Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Special Track on AI in FinTech

A Two-level Reinforcement Learning Algorithm for Ambiguous Mean-variance Portfolio Selection Problem

Xin Huang[1] and **Duan Li**[2*]
[1]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong
[2]School of Data Science, City University of Hong Kong, Hong Kong
huangxin@se.cuhk.edu.hk, dli226@cityu.edu.hk

Abstract

Traditional modeling on the mean-variance portfolio selection often assumes a full knowledge on statistics of assets' returns. It is, however, not always the case in real financial markets. This paper deals with an ambiguous mean-variance portfolio selection problem with a mixture model on the returns of risky assets, where the proportions of different component distributions are assumed to be unknown to the investor, but being constants (in any time instant). Taking into consideration the updates of proportions from future observations is essential to find an optimal policy with active learning feature, but makes the problem intractable when we adopt the classical methods. Using reinforcement learning, we derive an investment policy with a learning feature in a two-level framework. In the lower level, the time-decomposed approach (dynamic programming) is adopted to solve a family of scenario subcases where in each case the series of component distributions along multiple time periods is specified. At the upper level, a scenario-decomposed approach (progressive hedging algorithm) is applied in order to iteratively aggregate the scenario solutions from the lower layer based on the current knowledge on proportions, and this two-level solution framework is repeated in a manner of rolling horizon. We carry out experimental studies to illustrate the execution of our policy scheme.

1 Introduction

The modern portfolio selection theory has started to play an important role in both academia and industry of finance, after the seminal work of [Markowitz, 1952] who proposed the renowned mean-variance (MV) portfolio selection framework. After almost half century of struggle, [Li and Ng, 2000] succeeded in extending the Markowitz's static model into a multi-period setting and derived an analytical optimal portfolio policy through a novel way that embeds the time-inconsistent dynamic MV portfolio selection problem into a series of time-consistent control problems that are able to be solved by dynamic programming (DP). Most of the existing results in portfolio selection, including the dynamic MV problem with no-shorting constraint [Cui *et al.*, 2014] and mean-CVaR [Strub *et al.*, 2019], are based on the assumption that the investors have a full knowledge about the statistics of returns, while in the pure MV case, the knowledge of means and covariances of returns of risky assets will be sufficient. However, this assumption is not often the case in the real investment practice. In fact, the market condition is always time-varying and even for a short stationary time period, its statistical nature could still be unclear to investors. Therefore, the portfolio policy with a learning feature becomes more demanding than the classical one relying only on the historical market information. Especially we can invoke nowadays, for this purpose, the type of reinforcement learning (RL) algorithms which have been rapidly developed on solving dynamic decision-making problems ([Sutton and Barto, 2018; Bertsekas, 2019]).

In this paper, we consider an ambiguous MV portfolio selection problem with a mixture distribution on the returns of risky assets, where the proportions of different component distributions are assumed to be *unknown* to the investor, but being *constants* (in any time instant). It is reasonable to adopt such a framework where all component distributions in the mixture model are *known* to the investor in advance while the proportions are not, as each component distribution could be linked with a certain market mode (such as bull, bear, or neutral), and investors could have knowledge on each specific market mode from their past experience. On the other hand, it is usually difficult to estimate the probability by which a certain mode will occur in the future. Such an assumption has been already adopted in the literature, see, for example, [Zhu *et al.*, 2014] who focused on a robust mean-CVaR problem. Moreover, given the prior belief of the investor, updating the proportions through foreseeable future observations during the process of solving such an ambiguous problem is indispensable for deriving an optimal policy with an *active*-learning feature. Unfortunately, such a coupling makes the solution process intractable by traditional methods such as DP, due to the very nonlinear and nonseparable nature of learning techniques in, for example, the expectation maximization (EM) algorithm [McLachlan and Peel, 2004] about upgrading the components' weights. We instead develop in this paper a two-level framework in order to derive an invest-

*Contact Author

ment policy with a learning feature through the combination of i) the time-decomposed approach DP in the lower level for solving in parallel a family of modified scenario subproblems where all the possible sequences of component distributions are enumerated and listed in different scenarios and ii) a scenario-decomposed method, progressive hedging algorithm of [Rockafellar and Wets, 1991], in the upper level in order to iteratively aggregate the scenario solutions from lower layer based on the ambiguous proportions. Learning is finally achieved in a manner of rolling horizon.

