivation of involving the discount factor in the variance metric is twofold. First, we calculate the *present* risk (deviation) value from a monetary point of view, which renders risks at multiple epochs comparable with a consistent measure. Second, a discounted variance puts more emphasis (weights  $\alpha^{t-1}$ ) on the transient behavior at the beginning of the process, which may account for a property of the risk: future risks are less critical than current one. Moreover, it is more tractable both computationally and analytically with a discount factor from a mathematical viewpoint.

For the steady-state variance, Filar et al. (1989) illustrate the reasonability of the steady-state variance in average MDPs with a simple example. They point out that when variance is concerned, the discounted problem is more difficult to analyze than its average counterpart. Furthermore, they model the two problems as convex quadratic programs and prove the existence of deterministic optimal policies. Sobel (1994) and Chung (1994) independently analyze the variance optimization problem with a mean performance constraint in unichain MDPs. With the aid of the extant theory on quasiconcave minimization, the

problem can be transformed to a linear program, and the relevant properties and Pareto optimality are studied. Prashanth and Ghavamzadeh (2013) propose actor-critic algorithms to estimate policy gradients of the return variance in discounted MDPs and the steady-state variance in average MDPs. With the ordinary differential equation approach, they prove the asymptotic local convergences of the algorithms. Gosavi (2014) proposes a model-free algorithm analogous to Q-learning for the mean-variance problem in average reinforcement learning (RL). The algorithm is validated with a numerical experiment, but the convergence analysis is missing. This gap is filled by Xia (2016), who proposes a policy iteration for variance minimization in average MDPs, regardless of the mean performance. With the aid of the sensitivity-based optimization theory (Cao 2007), Xia derives a variance performance difference formula (PDF), which quantifies the variance difference between MDPs under any two policies. With the PDF, the local convergence of the proposed policy iteration is proved. This work is later extended to the mean-variance optimization in average MDPs (Xia 2020). Bisi et al. (2020) study the discounted mean-variance in RL, where the steady-state variance is evaluated to bound the limiting average variance. They develop a gradient-based trust region policy optimization (originally proposed by Schulman et al. (2015)) algorithm with a monotonic policy improvement. Zhang et al. (2021) focus on the mean-variance optimization in both discounted and average MDPs, and develop a policy iteration algorithm and a gradient-based RL algorithm. To deal with the policy-dependent reward function, they reformulate variance with its Legendre-Fenchel dual with an extra variable introduced. Virtually, this variance reformulation is similar to the ones with pseudo variances defined by Xia (2016, 2020), where more detailed analyses are given on variance and mean-variance optimization problem equivalences and local convergences in average MDPs, respectively.

Two problems emerge from the literature review. One is that the mean-variance optimization has been studied in discounted and average MDPs separately, and the relationship between the two cases has not been revealed. The other problem is that various optimization algorithms are proposed without con-

vergence analyses, which is partially because the variance-related criteria do not fit in with a standard MDP model. Variance is a quadratic function of mean, and a variance-related problem is equivalent to an MDP with a special reward function, whose value for each state depends on policy instead of action, i.e., the variance value function at a current state will be affected by actions chosen at not only the current epoch but also future epochs. This dependency deprives the time-consistency property and revokes traditional dynamic programming (DP) methods (Puterman 2005, Eckstein et al. 2016, Bisi et al. 2020). In other words, the Bellman optimality equation does not optimize over the admissible action set for a state  $x$  ( $\max_{a \in A(x)}$ ), but over the policy space for the whole state space  $S$  ( $\max_{d \in D}$ ), and we can not divide and conquer this problem in a traditional manner. Most of the relevant studies resort to program-based or gradient-based methods, but the first type of methods cannot deal with problems with large state and action spaces, and the second usually suffers from intrinsic deficiencies: slow convergence, large variance of gradient estimates, and sensitivity to step sizes (Zhao et al. 2012). Xia (2016, 2020) proposes policy iterations for the risk-averse variance and mean-variance optimizations in average MDPs, which offer a new perspective for variance-related problems.

In this paper, we first unify the two mean-variance optimization problems with the continuity property of the discounted mean-variance metric, and we show that the average problem formulation can be viewed as a special case when  $\alpha \uparrow 1$ . For details see Remark 1. This formulation unification offers a systematic perspective in contrast to the previous works. To deal with the difficulty caused by policy dependency, we analyze it in the theory of sensitivity-based optimization (Cao 2007), which stems from the perturbation analysis theory (Ho and Cao 1991) and has been largely extended to stochastic dynamic systems including Markov models. This theory is applicable to general Markov systems, even including the cases without the time-consistency property. With the sensitivity-based optimization theory, we derive a discounted mean-variance PDF, based on which we propose a unified algorithm framework with a bilevel optimization structure, where the inner problem concerns a standard MDP with a fixed

pseudo mean, and the outer problem refers to a one-dimensional optimization of the pseudo mean. Different algorithms can be developed with convergence analyses, which are crucial for the risk-averse variance-related problems. Taking the value iteration as an example, we propose a discounted mean-variance value iteration (DMVVI) algorithm and prove its local convergence. In addition, we present a Bellman local-optimality equation, which presents a necessary and sufficient condition for local optimality of a policy. Finally, we apply the DMVVI algorithm to a portfolio management problem and illustrate its validity.

The contributions of our paper are twofold. First, we present a unified algorithm framework for the risk-averse (mean-)variance optimization problem in discounted and average MDPs. This unified framework can unify the algorithms in relevant works and provide a new perspective for the dynamic optimization concerning steady-state variance metrics. Second, we develop a DMVVI algorithm and prove its convergence, which can provide a foundation for further developing efficient temporal-difference learning methods, such as Q-learning, SARSA (Sutton and Barto 2018), and RL with neural networks, to variance-related optimization problems. We believe that the algorithm framework and the DMVVI algorithm can complement the steady-state variance optimization theory together with the existing works of policy iterations (Xia 2016, 2020).

The remainder of the paper proceeds as follows. Section 2 formulates the risk-averse mean-variance optimization in discounted MDPs. Section 3 proposes a unified algorithm framework with a bilevel optimization structure. Several algorithms can be developed in this framework, and we propose a DMVVI as an example and prove its convergence. Section 4 gives a numerical experiment on financial dynamic portfolio management to validate the DMVVI algorithm. Section 5 presents concluding comments.

## 2. Problem formulation

In this paper, we focus on infinite-horizon discrete-time MDPs, which can be represented by  $\mathcal{M} = \langle S, A, r, p, \mu, \alpha \rangle$ , in which  $S = \{s_1, s_2, \dots, s_{|S|}\}$  is a finite

state space, and  $X_t \in S$  represents the state at (decision) epoch  $t \in \mathbb{N}^+ = \{1, 2, \dots\}$ ;  $A(x)$  is the admissible action set for  $x \in S$ ,  $A = \bigcup_{x \in S} A(x)$  is a finite action space, and  $K_t \in A$  represents the action at  $t$ ;  $r : S \times A \rightarrow \mathbb{R}$  is a bounded reward function;  $p(y \mid x, a) = \mathbb{P}(X_{t+1} = y \mid X_t = x, K_t = a)$  denotes the homogeneous transition probability;  $\mu : S \rightarrow [0, 1]$  is the initial state probability mass function, and  $\boldsymbol{\mu}$  is an  $|S|$ -dimensional row vector with the  $n$ -th entry  $\mu(s_n)$  for any  $n \in \{1, \dots, |S|\}$ ; and  $\alpha \in (0, 1)$  is the discount factor.

A policy describes how to choose actions sequentially. It is stationary when it is independent of time, and deterministic if it determines an action for each state. In this study, we focus on stationary deterministic policy space  $D$  only. For a given MDP, a policy  $d : S \rightarrow A$  induces a Markov reward process. We denote its transition probability by an  $|S|$ -by- $|S|$  matrix  $\mathbf{P}_d$  with  $P_d(x, y) = p(y \mid x, d(x))$ , and its reward function by an  $|S|$ -dimensional column vector  $\mathbf{r}_d$  with  $r_d(x) = r(x, d(x))$ , for  $x, y \in S$ . We further denote its stationary distribution by an  $|S|$ -dimensional row vector  $\boldsymbol{\pi}_d$ , with the  $n$ -th entry  $\pi_d(s_n)$  for any  $n \in \{1, \dots, |S|\}$ . For notational simplicity, we omit the subscript “ $d$ ” when it is clear in the context.

In this study, we concern a risk-averse discounted mean-variance objective, where the discounted variance refers to the (normalized) cumulative discounted reward deviations from the discounted mean. Firstly, we denote the discounted mean value function under a policy  $d \in D$  by

$$v(x) = v_d(x) := (1 - \alpha) \mathbb{E}_x^d \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} r(X_t) \right\}, \quad x \in S, \quad (1)$$

where  $\mathbb{E}_x^d$  stands for the expectation given the initial state  $X_1 = x$  under the policy  $d$ , and we derive the following matrix form

$$\mathbf{v} = (1 - \alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{r}.$$

Considering the initial state distribution  $\mu$ , we have the discounted mean as

$$\eta = \eta_d := (1 - \alpha) \mathbb{E}_{\mu}^d \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} r(X_t) \right\} = \boldsymbol{\mu} \mathbf{v}, \quad (2)$$

where  $\mathbb{E}_\mu^d$  stands for the expectation given  $X_1 \sim \mu$  under the policy  $d$ .