2 Problem Formulation

Consider a financial market with N risky assets and one risk-free asset, and a finite investment horizon T. For $t = 0, 1, \ldots, T-1$, the total return of the riskless asset at time t, denoted by r_t, is deterministic and known, while the total return of risky assets, denoted by $\mathbf{e}_t = (e_t^1, \ldots, e_t^N)' \in \mathbb{R}^N$, is random and assumed to be independently and identically distributed with a finite mixture model of M components,

$$\mathbf{e}_t \sim \sum_{m=1}^{M} p_m^* D_m, \tag{1}$$

where D_m stands for the mth component distribution with given mean μ_m and covariance Σ_m, and all $p_m^* \in [0,1]$ as the corresponding mixed proportions satisfying $\sum_m p_m^* = 1$ are assumed to be *fixed* but *unknown* to the investor, leading to an uncertain statistics of the random returns. In this ambiguous situation, an MV investor seeks to develop some investment strategy $\mathbf{u}_t = (u_t^1, \ldots, u_t^N)' \in \mathbb{R}^N$, where each u_t^n represents the dollar amount invested in the risky asset n (hence $(x_t - \mathbf{1}'\mathbf{u}_t)$ goes to the risk-free asset for the wealth level x_t and $\mathbf{1} \in \mathbb{R}^N$ is the all-one column vector), so that the following dynamic MV problem with ambiguity is optimized,

$$\begin{aligned}\big(AMV(\epsilon)\big) \quad & \min_{\mathbf{u}_t, \forall t} \ \mathrm{Var}_{\mathbf{e},\mathfrak{p}}(x_T|I_0) \\ & \text{s.t.} \ \ \mathbb{E}_{\mathbf{e},\mathfrak{p}}[x_T|I_0] \geq \epsilon \\ & \qquad x_{t+1} = r_t x_t + \mathbf{P}_t'\mathbf{u}_t \\ & \qquad \mathfrak{p}_{t+1} = g(I^{t+1}),\ t = 0, 1, \ldots, T-1,\end{aligned}$$

where ϵ is the threshold of the expected final wealth specified by the investor, $\mathbf{P}_t = (e_t^1 - r_t, \ldots, e_t^N - r_t)' = \mathbf{e}_t - r_t\mathbf{1} \in \mathbb{R}^N$ is known as the random excess total return at time t, and the notations $\mathrm{Var}_{\mathbf{e},\mathfrak{p}}$ and $\mathbb{E}_{\mathbf{e},\mathfrak{p}}$ are used to emphasize that the variance and expectation are conducted not only on the randomness from the returns but also on the ambiguity from the proportions of the underlying mixture distribution. Note that the former uncertainty is *irreducible* as it is inherent in the risky asset by nature, while the latter is due to the lack of knowledge of the investor thus being *reducible* through learning. Besides, $I_0 = \{(D_m)_m, x_0, \mathfrak{p}_0\}$ represents the initial information set at time 0 which are given and contains knowledge of each component distribution, the initial wealth, and the prior belief of the investor, denoted by $\mathfrak{p}_0 = (p_{01}, \ldots, p_{0M})'$, on the ambiguous component proportions. For example, the investor may start with $p_{0m} = 1/M$ for all $m = 1, \ldots, M$. During the whole investment period, the wealth level and assets' prices will be observed. Given the pairs of newly-observed asset's price S_{t+1}^n and the previous one S_t^n for every risky asset n, the total return at time $t+1$ can be computed through

$$\mathbf{e}_{t+1} = \Big(\frac{S_{t+1}^1}{S_t^1}, \ldots, \frac{S_{t+1}^N}{S_t^N}\Big)' \in \mathbb{R}^N. \tag{2}$$

For simplicity, the new information set will directly include the realized return instead of the observed asset price and thus it satisfies the dynamics $I_{t+1} = \{I_t, \mathbf{e}_{t+1}, x_{t+1}, \mathbf{u}_t\}$. The posterior belief on the ambiguous proportions, denoted by $\mathfrak{p}_{t+1}$, is then updated based on I_{t+1} as in the nonlinear function g through EM algorithm that has been widely adopted to numerically estimate the unknown parameters (here for the unknown proportions) in the mixture model.

To solve $\big(AMV(\epsilon)\big)$, we first borrow the idea from [Li and Ng, 2000] to convert the problem into an equivalent form,

$$\begin{aligned}\big(AE(\epsilon)\big) \quad & \min_{\mathbf{u}_t, \forall t} \ \mathbb{E}_{\mathbf{e},\mathfrak{p}}[(x_T - \epsilon)^2|I_0] \\ & \text{s.t.} \ \ \mathbb{E}_{\mathbf{e},\mathfrak{p}}[x_T|I_0] = \epsilon, && (3) \\ & \qquad x_{t+1} = r_t x_t + \mathbf{P}_t'\mathbf{u}_t, && (4) \\ & \qquad \mathfrak{p}_{t+1} = g(I_{t+1}), && (5) \\ & \qquad t = 0, 1, \ldots, T-1.\end{aligned}$$

However, (5) is still intractable mainly due to the very nonlinear and nonseparable characteristics in the EM learning process. Thus, we bypass the difficulty by dropping this learning constraint (5) and replacing it with a rolling-horizon way for proportions updating. The two-level framework we are going to propose in the next section is designed to solve a series of truncated problems starting from time τ, $\tau = 0, 1, \ldots, T-1$, to the fixed terminal time T as given below,

$$\begin{aligned}\big(AE_\tau^T(\epsilon)\big) \quad & \min_{\mathbf{u}_t, \forall t} \ \mathbb{E}_{\mathbf{e},\mathfrak{p}}[(x_T - \epsilon)^2|I_\tau] \\ & \text{s.t.} \ \ \mathbb{E}_{\mathbf{e},\mathfrak{p}}[x_T|I_\tau] = \epsilon \\ & \qquad x_{t+1} = r_t x_t + \mathbf{P}_t'\mathbf{u}_t \\ & \qquad t = \tau, \tau+1, \ldots, T-1.\end{aligned}$$

3 A Two-level Portfolio Policy Scheme

The core feature of the two-level idea generally lies in separating the ambiguity on proportions from the randomness in returns, instead of coupling them together when dealing with the overall problem. Since the two-level policy scheme relies not only on the well-known DP for solving portfolio problems under each specific scenario but also on the progressive hedging algorithm (PHA) with which some readers may not be familiar, we will first provide some preliminaries on both topics and for more details we suggest referring to [Li and Ng, 2000] and [Rockafellar and Wets, 1991], respectively.

3.1 Classical Mean-variance Portfolio Policy

If there is no ambiguity on the proportions p_m^*, $\big(AMV(\epsilon)\big)$ reduces to a classical problem with *sufficient* knowledge on the returns (i.e., the mean and covariance of the mixture distribution), and can be solved analytically by following [Li and Ng, 2000] directly. Obviously, each scenario subproblem of $\big(AE_0^T(\epsilon)\big)$ under our ambiguous setting can be just counted as one special case of [Li and Ng, 2000]. More precisely, a

scenario j in our paper represents a sequence of component distributions along the time horizon of length T. For instance, one extreme scenario could be that the first component distribution happens T times, of which the scenario probability is $(p_1^*)^T$ if we know its exact proportion in the underlying mixture distribution. On top of that, we actually do not know at time 0 which scenario will really occur in the future. However, once given a specific scenario j, we are fully aware of the values of $\mathbb{E}_j[\mathbf{e}_t]$ and $\mathbb{E}_j[\mathbf{e}_t\mathbf{e}_t']$ for all t as we assume, and the scenario subproblem we are facing is