Next, we define the discounted steady-state variance with a second moment value function. The (normalized) discounted second moment value function under a policy  $d$  is

$$w(x) = w_d(x) := (1 - \alpha) \mathbb{E}_x^d \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} r^2(X_t) \right\}, \quad x \in S, \quad (3)$$

and in matrix form, we have

$$\mathbf{w} = (1 - \alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}(\mathbf{r})_{\odot}^2,$$

where “ $\odot$ ” refers to the Hadamard product, i.e.,  $(\mathbf{r})_{\odot}^2 = (r^2(s_1), \dots, r^2(s_{|S|}))^T$ . Considering the initial state distribution  $\mu$ , we have the (normalized) discounted steady-state variance as

$$\zeta = \zeta_d := (1 - \alpha) \mathbb{E}_\mu^d \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} [r(X_t) - \eta]^2 \right\} = \boldsymbol{\mu}(\mathbf{w} - 2\eta\mathbf{v} + \eta^2\mathbf{e}),$$

where  $\mathbf{e} = (1, \dots, 1)^T$ .

After defining the two value functions, we define the discounted mean-variance value function by

$$u(x) = u_d(x) := v(x) - \beta[w(x) - 2\eta v(x) + \eta^2], \quad x \in S, \quad (4)$$

where  $\beta > 0$  is a risk-aversion parameter. We may consider the discounted mean-variance optimization problem as a discounted MDP with a special *policy-dependent* reward function represented by

$$f(x) = f_d(x) := r(x) - \beta[r(x) - \eta]^2, \quad x \in S, \quad (5)$$

where  $\eta$  depends on the policy. In matrix form, we have

$$\mathbf{f} = \mathbf{r} - \beta(\mathbf{r} - \eta\mathbf{e})_{\odot}^2,$$

and then we have the discounted mean-variance value function in matrix form as

$$\mathbf{u} = \mathbf{v} - \beta(\mathbf{w} - 2\eta\mathbf{v} + \eta^2\mathbf{e}) = (1 - \alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{f}.$$



Considering the initial state distribution  $\mu$ , we have the discounted mean-variance as

$$\begin{aligned}\xi = \xi_d &:= (1 - \alpha) \mathbb{E}_\mu^d \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} \{r(X_t) - \beta[r(X_t) - \eta]^2\} \right\} \\ &= (1 - \alpha) \mathbb{E}_\mu^d \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} f(X_t) \right\} = \eta - \beta\zeta = \boldsymbol{\mu}\mathbf{u}.\end{aligned}\tag{6}$$

Our objective is to find a deterministic policy  $d \in D$  to maximize the discounted mean-variance, i.e.,

$$\begin{aligned}\xi^* &:= \max_{d \in D} \{\xi\}, \\ d^* &\in \arg \max_{d \in D} \{\xi\}.\end{aligned}\tag{7}$$

The two equations together define the risk-averse discounted mean-variance optimization problem. We may consider this problem as a discounted MDP with a reward function defined in (5), where  $\eta$  is the discounted mean defined in (2). The value of the variance part in this special reward function for each state depends on policy instead of action, i.e., the performance at a current state will be affected by actions chosen at not only the current epoch but also future epochs. This dependency deprives the discounted variance metric of the time-consistency property. In this case, the Bellman optimality equation does not optimize over the admissible action set for a state, but over the policy space for the whole state space, which revokes the divide-and-conquer DP methods. In the next section, we will turn to the sensitivity-based optimization theory to solve this problem.

*Remark 1* (Problem formulation unification by discounting). Besides the rationales given in the introduction section, the involvement of a discount factor unifies variance-related problems in both discounted and average MDPs. This unification is embodied in the continuities of the discounted mean and variance at  $\alpha = 1$  with

$$\lim_{\alpha \uparrow 1} (1 - \alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1} = \mathbf{e}\boldsymbol{\pi}.$$

For details, see Chapter 2 in (Cao 2007). In other words, with respect to an expected return/variance/mean-variance objective, the three scenarios are

equivalent:

1. an MDP with a discount factor  $\alpha \uparrow 1$ ;
2. a discounted MDP with a special initial state distribution  $\boldsymbol{\mu} = \boldsymbol{\pi}$  (see Lemma 2.1); and
3. an average MDP.

In particular, the discount factor is trivial when the steady-state distribution equals the initial state distribution, for which we have the following lemma.

**Lemma 2.1** (The futility of discounting). *For a given policy  $d \in D$ , the discounted mean-variance is independent of the discount factor if  $\boldsymbol{\mu} = \boldsymbol{\pi}$ , i.e., the initial state distribution equals the stationary distribution.*

*Proof.* Since  $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ , for the discounted mean with the initial state distribution  $\boldsymbol{\mu} = \boldsymbol{\pi}$ , we have

$$\begin{aligned}
\eta &= \boldsymbol{\pi}(1 - \alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{r} \\
&= \boldsymbol{\pi}(1 - \alpha) \left( \sum_{t=1}^{\infty} \alpha^{t-1} \mathbf{P}^{t-1} \right) \mathbf{r} \\
&= (1 - \alpha) \left( \sum_{t=1}^{\infty} \alpha^{t-1} \boldsymbol{\pi} \mathbf{P}^{t-1} \right) \mathbf{r} \\
&= (1 - \alpha) \sum_{t=1}^{\infty} \alpha^{t-1} \boldsymbol{\pi} \mathbf{r} \\
&= \boldsymbol{\pi} \mathbf{r}.
\end{aligned}$$

Hence, the discounted mean is independent of the discount factor when  $\boldsymbol{\mu} = \boldsymbol{\pi}$  (Sutton and Barto 2018), and we can similarly derive that  $\zeta = \boldsymbol{\pi}(\mathbf{r} - \eta\mathbf{e})_{\odot}^2$  in this case. Therefore, we have  $\xi = \boldsymbol{\pi}\mathbf{f}$ , i.e., the discounted mean-variance is independent of the discount factor when  $\boldsymbol{\mu} = \boldsymbol{\pi}$ .  $\square$

### 3. Unified algorithm framework and discounted mean-variance value iteration

In this section, we propose a unified algorithm framework for the risk-averse mean-variance optimization and give a value iteration algorithm as an exam-

ple. First, we introduce the pseudo mean to remove the policy dependency of the reward function, and derive the discounted mean-variance PDF, which has a square term to handle the error from the introduction of pseudo mean. Next, we propose a unified algorithm framework with a bilevel optimization structure, where the inner problem refers to a standard MDP with a reward function dependent on a fixed pseudo mean, and the outer problem concerns a one-dimensional optimization of the pseudo mean. We show that risk-averse (mean-)variance optimization can be solved by algorithm variants in the proposed framework. Finally, we develop a value iteration in the framework for the discounted mean-variance problem. Furthermore, we prove its local convergence with a Bellman local-optimality equation, which is a necessary and sufficient condition for local optimality of a policy.

### 3.1. Performance difference formula

One key result of the sensitivity-based optimization theory is the performance difference formula (PDF). Based on the performance sensitivity analysis, a PDF quantifies the difference between system performances under any two policies. This theory is valid even for unorthodox Markov systems where the traditional DP methods fail (Cao 2007). For the concerned mean-variance optimization, Equation (5) shows that the reshaped reward function depends on the discounted mean  $\eta$ , which is unknown and affected by future actions. To handle this policy dependency, we firstly replace  $\eta$  with a pseudo mean  $\lambda \in \mathbb{R}$ , and define a pseudo reward function for any  $d \in D$  by

$$f_\lambda(x) = f_{\lambda,d}(x) := r(x) - \beta[r(x) - \lambda]^2, \quad x \in S, \quad (8)$$

and in matrix form, we have

$$\mathbf{f}_\lambda = \mathbf{r} - \beta(\mathbf{r} - \lambda \mathbf{e})_{\odot}^2 = (r(s_1) - \beta[r(s_1) - \lambda]^2, \dots, r(s_{|S|}) - \beta[r(s_{|S|}) - \lambda]^2)^T.$$

The corresponding pseudo discounted mean-variance value function under policy  $d \in D$  is

$$u_\lambda(x) := (1 - \alpha) \mathbb{E}_x^d \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} f_\lambda(X_t) \right\}, \quad x \in S,$$

and in matrix form, we have

$$\mathbf{u}_\lambda = (1 - \alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{f}_\lambda. \quad (9)$$

Considering the initial state distribution  $\mu$ , we have the pseudo discounted mean-variance as

$$\xi_\lambda = \xi_{\lambda,d} := (1 - \alpha)\mathbb{E}_\mu^d \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} f_\lambda(X_t) \right\} = \mu \mathbf{u}_\lambda. \quad (10)$$

Now we have a standard MDP with the pseudo reward function (8), and the difference between the pseudo discounted mean-variance  $\xi_\lambda$  and the discounted mean-variance  $\xi$  can be measured.