$$\begin{aligned}\big(AE_j(\epsilon)\big) \quad & \min_{\mathbf{u}_t,\forall t} \ \mathbb{E}_j[(x_T-\epsilon)^2|I_0] \\ & \text{s.t.} \ \ \mathbb{E}_j[x_T|I_0]=\epsilon \qquad (6)\\ & \quad x_{t+1}=r_t x_t+\mathbf{P}_t'\mathbf{u}_t,\ t=0,1,\dots,T-1,\end{aligned}$$

where the notation $\mathbb{E}_j$ is used to emphasize that the expectation on each $\mathbf{e}_t$ is computed under the scenario j. We continue to follow the steps in [Li and Ng, 2000] to attach the equality constraint (6) of the final wealth into the objective function and generate the corresponding Lagrangian problem with a Lagrangian multiplier $\lambda_j > 0$,

$$\begin{aligned}&\big(L_j(\epsilon,\lambda_j)\big) \qquad h_j(\lambda_j;\epsilon):=\\ &\min_{\mathbf{u}_t,\forall t} \ \mathbb{E}_j[(x_T-\epsilon)^2|I_0]-\lambda_j(\mathbb{E}_j[x_T|I_0]-\epsilon)\\ &=\mathbb{E}_j\left[x_T^2-(2\epsilon+\lambda_j)x_T|I_0\right]+\epsilon^2+\lambda_j\epsilon\\ &\text{s.t.} \ \ x_{t+1}=r_t x_t+\mathbf{P}_t'\mathbf{u}_t,\ t=0,1,\dots,T-1.\end{aligned}$$

Solving $\big(L_j(\epsilon,\lambda_j)\big)$ is a standard application of DP that can be found in [Li *et al.*, 1998]. Furthermore, due to the convexity of the problem, maximizing the dual function $h_j(\lambda_j;\epsilon)$ w.r.t. λ_j would give rise the optimal feedback policy to $\big(AE_j(\epsilon)\big)$. We list the final results as follows,

$$\mathbf{u}_{tj}(x_t;\epsilon,\lambda_j^*)=-\mathbf{K}_{tj}x_t+\mathbf{V}_{tj}(2\epsilon+\lambda_j^*), \qquad (7)$$

where

$$\mathbf{K}_{tj} = \mathbb{E}_j^{-1}\left[\mathbf{P}_t\mathbf{P}_t'\right]\mathbb{E}_j[\mathbf{P}_t]r_t, \qquad (8)$$

$$\mathbf{V}_{tj} = \frac{\mathbb{E}_j^{-1}\left[\mathbf{P}_t\mathbf{P}_t'\right]\mathbb{E}_j\left[\mathbf{P}_t\right]}{2\left(\prod_{\tilde{t}=t+1}^{T-1} r_{\tilde{t}}\right)}, \text{ with } \prod_{\tilde{t}=T}^{T-1}(\cdot):=1, \qquad (9)$$

$$\lambda_j^* = \frac{\epsilon\,(1-\theta_j)-\delta_j x_0}{\theta_j/2}, \text{ and} \qquad (10)$$

$$\delta_j = \prod_{t=0}^{T-1} r_t\Big(1-\mathbb{E}_j[\mathbf{P}_t']\mathbb{E}_j^{-1}\left[\mathbf{P}_t\mathbf{P}_t'\right]\mathbb{E}_j[\mathbf{P}_t]\Big),$$

$$\theta_j = \sum_{t=0}^{T-1}\Big[\mathbb{E}_j[\mathbf{P}_t']\mathbb{E}_j^{-1}\left[\mathbf{P}_t\mathbf{P}_t'\right]\mathbb{E}_j[\mathbf{P}_t] \cdot \prod_{\tilde{t}=t+1}^{T-1}\Big(1-\mathbb{E}_j[\mathbf{P}_{\tilde{t}}']\mathbb{E}_j^{-1}\left[\mathbf{P}_{\tilde{t}}\mathbf{P}_{\tilde{t}}'\right]\mathbb{E}_j[\mathbf{P}_{\tilde{t}}]\Big)\Big],$$

Besides, $\mathbb{E}_j[\mathbf{P}_t]=\mathbb{E}_j[\mathbf{e}_t]-r_t\mathbf{1}$ and $\mathbb{E}_j\left[\mathbf{P}_t\mathbf{P}_t'\right]=\mathbb{E}_j\left[\mathbf{e}_t\mathbf{e}_t'\right]-r_t\mathbb{E}_j[\mathbf{e}_t]\mathbf{1}'-r_t\mathbf{1}\mathbb{E}_j[\mathbf{e}_t']+r_t^2\mathbf{1}\cdot\mathbf{1}'$.

However, the above results are in general *not* the feasible solutions to $\big(AE_0^T(\epsilon)\big)$, as they already presume the realizations of component distributions hereafter as if "stealing" the *future* information. The PHA we are going to introduce next is a general method to fix this issue.

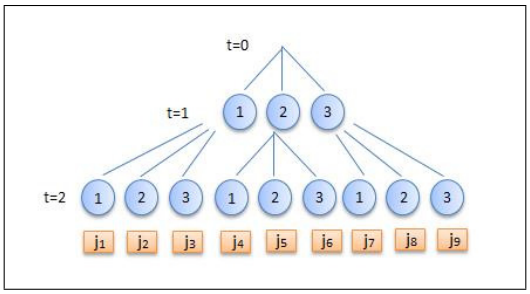

Figure 1: Permutation scenario tree

3.2 Brief Introduction on Progressive Hedging Algorithm

Let us consider an optimization problem expressed in a multi-scenario form,

$$(\mathcal{P}) \quad \min \ F(\mathbf{y})=\sum_{j\in S} s_j f_j(\mathbf{y}(j)) \text{ over all } \mathbf{y}\in\mathcal{A}\cap\mathcal{N},$$

where S represents a finite-scenario set and f_j is a specific objective function if the scenario j happens. A policy $\mathbf{y}$ is defined as a mapping from S to $\mathbb{R}^{NT}$, and $\mathbf{y}(j)=(y_0'(j),y_1'(j),\dots,y_{T-1}'(j))'$ denotes the actions taken at each time under the scenario j. The overall objective is adjusted by the scenario probabilities s_j's that are evaluated at the beginning. For the mixture distribution with M members, the scenario set in rolling horizon is time-varying but always contains a *permutation* of all components randomly allocated to returns of $(T-\tau)$ periods, therefore we have $|S_\tau|=M^{T-\tau}$ at each time τ, and the scenario probability is nothing but just the product of the estimated proportions of corresponding components due to the statistical independence of returns among different time periods. Figure 1 exhibits a permutation scenario tree at $\tau=0$ from an example of two-period problem with a three-component mixture distribution thus nine scenarios, where the numbers in each circle node represent the indices of the components that returns follow at related times. It is clear that, for example, the scenario j_6's probability is $s_{j_6}=p_{02}\cdot p_{03}$.

The key feature of a qualified policy of *multistage* decision-making problems, compared with those from static cases, is that apart from satisfying the admissible set $\mathcal{A}$ as usual, it must avoid leveraging on the anticipative message since the information flow comes by time, i.e., it should also satisfy the *non-anticipative* constraint $\mathcal{N}$ so that it can be *implemented* in reality. The approach used in PHA to force scenario-specific policies into an *implementable* one is to take *conditional* expectations based on the *subtree* starting from each node within each scenario *bundle*. To see this, let us revisit the problem in Figure 1. The scenario-based solution, denoted by $\mathbf{y}^0$, satisfies $\mathbf{y}^0(j)=\arg\min\ (\mathcal{P}_j)$ for every $j\in S=\{j_1,\dots,j_9\}$, where $(\mathcal{P}_j):=\{\min_y\ f_j(y) \text{ over } y\in A_j\}$ is the scenario subproblem (without foreseen restriction) and A_j constitutes $\mathcal{A}:=\{\mathbf{y}\mid\mathbf{y}(j)\in A_j\ \forall j\in S\}$. Then the implementable policy in this example, denoted by $\hat{\mathbf{y}}^0=((\hat{y}_0^0)',(\hat{y}_1^0)')':S\to\mathbb{R}^{2N}$, is computed such that i) when $t=0$, $\hat{y}_0^0(j)=\sum_{k\in S}s_k y_0^0(k)$ for each $j\in S$ and