**Lemma 3.1** (Deviation of pseudo discounted mean-variance). *The pseudo discounted mean-variance and the discounted mean-variance have the following relation*

$$\xi_\lambda = \xi - \beta(\eta - \lambda)^2.$$

*Proof.* From (10), we have

$$\begin{aligned} \xi_\lambda &= (1 - \alpha)\mu(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{f}_\lambda \\ &= (1 - \alpha)\mu(\mathbf{I} - \alpha\mathbf{P})^{-1}[\mathbf{r} - \beta(\mathbf{r} - \lambda\mathbf{e})_\odot^2] \\ &= (1 - \alpha)\mu(\mathbf{I} - \alpha\mathbf{P})^{-1}[\mathbf{r} - \beta(\mathbf{r} - \eta\mathbf{e} + \eta\mathbf{e} - \lambda\mathbf{e})_\odot^2] \\ &= (1 - \alpha)\mu(\mathbf{I} - \alpha\mathbf{P})^{-1} \{ [\mathbf{r} - \beta(\mathbf{r} - \eta\mathbf{e})_\odot^2] - \beta[2\eta\mathbf{r} - 2\lambda\mathbf{r} - \eta^2\mathbf{e} + \lambda^2\mathbf{e}] \} \\ &= \xi - \beta(1 - \alpha)\mu(\mathbf{I} - \alpha\mathbf{P})^{-1}[2\eta\mathbf{r} - 2\lambda\mathbf{r} - \eta^2\mathbf{e} + \lambda^2\mathbf{e}]. \end{aligned}$$

With (2) and noticing that  $(1 - \alpha)\mu(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{e} = 1$ , we have

$$\begin{aligned} \xi_\lambda &= \xi - \beta(2\eta^2 - 2\lambda\eta - \eta^2 + \lambda^2) \\ &= \xi - \beta(\eta - \lambda)^2. \end{aligned}$$

□

*Remark 2* (Means in discounted variance). One may be tempted to set the long-run average  $\eta_a = \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}\{\sum_{t=1}^T r(X_t)\}$  as the value from which the

deviations are measured. Though it has a straightforward physical meaning, it is not the real mean in this discounted setting. When a discount factor is involved, it implies that the underpinned occupation measure of state-action pairs is in a discounted form, so the real “baseline” is the first central moment—the discounted mean  $\eta$ . This claim is supported by Lemma 3.1 as well, since the real mean should minimize variance (maximize mean-variance).

To construct the discounted mean-variance PDF, we start by quantifying the difference between any two pseudo discounted mean-variance functions under two policies with respect to a pseudo mean  $\lambda \in \mathbb{R}$ . From (9), for  $d \in D$  we have

$$\mathbf{u}_\lambda = (1 - \alpha)\mathbf{f}_\lambda + \alpha\mathbf{P}\mathbf{u}_\lambda.$$

Denote the other pseudo discounted mean-variance function by  $\mathbf{u}'_\lambda$  under  $d' \in D$ , with the transition matrix  $\mathbf{P}'$  and the pseudo reward  $\mathbf{f}'_\lambda$ , and then we have the difference as

$$\begin{aligned} \mathbf{u}'_\lambda - \mathbf{u}_\lambda &= (1 - \alpha)(\mathbf{f}'_\lambda - \mathbf{f}_\lambda) + \alpha(\mathbf{P}'\mathbf{u}'_\lambda - \mathbf{P}\mathbf{u}_\lambda) \\ &= (1 - \alpha)(\mathbf{f}'_\lambda - \mathbf{f}_\lambda) + \alpha(\mathbf{P}'\mathbf{u}'_\lambda - \mathbf{P}\mathbf{u}_\lambda + \mathbf{P}'\mathbf{u}_\lambda - \mathbf{P}'\mathbf{u}_\lambda) \\ &= (1 - \alpha)(\mathbf{f}'_\lambda - \mathbf{f}_\lambda) + \alpha(\mathbf{P}' - \mathbf{P})\mathbf{u}_\lambda + \alpha\mathbf{P}'(\mathbf{u}'_\lambda - \mathbf{u}_\lambda) \end{aligned} \quad (11)$$

$$= (1 - \alpha)(\mathbf{I} - \alpha\mathbf{P}')^{-1}(\mathbf{f}'_\lambda - \mathbf{f}_\lambda) + \alpha(\mathbf{I} - \alpha\mathbf{P}')^{-1}(\mathbf{P}' - \mathbf{P})\mathbf{u}_\lambda \quad (12)$$

$$= (\mathbf{I} - \alpha\mathbf{P}')^{-1}[(1 - \alpha)(\mathbf{f}'_\lambda - \mathbf{f}_\lambda) + \alpha(\mathbf{P}' - \mathbf{P})\mathbf{u}_\lambda], \quad (13)$$

noticing from (11) to (12), we have  $(\mathbf{I} - \alpha\mathbf{P}')(\mathbf{u}'_\lambda - \mathbf{u}_\lambda) = (1 - \alpha)(\mathbf{f}'_\lambda - \mathbf{f}_\lambda) + \alpha(\mathbf{P}' - \mathbf{P})\mathbf{u}_\lambda$ . Equation (13) checks the update rule of the standard value iteration from the perspective of PDF, and it explains why the standard value iteration converges to a global optimum. For the discounted mean-variance optimization, multiply the initial state distribution  $\mu$  on both sides, and then we have the PDF for the pseudo discounted mean-variance as

$$\xi'_\lambda - \xi_\lambda = \mu(\mathbf{I} - \alpha\mathbf{P}')^{-1}[(1 - \alpha)(\mathbf{f}'_\lambda - \mathbf{f}_\lambda) + \alpha(\mathbf{P}' - \mathbf{P})\mathbf{u}_\lambda].$$

Furthermore, with Lemma 3.1 we have the PDF for the discounted mean-

variance as

$$\xi' - \xi = \boldsymbol{\mu}(\mathbf{I} - \alpha \mathbf{P}')^{-1} [(1 - \alpha)(\mathbf{f}'_{\lambda} - \mathbf{f}_{\lambda}) + \alpha(\mathbf{P}' - \mathbf{P})\mathbf{u}_{\lambda}] + \beta(\eta' - \lambda)^2 - \beta(\eta - \lambda)^2. \quad (14)$$

Equation (14) quantifies the difference between the mean-variance performances under two policies in an MDP with a pseudo reward. Based on (14), it is straightforward to develop a policy iteration for the risk-averse (mean-)variance optimization problems. For the problems in average MDPs see (Xia 2016, 2020).

The involvement of the pseudo mean brings in the last two terms in (14) and makes the variance-related optimization converge to a local optimum. To clarify the local optimality, we present the discounted mean-variance performance derivative formula, which is another fundamental concept in the theory of sensitivity-based optimization. Different from the PDF, the derivative formula captures the behavior when the policy changes in a small local region. To see that, we first define a mixed policy space with the concept of mixed policy. For any two policies  $d, d' \in D$ , we define a mixed policy  $d^{\delta, d'}$  for  $\delta \in (0, 1)$ , which follows  $d$  with probability  $1 - \delta$  and follows  $d'$  in the rest. It is easy to verify that  $\mathbf{P}^{\delta} = \mathbf{P} + \delta(\mathbf{P}' - \mathbf{P})$  and  $\mathbf{f}_{\lambda}^{\delta} = \mathbf{f}_{\lambda} + \delta(\mathbf{f}'_{\lambda} - \mathbf{f}_{\lambda})$ . Substituting them into (14), we derive the performance difference between  $d^{\delta, d'}$  and  $d$  as

$$\xi^{\delta} - \xi = \boldsymbol{\mu}(\mathbf{I} - \alpha \mathbf{P}^{\delta})^{-1} \delta [(1 - \alpha)(\mathbf{f}'_{\lambda} - \mathbf{f}_{\lambda}) + \alpha(\mathbf{P}' - \mathbf{P})\mathbf{u}_{\lambda}] + \beta(\eta^{\delta} - \lambda)^2 - \beta(\eta - \lambda)^2.$$

Letting  $\delta \rightarrow 0$ , we obtain the derivative formula in the mixed policy space,

$$\frac{d\xi}{d\delta} = \boldsymbol{\mu}(\mathbf{I} - \alpha \mathbf{P})^{-1} [(1 - \alpha)(\mathbf{f}'_{\lambda} - \mathbf{f}_{\lambda}) + \alpha(\mathbf{P}' - \mathbf{P})\mathbf{u}_{\lambda}] + 2\beta(\eta - \lambda) \frac{d\eta}{d\delta}, \quad (15)$$

where  $\lim_{\delta \rightarrow 0} \mathbf{P}^{\delta} = \mathbf{P}$  and  $\lim_{\delta \rightarrow 0} \frac{d(\eta^{\delta} - \lambda)^2}{d\delta} = 2(\eta - \lambda) \frac{d\eta}{d\delta}$ . Next, we present a unified algorithm framework for the risk-averse discounted mean-variance optimization and develop a value iteration algorithm with a provable local convergence.