ii) when $t = 1$ there are three bundles (subtrees) and for the scenarios belonging to the first bundle $j \in \mathcal{J}_1 = \{j_1, j_2, j_3\}$, we have $\hat{y}_1^0(j) = \sum_{k\in\mathcal{J}_1}[s_k/(s_{j_1} + s_{j_2} + s_{j_3})]y_1^0(k)$, and similar procedures to obtain the values of $\hat{y}_1^0$ on the bundles $\mathcal{J}_2 = \{j_4, j_5, j_6\}$ and $\mathcal{J}_3 = \{j_7, j_8, j_9\}$.

Based on the implementable policy $\hat{\mathbf{y}}^0$ aggregated from the scenario-specific solutions, PHA then starts to iteratively ($\nu = 0, 1, \ldots$) solve the augmented Lagrangian problems,

$$(\mathcal{P}_j^\nu)\ \min\ f_j(y) + y'\mathbf{w}^\nu(j) + \frac{1}{2}\eta|y - \hat{\mathbf{y}}^\nu(j)|_2^2 \text{ over } y \in A_j,$$

for every $j \in S$, where $\mathbf{w}^\nu$ serves as the Lagrangian multiplier with initial value $\mathbf{w}^0(j) = \mathbf{0}\ \forall j$, $|\cdot|_2$ denotes the 2-norm, and $\eta > 0$ is the predetermined penalty parameter. The optimal solutions of all $(\mathcal{P}_j^\nu)$ form $\mathbf{y}^{\nu+1}$, which will be aggregated again into an implementable policy $\hat{\mathbf{y}}^{\nu+1}$ used in $(\mathcal{P}_j^{\nu+1})$ for all j in the next iteration. The Lagrangian multiplier is then updated via $\mathbf{w}^{\nu+1} = \mathbf{w}^\nu + \eta(\mathbf{y}^{\nu+1} - \hat{\mathbf{y}}^{\nu+1})$. Eventually, $\hat{\mathbf{y}}^{\nu+1}$ will converge to the real optimum $\mathbf{y}^*$ of the primal problem $(\mathcal{P})$ if $f_j(y)$ is convex w.r.t. y for every j.

3.3 Two-level Reinforcement Learning Framework

Now it is ready to come up with our two-level (TL) portfolio policy algorithm. It is actually an online rolling-horizon scheme. Since the policy considered in this paper belongs to the type of pre-committed ones, we fix at each time τ the value of λ, the Lagrangian multiplier to deal with the equality constraint (3), at λ_0 that will be decided by a special scenario j_0 using only the prior knowledge of the investor. More specifically, let $\lambda_0 = \lambda_{j_0}^*$ in (10), where δ_{j_0} and θ_{j_0} are calculated based on $\mathbb{E}_{j_0}[\mathbf{e}_t] = \sum_m p_{0m}\mu_m$ and $\mathbb{E}_{j_0}[\mathbf{e}_t\mathbf{e}_t'] = \sum_m p_{0m}(\mu_m\mu_m' + \Sigma_m)$ for all $t = 0, 1, \ldots, T-1$. Note that p_{0m} is the prior belief of the investor on the mth component of the mixture distribution. The TL procedure is presented through multiple steps as follows:

Step 0: Importing Initial Information
We specify x_0 and ϵ, configure λ_0, and set $\tau = 0$.

Step 1: Solving the Truncated Problem at Time τ
This is the key part of the TL idea in order to solve $\big(AE_\tau^T(\epsilon)\big)$.

Step 1.1: lower level for scenario subproblems. We solve the problem $\big(AE_j(\epsilon)\big)$ subject to $t = \tau, \tau+1, \ldots, T-1$ by the classical techniques as summarized in the Subsection 3.1 for each scenario j. Let us denote the feedback solution by

$$\mathbf{u}_{tj}^0(x_t; \epsilon, \lambda_0) = -\mathbf{K}_{tj}^0 x_t + \mathbf{V}_{tj}^0(2\epsilon + \lambda_0). \tag{11}$$

Step 1.2: upper level for iterative aggregations. These are done in terms of policies' coefficients:

Step 1.2.0: initialization. Set $\nu = 0$, the penalty parameter η, and the tolerance level tol. The initial implementable feedback policy at time t, $\hat{\mathbf{u}}_t^0$, is just given by

$$\hat{\mathbf{u}}_t^0(x_t; \epsilon, \lambda_0) = -\hat{\mathbf{K}}_t^0 x_t + \hat{\mathbf{V}}_t^0(2\epsilon + \lambda_0), \tag{12}$$

where

$$\hat{\mathbf{K}}_t^0 = \sum_{j\in\mathcal{J}_t}(s_{\tau j}/\sum_{\tilde{j}\in\mathcal{J}_t} s_{\tau\tilde{j}})\mathbf{K}_{tj}^0, \tag{13}$$

$$\hat{\mathbf{V}}_t^0 = \sum_{j\in\mathcal{J}_t}(s_{\tau j}/\sum_{\tilde{j}\in\mathcal{J}_t} s_{\tau\tilde{j}})\mathbf{V}_{tj}^0, \tag{14}$$

and $\mathbf{K}_{tj}^0$ and $\mathbf{V}_{tj}^0$ are coming from the lower level in (11), and $s_{\tau j}$ is the scenario probability calculated based on the belief $\mathfrak{p}_\tau$ of the proportions at time τ, and $\mathcal{J}_t$ is a scenario bundle at time t of a $(T-\tau)$-stage permutation scenario tree. In fact, there is no difference for the aggregated results on different bundles from the same time stage of the permutation scenario tree structure, since the remaining subtrees starting from any node at the same time stage are actually identical. Therefore, we do not need to specify a bundle-oriented subscript for the coefficients $\hat{\mathbf{K}}_t^0$ and $\hat{\mathbf{V}}_t^0$ in (13) and (14), while only a time index t is enough for them. For simplicity, we could always select $\mathcal{J}_t$ as the first scenario bundle at each time stage t. The initial $\mathbf{w}_{tj}^0$ is set to be $\mathbf{0}$ for all t and all j. In fact, $\mathbf{w}_{tj}^\nu$ can be deemed as an affine function w.r.t. x_t given the parameters ϵ and λ_0, that is, $\mathbf{w}_{tj}^\nu(x_t; \epsilon, \lambda_0) = \mathbf{W}_{tj}^\nu x_t + \mathbf{X}_{tj}^\nu(2\epsilon + \lambda_0)$ with $\mathbf{W}_{tj}^0$ and $\mathbf{X}_{tj}^0$ being zero vectors for $\nu = 0$.

Step 1.2.1: solve and aggregate. In the following we are going to deal with the scenario-based augmented Lagrangian problem formed at time τ for different scenarios (note that we have already attached the equality constraint $\mathbb{E}_j[x_T|I_\tau] = \epsilon$ into the objective by the Lagrangian multiplier λ_0 and simplified the expression of the following),

$$\begin{aligned}\big(LAE_j^\nu(\epsilon)\big)\quad \min_{\mathbf{u}_t,\forall t}\ & \mathbb{E}_j\left[x_T^2 - (2\epsilon + \lambda_0)x_T|I_\tau\right] + \epsilon^2 + \lambda_0\epsilon \\ & + \sum_{t=\tau}^{T-1}\mathbf{u}_t'\mathbf{w}_{tj}^\nu + \sum_{t=\tau}^{T-1}\frac{1}{2}\eta|\mathbf{u}_t - \hat{\mathbf{u}}_t^\nu|_2^2 \\ \text{s.t.}\quad & x_{t+1} = r_t x_t + \mathbf{P}_t'\mathbf{u}_t, \\ & t = \tau, \tau+1, \ldots, T-1.\end{aligned}$$