### 3.2. Unified algorithm framework

In this subsection, we propose a unified algorithm framework for the risk-averse discounted mean-variance optimization. This framework has a bilevel

optimization structure, where the inner problem refers to a standard MDP with a reward function dependent on a fixed pseudo mean, and the outer problem concerns a one-dimensional optimization of the pseudo mean. In particular, for variance-related problems (risk-averse discounted/average variance/mean-variance, etc.), the outer problem has a closed-form solution. Different algorithm variants can be developed with different solvers to the inner problem. Moreover, the proposed framework is applicable to some other variance-related optimality criteria as well.

The difficulty of the risk-averse (mean-)variance optimization lies in that the variance metric is a function of the discounted mean. This dependency suppresses the time-consistency property, and so that the traditional DP methods are not applicable. To remove the dependency, we introduce the pseudo mean to transform the special MDP to a standard one, where traditional DP methods can be applied. Mathematically, the introduction of the pseudo mean results in a bilevel optimization problem, which can be further extended to such problem equivalences.

**Lemma 3.2** (Problem equivalences with pseudo mean).

$$\xi^* = \max_{d \in D} \{\xi\} = \max_{d \in D} \left\{ \max_{\lambda \in \mathbb{R}} \{\xi_\lambda\} \right\} \quad (16)$$

$$= \max_{\lambda \in \mathbb{R}} \left\{ \max_{d \in D} \{\xi_\lambda\} \right\} \quad (17)$$

$$= \max_{\lambda \in \mathbb{R}} \left\{ \max_{d \in D} \{\boldsymbol{\mu} \mathbf{u}_\lambda\} \right\} \quad (18)$$

$$= \max_{\lambda \in \mathbb{R}} \left\{ \left\langle \left( \max_{d \in D} \{u_\lambda(x)\} \right)_{x \in S}^T, \boldsymbol{\mu}^T \right\rangle \right\}. \quad (19)$$

*Proof.* Lemma 3.1 implies that (16) holds with  $\lambda = \arg \max_{\lambda \in \mathbb{R}} \{\xi_\lambda\} = \eta_d$ . Since the outer and inner operators are both maximum, the two are exchangeable and (17) holds. Equation (18) comes from (10). Noticing that for a given  $\lambda$ , the inner optimization refers to a standard MDP, and the optimal mean results from the optimal value function  $u_\lambda$ .  $\square$

Equation (17) underpins the bilevel algorithm framework. By introducing the pseudo mean  $\lambda$ , the original problem is transformed to a bilevel problem,

where the inner problem concerns a standard MDP  $\mathcal{M}_\lambda = \langle S, A, f_\lambda, p, \mu, \alpha \rangle$ , and the outer problem refers to a one-dimensional optimization of the variable  $\lambda$ . The framework is shown in Algorithm 1.

---

**Algorithm 1** A unified algorithm framework for discounted mean-variance optimization

---

**Input:** The MDP  $\mathcal{M}$ ; initialize two pseudo means  $\lambda, \lambda' \in \mathbb{R}$  with  $\lambda' \neq \lambda$

**Output:** A local optimal policy  $d$  and the discounted mean-variance optimum

$\xi_d$

**while**  $\lambda \neq \lambda'$  **do**

$\lambda \leftarrow \lambda'$

Construct a standard MDP  $\mathcal{M}_\lambda$ . By using standard/optimistic algorithms (e.g., policy iteration, value iteration, or policy gradient), solve or partly solve  $\mathcal{M}_\lambda$  to obtain an improved policy  $d$  and  $\lambda'$  ▷ Inner optimization

**end while**

Return  $d$  and  $\xi_d = \mu \mathbf{u}_\lambda$

---

In the bilevel framework, the inner optimization helps optimize  $\lambda$ , which means to keep updating it with a value closer to the discounted mean of any local optimum. For any fixed  $\lambda$ , the resultant  $\mathcal{M}_\lambda$  is a standard MDP, so there are two threads to optimize  $\lambda$ . One is to calculate the optimal discounted mean with a standard DP algorithm. The convergence of algorithms stemming from this thread is guaranteed by the convergence of the involved DP algorithm and Lemma 3.1, which claims that the pseudo discounted mean-variance  $\xi_\lambda$  equals the discounted mean-variance  $\xi$  when  $\lambda = \eta$ . This thread is straightforward but could be conservative. The other thread is to improve  $\lambda$  with an intermediate value during its process converging to  $\eta_\lambda$ . The variant of policy iteration implementing in this thread is well known as the optimistic policy iteration (Sutton and Barto 2018), which has been studied for solving the risk-averse variance and mean-variance optimizations in average MDPs in (Xia 2016, 2020), respectively. Here we give simplified descriptions on variants of policy iteration and



value iteration for the inner optimization as examples in Algorithms 2 and 3<sup>1 2</sup>. The inputs of these four algorithm variants are the MDP  $\mathcal{M}$  and the current pseudo mean  $\lambda$ , and the outputs are the updated policy  $d$  and pseudo mean  $\lambda$ . For some realizations of the algorithm variants, we need to add the additional initializations to the input of the framework. It is worth noting that, though a set of DP algorithms, such as policy gradient, linear programming, and policy iteration, can be applied in the first thread, the convergences of their optimistic counterparts need further deliberations.

---

**Algorithm 2** Policy iteration variants for inner optimization in Algorithm 1

---

*Standard version:*

Initialize  $d, d' \in D$  with  $d \neq d'$

**while**  $d \neq d'$  **do**

$d \leftarrow d'$

$\mathbf{u}_\lambda \leftarrow (1 - \alpha)(\mathbf{I} - \alpha \mathbf{P}_d)^{-1} \mathbf{f}_{\lambda, d}$

$d' \in \arg \max_{d \in D} \{(1 - \alpha) \mathbf{f}_{\lambda, d} + \alpha \mathbf{P}_d \mathbf{u}_\lambda\}$

**end while**

$\lambda' \leftarrow (1 - \alpha) \boldsymbol{\mu} (\mathbf{I} - \alpha \mathbf{P}_d)^{-1} \mathbf{r}_d$

---

*Optimistic version:*

**Add. Init.:**  $d \in D$

$\mathbf{u}_\lambda \leftarrow (1 - \alpha)(\mathbf{I} - \alpha \mathbf{P}_d)^{-1} \mathbf{f}_{\lambda, d}$

$d \in \arg \max_{d \in D} \{(1 - \alpha) \mathbf{f}_{\lambda, d} + \alpha \mathbf{P}_d \mathbf{u}_\lambda\}$

$\lambda' \leftarrow (1 - \alpha) \boldsymbol{\mu} (\mathbf{I} - \alpha \mathbf{P}_d)^{-1} \mathbf{r}_d$

A variety of algorithms for variance-related optimization in previous works can be unified and analyzed in the proposed framework. When the equivalence in (18) is concerned, we have the policy gradient for the mean-variance optimizations in discounted MDPs (Bisi et al. 2020), and the policy iteration for the mean-variance optimization in discounted and average MDPs (Zhang et al. 2021). However, no convergence analysis is given in either of the works, such as the analyses for the policy iterations in (Xia 2016, 2020). Since most, if not all, of the algorithms exploit the variance property described in Lemma 3.1 and result in local optima, a convergence analysis is crucial. In the proposed

---

<sup>1</sup>The norm used through the paper could be  $p$ -norm for  $p \in \{1, 2, +\infty\}$ .

<sup>2</sup>For the value iteration variants, the policy  $d$  should be derived at the end of Algorithm 1, and here we put it in the descriptions for structure unification only.

---

**Algorithm 3** Value iteration variants for inner optimization in Algorithm 1

---

<i>Standard version:</i>	<i>Optimistic version:</i>
<b>Add. Init.:</b> a small constant $\theta > 0$	<b>Add. Init.:</b> value functions $\mathbf{v}, \mathbf{u}_\lambda \in \mathbb{R}^{ S }$
Initialize value functions $\mathbf{v}, \mathbf{u}_\lambda, \mathbf{u}'_\lambda \in \mathbb{R}^{ S }$	
$\mathbb{R}^{ S }$ with $\ \mathbf{u}_\lambda - \mathbf{u}'_\lambda\  > \theta$	$d \in \arg \max_{d \in D} \{(1 - \alpha)\mathbf{f}_{\lambda,d} + \alpha \mathbf{P}_d \mathbf{u}_\lambda\}$
<b>while</b> $\ \mathbf{u}_\lambda - \mathbf{u}'_\lambda\  > \theta$ <b>do</b>	$\mathbf{u}_\lambda \leftarrow (1 - \alpha)\mathbf{f}_{\lambda,d} + \alpha \mathbf{P}_d \mathbf{u}_\lambda$
$\mathbf{u}_\lambda \leftarrow \mathbf{u}'_\lambda$	$\mathbf{v} \leftarrow (1 - \alpha)\mathbf{r}_d + \alpha \mathbf{P}_d \mathbf{v}$
$d \in \arg \max_{d \in D} \{(1 - \alpha)\mathbf{f}_{\lambda,d} + \alpha \mathbf{P}_d \mathbf{u}_\lambda\}$	$\lambda' \leftarrow \mu \mathbf{v}$
$\mathbf{u}'_\lambda \leftarrow (1 - \alpha)\mathbf{f}_{\lambda,d} + \alpha \mathbf{P}_d \mathbf{u}_\lambda$	
$\mathbf{v} \leftarrow (1 - \alpha)\mathbf{r}_d + \alpha \mathbf{P}_d \mathbf{v}$	
<b>end while</b>	
$\lambda' \leftarrow \mu \mathbf{v}$	

---

framework, a convergence analysis can be developed with the aid of a PDF.

*Remark 3* (Unified algorithm framework). The unification in the algorithm framework is twofold.

1. A set of problems potentially solvable by algorithms developed in the framework. These problems include, but not limited to, the risk-averse variance and mean-variance optimizations in discounted and average MDPs, and these four metrics can be covered by (6). When the discount factor  $\alpha \uparrow 1$ , the problem turns into the mean-variance maximization in average MDPs (Xia 2020, Gosavi 2014) (see Remark 1). When the risk-aversion parameter  $\beta$  is large enough with respect to the mean, the problem degrades to the variance minimization problem (for average MDPs, see (Xia 2016)).
2. For the unified set of problems, a set of algorithms can be developed and analyzed in the framework, such as the policy gradients (Prashanth and Ghavamzadeh 2013, Bisi et al. 2020), the policy iterations (Xia 2016, 2020, Zhang et al. 2021), and the value iteration (Gosavi 2014). The missing convergence analyses in some previous works can be developed as well.

Moreover, both standard and optimistic versions of the DP algorithms can be studied with deliberations on their convergences.

*Remark 4* (Convergence rate and complexity). In the bilevel optimization framework, the inner problem is a standard MDP for a given pseudo mean  $\lambda$ . The convergence rate relies on the solver to the inner problem. For example, the convergence of the value iteration is linear at rate  $\beta$ . However, since the mean value function  $v$  is different from the mean-variance value function  $u$ , the convergence rate of  $\lambda$  cannot be analyzed similarly. The complexity of an algorithm in the bilevel optimization framework depends as well. Taking value iteration for example, the complexity for each iteration is  $\mathcal{O}(|S|^2|A|)$ . A lower bound for the number of iterations needed can be estimated with an error bound at the  $n$ -th iteration. For any  $d \in \arg \max_{d \in D} \{\mathbf{f}_{\lambda, \mathbf{d}} + \beta \mathbf{P}_{\mathbf{d}} \mathbf{u}_{\lambda, n}\}$ , where  $\mathbf{u}_{\lambda, n}$  is the pseudo mean-variance value function at  $n$ -th iteration, we have

$$\|\mathbf{u}_{\lambda}^{\mathbf{d}} - \mathbf{u}_{\lambda}^*\| \leq \frac{2\beta^{n-1}}{1-\beta} \|\max_{d \in D} \{\mathbf{f}_{\lambda, \mathbf{d}} + \beta \mathbf{P}_{\mathbf{d}} \mathbf{u}_{\lambda, 1}\} - \mathbf{u}_{\lambda, 1}\|,$$

where  $\mathbf{u}_{\lambda}^{\mathbf{d}}$  is the pseudo mean-variance value function under  $d$ , and  $\mathbf{u}_{\lambda, 1}$  is the initial pseudo mean-variance value function (Puterman 2005). To seek  $\epsilon$ -optimal policies, we have

$$\begin{aligned} \frac{2\beta^{n-1}}{1-\beta} \|\max_{d \in D} \{\mathbf{f}_{\lambda, \mathbf{d}} + \beta \mathbf{P}_{\mathbf{d}} \mathbf{u}_{\lambda, 1}\} - \mathbf{u}_{\lambda, 1}\| &\leq \epsilon, \\ \Leftrightarrow n &\geq \log_{\beta} \left\{ \frac{\epsilon(1-\beta)}{2\|\max_{d \in D} \{\mathbf{f}_{\lambda, \mathbf{d}} + \beta \mathbf{P}_{\mathbf{d}} \mathbf{u}_{\lambda, 1}\} - \mathbf{u}_{\lambda, 1}\|} \right\} + 1. \end{aligned}$$

### 3.3. Discounted mean-variance value iteration

In this subsection, we develop a discounted mean-variance value iteration (DMVVI) in the proposed framework as an example. We show the relationship between the original problem and the one with a pseudo mean. We prove the local convergence of the DMVVI with a Bellman local-optimality equation, which is a necessary and sufficient condition for local optimality of a policy. We believe that the DMVVI algorithm provides a foundation for model-free RL methods, such as Q-learning and SARSA, to the variance-related optimizations.

Equation (19) is the foundation for a value iteration to the risk-averse variance-related optimization. By further extending (19), we derive an optimality equation as

$$\xi^* = \max_{\lambda \in \mathbb{R}} \left\{ \left\langle \left( \max_{a \in A(x)} \{ (1 - \alpha) f_\lambda(x, a) + \alpha \sum_{y \in S} p(y | x, a) u_\lambda^*(y) \} \right)_{x \in S}^T, \boldsymbol{\mu}^T \right\rangle \right\}. \quad (20)$$

Equation (20) forms a Bellman optimality equation parameterized by  $\lambda$  for the inner standard MDP with a fixed  $\lambda$ . To improve the discounted mean-variance value function and evaluate the mean simultaneously, we maintain two value functions in the iteration: the discounted mean value function  $v$  and the second moment value function  $w$ . Given a policy  $d \in D$ , we define the value function updates with two Bellman operators:

$$\mathbf{v}' = \mathcal{T}_{v,d} \mathbf{v} := (1 - \alpha) \mathbf{r} + \alpha \mathbf{P} \mathbf{v}$$

and

$$\mathbf{w}' = \mathcal{T}_{w,d} \mathbf{w} := (1 - \alpha) (\mathbf{r})_{\odot}^2 + \alpha \mathbf{P} \mathbf{w}.$$

Now we give the value iteration for the risk-averse discounted mean-variance optimization in Algorithm 4, which is a detailed description of the standard value iteration in Algorithm 3 in the framework.

Although the problem with a pseudo mean is different from the original problem, we have the following theorem to relate these two problems in an iterative algorithm.

**Theorem 3.3** (Relationship between the two problems). *For a fixed pseudo mean-variance value function  $u_\lambda$ , compute the pseudo mean-variance  $\xi_\lambda$ , the policy  $d \in D$  and  $\eta = \eta_d$ , and then set  $\lambda = \eta$ . In the next iteration, if we have  $\xi'_\lambda \geq \xi_\lambda$ , then we have  $\xi' \geq \xi$ . We have  $\xi' > \xi$  if the first inequality strictly holds.*

*Proof.* With Lemma 3.1, we have

$$\xi' - \xi = [\xi'_\lambda + \beta(\eta' - \lambda)^2] - [\xi_\lambda + \beta(\eta - \lambda)^2].$$

---

**Algorithm 4** The discounted mean-variance value iteration (DMVVI)

---

**Input:** The MDP  $\mathcal{M}$ ; a small threshold  $\theta > 0$ ; initialize  $\lambda, \lambda' \in \mathbb{R}$  with  $\lambda \neq \lambda'$ ,  
two discounted mean value functions, two second moment value functions  
and a pseudo mean-variance value function  $\mathbf{v}, \mathbf{v}', \mathbf{w}, \mathbf{w}', \mathbf{u}_\lambda \in \mathbb{R}^{|S|}$ , with  
 $\|\mathbf{u}_\lambda - [\mathbf{v}' - \beta(\mathbf{w}' - 2\lambda\mathbf{v}' + \lambda^2\mathbf{e})]\| > \theta$

**Output:** Local optimal policy  $d$  and the local optimum  $\xi$

```

1: while  $\lambda' \neq \lambda$  do
2:    $\lambda \leftarrow \lambda'$ 
3:   while  $\|\mathbf{u}_\lambda - [\mathbf{v}' - \beta(\mathbf{w}' - 2\lambda\mathbf{v}' + \lambda^2\mathbf{e})]\| > \theta$  or  $\|\mathbf{v}' - \mathbf{v}\| > \theta$  do
4:      $\mathbf{v} \leftarrow \mathbf{v}'$ 
5:      $\mathbf{w} \leftarrow \mathbf{w}'$ 
6:      $\mathbf{u}_\lambda \leftarrow \mathbf{v} - \beta(\mathbf{w} - 2\lambda\mathbf{v} + \lambda^2\mathbf{e})$ 
7:     for  $x \in S$  do
8:       
$$d(x) \in \arg \max_{a \in A(x)} \left\{ (1 - \alpha)f_\lambda(x, a) + \alpha \sum_{y \in S} p(y \mid x, a)u_\lambda(y) \right\}$$

9:     end for
10:     $\mathbf{v}' \leftarrow \mathcal{T}_{v,d}\mathbf{v}$ 
11:     $\mathbf{w}' \leftarrow \mathcal{T}_{w,d}\mathbf{w}$ 
12:   end while
13:    $\lambda' \leftarrow \mu\mathbf{v}'$ 
14: end while
15:  $\xi = \mu\mathbf{u}_\lambda$ 

```

---

By setting  $\lambda = \eta$ , we have

$$\xi' - \xi = \xi'_\lambda - \xi_\lambda + \beta(\eta - \eta')^2.$$

Therefore, if  $\xi'_\lambda \geq (>)\xi_\lambda$ , we have  $\xi' \geq (>)\xi$ . □

Though the error term in Lemma 3.1 can gracefully handle the policy dependency of the variance metric, it takes a toll as well. Next, we show the condition

that the global optimum cannot be reached in one iteration of the DMVVI.

**Lemma 3.4** (Unreachability of global optimum in one iteration). *In one outer iteration of the DMVVI, ignoring the estimation error, given the current discounted mean-variance  $\xi$  with the discounted mean  $\eta$ , and the global optimum  $\xi^*$  with*

$$\eta^* = \eta_{d^*}, \quad d^* \in \arg \min_{d \in D^*} \{|\eta_d - \eta|\}, \quad (21)$$

where  $D^*$  is the set of global optimal policies, if

$$(\xi^* - \xi) \leq \beta(\eta^* - \eta)^2, \quad (22)$$

the global optimum cannot be reached in the current iteration.

*Proof.* For the current discounted mean-variance function  $u$ , we have the transition probability matrix  $\mathbf{P}$  and the discounted mean-variance reward function  $f$ . Denote  $\xi_\eta^*$  the pseudo discounted mean-variance under a global optimal policy  $d^* \in D^*$ , with  $\mathbf{P}^*$  and  $f_\eta^*$ . From Lemma 3.1, we have

$$\begin{aligned} (\xi^* - \xi) - \beta(\eta^* - \eta)^2 &= \xi_\eta^* - \xi \\ &= \boldsymbol{\mu}(\mathbf{I} - \alpha\mathbf{P}^*)^{-1}[(1 - \alpha)(\mathbf{f}_\eta^* - \mathbf{f}) + \alpha(\mathbf{P}^* - \mathbf{P})\mathbf{u}] \leq 0. \end{aligned}$$

Since each entry of  $\boldsymbol{\mu}(\mathbf{I} - \alpha\mathbf{P}^*)^{-1}$  is nonnegative, there exists at least one  $x \in S$ , such that

$$\begin{aligned} [(1 - \alpha)\mathbf{f}_\eta^* + \alpha\mathbf{P}^*\mathbf{u}](x) &\leq [(1 - \alpha)\mathbf{f} + \alpha\mathbf{P}\mathbf{u}](x) \\ &\leq \max_{a \in A(x)} \left\{ (1 - \alpha)[r(x, a) - \beta(r(x, a) - \eta)^2] + \alpha \sum_{y \in S} p(y \mid x, a)u(y) \right\}, \end{aligned}$$

which means that, in this iteration, the value function will not converge to the optimal value function under any  $d^* \in D^*$ , so the global optimum cannot be reached in the current iteration.  $\square$

Lemma 3.4 claims that, given  $\xi$ ,  $\eta$ , and the optimal policy  $d^*$  whose discounted mean performance  $\eta^*$  is closest to  $\eta$  (see (21)) in the current iteration, if the difference between  $\eta$  and  $\eta^*$  is relatively large (see (22)), then the pseudo

mean-variance under  $d^*$  will deteriorate to  $\xi_\eta^* \leq \xi$  because of  $\eta$ , and the optimum cannot be reached in this iteration. With the aid of Lemma 3.4, we can prove the local convergence of the DMVVI. We first give the definition of local optimality for the discounted mean-variance optimization, along with the Bellman local-optimality equation.

**Definition 3.1** (Local optimality). For a policy  $d \in D$ , if there exists  $\Delta \in (0, 1)$ , we always have  $\xi_d \geq \xi_{d^\delta, d'}$  for any  $\delta \in (0, \Delta)$ , then we say  $d$  is a local optimum in the mixed policy space.

**Definition 3.2** (Bellman local-optimality equation). A policy  $d^\# \in D$  is a local optimal policy if and only if its value function  $u^\#$  satisfies the Bellman local-optimality equation

$$u^\#(x) = \max_{a \in A(x)} \left\{ (1 - \alpha)[r(x, a) - \beta(r(x, a) - \eta_{d^\#})^2] + \alpha \sum_{y \in S} p(y | x, a) u^\#(y) \right\}. \quad (23)$$

The local convergence of DMVVI is established with the performance derivative formula (15), which shows that the converged value function has a nonpositive gradient in any feasible directions. Next, we give the local convergence proof of the DMVVI.

**Theorem 3.5** (Local convergence of DMVVI). *The DMVVI converges to a local optimum.*

*Proof.* First, we prove the convergence of the DMVVI. At  $t$ -th step of the outer iteration, we have a pseudo mean  $\lambda_t \in \mathbb{R}$  and a discounted mean-variance  $\mu u_{\lambda_t, t}$  (i.e.,  $\xi_{\lambda_t, t}$ ) at the beginning of the inner optimization (at Line 3 in Algorithm 4). After one inner optimization, we have  $\mu u_{\lambda_t, t+1} \geq \mu u_{\lambda_t, t}$  (i.e.,  $\xi_{\lambda_t, t+1} \geq \xi_{\lambda_t, t}$ , at Line 1), which is guaranteed by the convergence of the standard value iteration (Puterman 2005). Derive  $\lambda_{t+1}$  (at Line 2) and then we have  $\mu u_{\lambda_{t+1}, t+1} \geq \mu u_{\lambda_t, t+1}$  (i.e.,  $\xi_{\lambda_{t+1}, t+1} \geq \xi_{\lambda_t, t+1}$ ) based on Lemma 3.1, which strictly holds if  $\lambda_{t+1} \neq \lambda_t$ . Since the policy space  $D$  is finite and

$\xi_{\lambda_{t+1}, t+1} \geq \xi_{\lambda_t, t+1} \geq \xi_{\lambda_t, t}$ , this algorithm will stop after a finite number of iterations. Thus, the convergence of the DMVVI is proved.

Second, we prove that the DMVVI converges to a local optimum. From Lemma 3.4, the DMVVI converges to the current discounted mean-variance  $\xi^\#$  if it is local optimum, i.e., it satisfies the Bellman local-optimality equation (23). Plugging  $\lambda = \eta_{d^\#}$  into (15), we derive that

$$\frac{d\xi_{d^\#}}{d\delta} = \boldsymbol{\mu}(\mathbf{I} - \alpha\mathbf{P}^\#)^{-1}[(1 - \alpha)(\mathbf{f}^{\#'} - \mathbf{f}^\#) + \alpha(\mathbf{P}' - \mathbf{P})\mathbf{u}^\#].$$

Noticing that the elements of  $\boldsymbol{\mu}(\mathbf{I} - \alpha\mathbf{P}^\#)^{-1}$  are always nonnegative, we conclude that  $\frac{d\xi_{d^\#}}{d\delta} \leq 0$  along any feasible changing direction, indicating that  $d^\#$  is a local optimum in the mixed policy space.  $\square$

Comparing with the standard value iteration, whose global convergence is guaranteed by (13), Theorem 3.5 explains why the DMVVI converges a local optimum—within one iteration, the value function update always depends on the former discounted mean, and the error term in Lemma 3.1 suppresses the global policy in that iteration (Lemma 3.4). Since the optimization procedure is deterministic, the sequence of  $\lambda$ 's depends on the initial  $\lambda$  only. If a pseudo mean derived from a local optimal policy is reached before one from a global optimal policy, then the DMVVI will converge to this local optimum, and in this case, the iteration is “trapped” by the Bellman local-optimality equation (23). This local convergence analysis can be generalized to other algorithm variants governed by the unified algorithm framework, such as the works in (Xia 2016, 2020, Zhang et al. 2021). Besides, if all policies share the same discounted mean, the error term vanishes, and the DMVVI will converge to the global optimum.

By introducing a pseudo mean, the PDF quantifies the performance difference between any two policies and provides a foundation for iterative algorithms. As shown in the introduction section, most of the relevant works focus on gradient-based methods, and to the best of our knowledge, only two works concern iterative algorithms for variance-related problems besides (Xia 2016, 2020). One is (Zhang et al. 2021), which reformulates the mean-variance



formulation in average MDPs with its Legendre-Fenchel dual, and derives a similar problem formulation as the one with a pseudo mean. The authors propose a stochastic block coordinate ascent algorithm (Cui et al. 2018), which can be unified as a policy iteration in our unified algorithm framework. No local convergence analysis is given in this work. The other is (Gosavi 2014), where a value iteration is proposed for the mean-variance optimization in average MDPs. However, relevant algorithm analyses, such as convergence and local optimality, are circumvented with assumptions. We believe that our work complements risk-sensitive optimization in MDPs from three aspects:

1. The mean-variance optimization theory is extended to discounted MDPs;
2. A unified algorithm framework is proposed, where a variety of algorithm variants can be unified and analyzed with the aid of PDF, and the framework works for a collection of risk-sensitive criteria including, but not limited to, several variance-related risk measures; and
3. The DMVVI is proposed with a convergence analysis and a Bellman local optimality equation, and it provides a foundation for model-free RL methods, such as Q-learning and SARSA, to the variance-related optimizations.

#### 4. Numerical experiment

In this section, we validate the proposed DMVVI by solving the discounted mean-variance optimization in a portfolio management problem (Tamar et al. 2012). We assume the dynamics of portfolio management to be a stationary stochastic process and model it as an MDP with an appropriate discretization of all relevant continuous variables.

A portfolio is usually composed of two types of assets. One is the liquid assets (e.g., short-term T-bills), each of which has a fixed interest rate  $r_l$  and can be sold at any epoch  $t \in \mathbb{N}^+$ . The other type is the non-liquid assets (e.g., low liquidity bonds or options), each of which can be sold only after a maturity period of  $M \in \mathbb{N}^+$  steps with a time-dependent interest rate  $r_n(t)$ . We assume that  $r_n(t)$  can take either  $r_n^{low}$  or  $r_n^{high}$ , and the transitions between these two

cases occur randomly with a switching probability  $p_s$ . In addition, a non-liquid asset has a default risk with probability  $p_r$ . To simplify the problem, we assume that the portfolio has one for each type of assets. Besides, we discretize the investor's total available cash into  $N \in \mathbb{N}^+$  units, and represent a state by a vector  $x = (x_0, x_1, \dots, x_{M+1}) \in \{0, \dots, N\}^{M+1} \times \{r_n^{low}, r_n^{high}\}$ , where  $x_0 \in \{0, \dots, N\}$  is the number of units invested in the liquid asset;  $x_1, \dots, x_M \in \{0, \dots, N\}$  are the numbers of units invested in the non-liquid assets with  $0, \dots, (M-1)$  time steps to maturity, respectively; and  $x_{M+1} \in \{r_n^{low}, r_n^{high}\}$  records the current non-liquid interest rate. At each epoch with a state  $x$ , the investor may change her portfolio by investing a number of units  $a \in A(x) = \{0, 1, \dots, (x_0 + x_1)\}$  in the non-liquid asset. We further assume that default can happen only at the maturity epoch. The dynamics of the investment among the liquid asset and the non-liquid assets with different maturity times is illustrated in Figure 1.

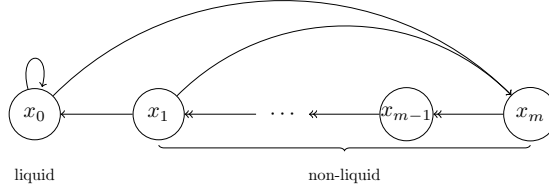


Figure 1: Dynamics of investment among liquid asset and non-liquid assets with different maturity times. The arrow “ $\rightarrow$ ” represents a controllable investment, and “ $\rightsquigarrow$ ” represents an uncontrollable state transition. Notice that the investment at  $x_1$  can be directly reinvested to the non-liquid asset since it is matured at the decision epoch.

To consider a small-scale problem, we set the discount factor  $\alpha = 0.95$ , the risk-aversion parameter  $\beta = 1$ , the maturity period  $M = 3$ , the total available cash units  $N = 3$ , the liquid asset interest rate  $r_l = 0.03$ , the low non-liquid asset interest rate  $r_n^{low} = 0.4$ , the high non-liquid asset interest rate  $r_n^{high} = 1$ , the interest switching probability  $p_s = 0.1$ , and the default risk probability  $p_r = 0.1$ . We assume that all units of cash are in the liquid asset at  $t = 1$ . For the given parameter setting, we construct an MDP to represent this portfolio management problem.

#### 4.1. Local convergence and laddered policy

We solve the risk-averse discounted mean-variance optimization in the portfolio management with the DMVVI (Algorithm 4) with  $\theta = 10^{-5}$ . As Theorem 3.5 states, the DMVVI converges to a local optimum. For different initial pseudo means, the value iteration may converge to different local optima. In this portfolio management problem with the specified setting, the algorithm converges to the global optimum if we initialize the pseudo mean by  $\lambda' = 1$ . In contrast, it converges to a local optimum if we initialize  $\lambda' = -1$ . The two convergences are compared in Figures 2 and 3.

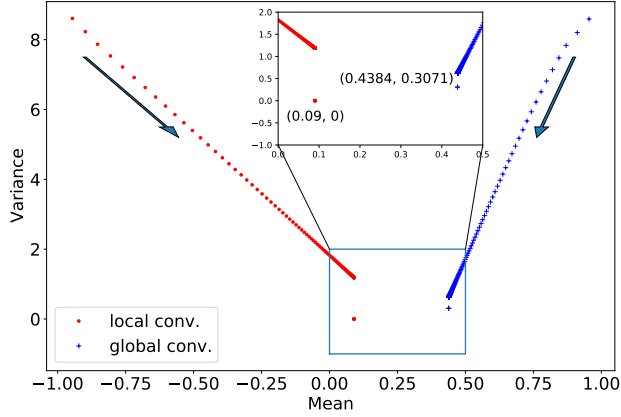


Figure 2: The local and global convergences in the mean-variance space with  $\lambda' = -1$  and 1, respectively.

In Figure 2 we show that when the pseudo mean is initialized differently, the algorithm will converge to different local optima. In this case, the local optimal policy is  $d(x) = 0, x \in S$ , which indicates that we should always invest in the liquid risk-free asset, which will deliver a deterministic revenue 0.09. Comparing with this conservative local optimal policy, the global optimum achieves a revenue with the discounted mean  $\mu \approx 0.4384$  and the discounted variance  $\zeta \approx 0.3071$ , which is better with  $\beta = 1$ . The discounted mean-variance values are 0.1313 and 0.09 for the global and local optima, and the convergences are

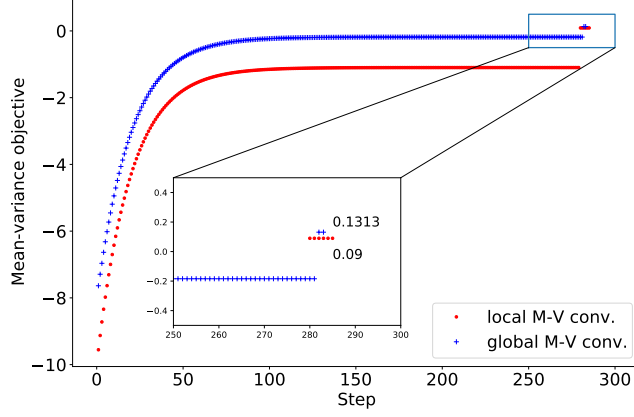


Figure 3: The local and global convergences along the time steps with  $\lambda' = -1$  and 1, respectively.

illustrated in Figure 3. In both Figures 2 and 3, we can see that there are “jumps” near the ends. That is because the pseudo means are updated in the outer optimization in the bilevel framework.

It is worth noting that, the global optimal policy is  $d(x) = 0$  for  $x \in \{x \in S \mid x_0 + x_1 = 0\}$  and  $d(x) = 1$  otherwise. This policy is a laddered (or laddering) strategy, which splits an investment to non-liquid assets into equal units and invests them in regular intervals consecutively in order to maintain a cash flow. Since investments are spread across several maturities, the laddered strategy also implies temporal diversification which reduces financial risks. Laddered strategies are widely used in portfolio management (Caldeira et al. 2016).

#### 4.2. Mean-variance trade-off along convex efficient frontier

A mean-variance metric is a combination of two metrics weighted with a risk-aversion parameter  $\beta$ . Different  $\beta$ 's reflect different trade-offs between profit ( $\mu$ ) and risk ( $\zeta$ ). By adjusting  $\beta$ , different policies represented by  $(\mu, \zeta)$ 's can be found with the DMVVI (assuming the global optimality is achieved), and these combinations establish a *convex efficient frontier*, which is the intersection of

the mean-variance (Pareto) efficient frontier and the convex hull of the mean-variance pairs. The convex hull, which is also known as the convex envelope or convex closure, is the smallest convex set that “contains” a given space. Each policy in the convex efficient frontier corresponds to a global optimal solution to a mean-variance optimization with a specific  $\beta$ . The convex efficient frontier of the discounted mean-variance problem along with some possible mean-variance pairs is shown in Figure 4. In this case, the convex efficient frontier is composed of five vertices, and every vertex represents a possible  $(\mu, \zeta)$  under an optimal policy with respect to some risk-aversion parameter.

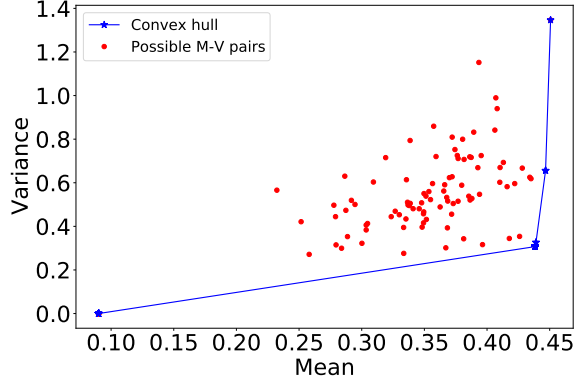


Figure 4: The convex hull of the risk-averse discounted mean-variance optimization.

#### 4.3. Risk-aversion versus risk-neutral

Another concern could be the difference between the results of risk-averse and risk-neutral optimizations—“how much profit is sacrificed” and “how much risk is eliminated” are crucial questions. To quantify the difference, we compare the mean-variance pairs under the optimal policies for the cases with  $\beta = 0$  and  $\beta = 1$ . The comparison is shown in Figure 5, with the Gaussian distribution used for presentation only. The mean and variance for the risk-neutral case is  $(0.4507, 1.3468)$ , while its risk-averse counterpart is  $(0.4384, 0.3071)$ . We can see that by sacrificing  $(0.4507 - 0.4384) = 0.0123$  ( $0.0123/0.4507 \times 100\% \approx$

2.73%) in the expected profit, we reduce the variance from 1.3468 to 0.3071 ( $0.3071/1.3468 \times 100\% \approx 22.80\%$ ). It means that in this case, we sacrifice 2.73% of the expected profit to eliminate  $(1 - 22.80\%) = 77.20\%$  of the risk, which could be a meaningful risk-averse decision.

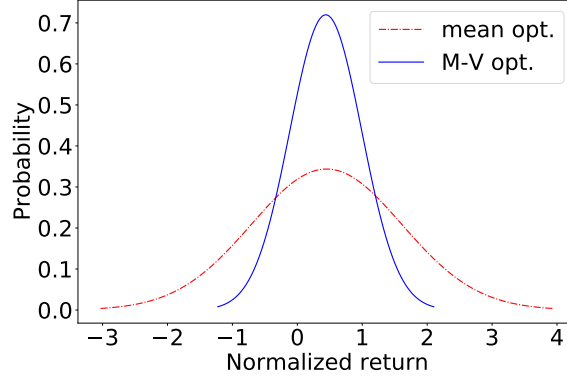


Figure 5: A comparison between the means and variances with  $\beta = 0$  and  $\beta = 1$ .

## 5. Summary and outlook

In this paper, we study the risk-averse discounted mean-variance optimization, in which the concerned variance refers to the steady-state variance formulated with a discount factor. The problem formulation unifies the (mean-)variance optimizations in discounted and average MDPs. Since the reward functions are policy-dependent in the variance-related problems, the MDPs are unorthodox, and the traditional DP methods cannot be applied directly. We proposed a unified algorithm framework to solve and analyze this problem. This framework has a bilevel optimization structure, where the inner problem refers to a standard MDP, and the outer problem concerns a one-dimensional optimization. For the concerned mean-variance optimization, the outer problem has a closed-form solution, i.e., the discounted mean with respect to a given policy determined by the inner optimization. This framework unifies a series

of algorithms for several variance-related optimizations in discounted and average MDPs, such as the policy gradients (Prashanth and Ghavamzadeh 2013, Bisi et al. 2020), the policy iterations (Xia 2016, 2020, Zhang et al. 2021), and the value iteration (Gosavi 2014). Furthermore, convergence analyses can be developed with the aid of the Bellman local-optimality equation. For the risk-averse mean-variance optimization in discounted MDPs, we take value iteration as an example and develop the DMVVI algorithm. A numerical experiment on a portfolio management problem is given to validate the proposed DMVVI.

Possibilities for future work include studies on the conditions of global convergence, i.e., when an algorithm for a variance-related optimization can converge to a global optimum with probability one. A first attempt could be a stochastic global convergence achieved with the exploratory mechanism in RL. The other future work could be model-free RL algorithms, such as Q-learning and SARSA, as online solutions to risk-averse variance-related problems. We believe that the proposed algorithm framework and one of its consequent algorithms, the DMVVI, provide a theoretic foundation and inspiration for future works.

## References

- Basak, S. and Chabakauri, G. (2010). Dynamic mean-variance asset allocation. *Review of Financial Studies*, 23(8):2970–3016.
- Best, M. J. and Grauer, R. R. (1991). Sensitivity analysis for mean-variance portfolio problems. *Management Science*, 37(8):980–989.
- Bisi, L., Sabbioni, L., Vittori, E., Papini, M., and Restelli, M. (2020). Risk-averse trust region optimization for reward-volatility reduction. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 4583–4589.
- Caldeira, J. F., Moura, G. V., and Santos, A. A. (2016). Bond portfolio optimization using dynamic factor models. *Journal of Empirical Finance*, 37:128–158.
- Cao, X.-R. (2007). *Stochastic Learning and Optimization - A Sensitivity-Based Approach*. Springer Science & Business Media.

- Chung, K. J. (1994). Mean-variance tradeoffs in an undiscounted MDP: The unichain case. *Operations Research*, 42(1):184–188.
- Cui, X., Sun, X., Zhu, S., Jiang, R., and Li, D. (2018). Portfolio optimization with nonparametric value at risk: A block coordinate descent method. *INFORMS Journal on Computing*, 30(3):454–471.
- Eckstein, J., Eskandani, D., and Fan, J. (2016). Multilevel optimization modeling for risk-averse stochastic programming. *INFORMS Journal on Computing*, 28(1):112–128.
- Filar, J. A., Kallenberg, L. C. M., and Lee, H. M. (1989). Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161.
- Gosavi, A. (2014). Variance-penalized Markov decision processes: Dynamic programming and reinforcement learning techniques. *International Journal of General Systems*, 43(6):649–669.
- Hernández-Lerma, O., Vega-Amaya, O., and Carrasco, G. (1999). Sample-path optimality and variance-minimization of average cost Markov control processes. *SIAM Journal on Control and Optimization*, 38(1):79–93.
- Ho, Y.-C. and Cao, X.-R. (1991). *Perturbation Analysis of Discrete Event Dynamic Systems*. Springer Science & Business Media.
- Kandel, S. and Stambaugh, R. F. (1989). A mean-variance framework for tests of asset pricing models. *Review of Financial Studies*, 2(2):125–156.
- Kouvelis, P., Pang, Z., and Ding, Q. (2018). Integrated commodity inventory management and financial hedging: A dynamic mean-variance analysis. *Production and Operations Management*, 27(6):1052–1073.
- Prashanth, L. A. and Ghavamzadeh, M. (2013). Actor-critic algorithms for risk-sensitive MDPs. *Advances in Neural Information Processing Systems*, 252–260.
- Levy, H. and Levy, M. (2003). Prospect theory and mean-variance analysis. *Review of Financial Studies*, 17(4):1015–1041.
- Li, D. and Ng, W. L. (2000). Optimal dynamic portfolio selection: Multiperiod mean-variance formulation. *Mathematical Finance*, 10(3):387–406.
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.
- Puterman, M. L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.



- Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. *Proceedings of the International Conference on Machine Learning*, 1889–1897.
- Sobel, M. J. (1982). The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19(4):794–802.
- Sobel, M. J. (1994). Mean-variance tradeoffs in an undiscounted MDP. *Operations Research*, 42(1):175–183.
- Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT press.
- Tamar, A., Di Castro, D., and Mannor, S. (2012) Policy gradients with variance related risk criteria. *Proceedings of the International Conference on Machine Learning*, 935–942.
- Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.
- Xia, L. (2016). Optimization of Markov decision processes under the variance criterion. *Automatica*, 73:269–278.
- Xia, L. (2020). Risk-sensitive Markov decision processes with combined metrics of mean and variance. *Production and Operations Management*, 29(12):2808–2827.
- Xie, T., Liu, B., Xu, Y., Ghavamzadeh, M., Chow, Y., Lyu, D., and Yoon, D. (2018). A block coordinate ascent algorithm for mean-variance optimization. *Advances in the Conference on Neural Information Processing Systems*, 1073–1083.
- Zhang, S., Liu, B., and Whiteson, S. (2021) Mean-variance policy iteration for risk-averse reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 10905–10913.
- Zhang, W.-G., Liu, Y.-J., and Xu, W.-J. (2012). A possibilistic mean-semivariance-entropy model for multi-period portfolio selection with transaction costs. *European Journal of Operational Research*, 222(2):341–349.
- Zhao, T., Hachiya, H., Niu, G., and Sugiyama, M. (2012). Analysis and improvement of policy gradient estimation. *Neural Networks*, 26:118–129.