It turns out that, given $\hat{\mathbf{u}}_t^\nu = -\hat{\mathbf{K}}_t^\nu x_t + \hat{\mathbf{V}}_t^\nu(2\epsilon + \lambda_0)$ and $\mathbf{w}_{tj}^\nu = \mathbf{W}_{tj}^\nu x_t + \mathbf{X}_{tj}^\nu(2\epsilon + \lambda_0)$, the optimal solution of $\big(LAE_j^\nu(\epsilon)\big)$ can be obtained analytically by DP [1] again, leading to the following result, denoted by $\mathbf{u}_{tj}^{\nu+1}(x_t; \epsilon, \lambda_0)$, which is treated as the scenario-specific policy from $\big(AE_j^\nu(\epsilon)\big)$,

$$\mathbf{u}_{tj}^{\nu+1}(x_t; \epsilon, \lambda_0) = -\mathbf{K}_{tj}^{\nu+1}x_t + \mathbf{V}_{tj}^{\nu+1}(2\epsilon + \lambda_0), \tag{15}$$

and the two coefficients satisfy the backward recursions derived from adopting DP on solving $\big(LAE_j^\nu(\epsilon)\big)$,

$$\begin{aligned}\mathbf{K}_{tj}^{\nu+1} &= \big(2a_{(t+1)j}^\nu\mathbb{E}_j\left[\mathbf{P}_t\mathbf{P}_t'\right] + \eta\mathbf{I}\big)^{-1} \\ &\quad\cdot(2a_{(t+1)j}^\nu r_t\mathbb{E}_j[\mathbf{P}_t] + \mathbf{W}_{tj}^\nu + \eta\hat{\mathbf{K}}_t^\nu), \quad (16)\\ \mathbf{V}_{tj}^{\nu+1} &= \big(2a_{(t+1)j}^\nu\mathbb{E}_j\left[\mathbf{P}_t\mathbf{P}_t'\right] + \eta\mathbf{I}\big)^{-1} \\ &\quad\cdot(b_{(t+1)j}^\nu\mathbb{E}_j[\mathbf{P}_t] - \mathbf{X}_{tj}^\nu + \eta\hat{\mathbf{V}}_t^\nu), \quad (17)\end{aligned}$$

where $\mathbf{I}$ is an identity matrix with size N, and $a_{(t+1)j}^\nu$ and $b_{(t+1)j}^\nu$ are two elements in our optimal cost-to-go function of $\big(LAE_j^\nu(\epsilon)\big)$ with the terminal conditions $a_{Tj}^\nu = b_{Tj}^\nu = 1$

[1] We ignore the derivation details and only list some necessary recursive formulas here, as this is a standard application of DP.

for all j and all ν, and they satisfy the backward recursions

$$\begin{aligned}
a_{tj}^{\nu+1} &= a_{(t+1)j}^{\nu}\Big[r_t^2 - 2r_t\mathbb{E}_j[\mathbf{P}_t']\mathbf{K}_{tj}^{\nu+1} \\
&\quad +(\mathbf{K}_{tj}^{\nu+1})'\mathbb{E}_j\left[\mathbf{P}_t\mathbf{P}_t'\right]\mathbf{K}_{tj}^{\nu+1}\Big] \\
&\quad +\frac{1}{2}\eta\Big[(\mathbf{K}_{tj}^{\nu+1})'\mathbf{K}_{tj}^{\nu+1} - 2(\mathbf{K}_{tj}^{\nu+1})'\hat{\mathbf{K}}_t^{\nu} \\
&\quad +(\hat{\mathbf{K}}_t^{\nu})'\hat{\mathbf{K}}_t^{\nu}\Big] - (\mathbf{K}_{tj}^{\nu+1})'\mathbf{W}_{tj}^{\nu}, \\
b_{tj}^{\nu+1} &= 2a_{(t+1)j}^{\nu}(\mathbf{K}_{tj}^{\nu+1})'\mathbb{E}_j\left[\mathbf{P}_t\mathbf{P}_t'\right]\mathbf{V}_{tj}^{\nu+1} + \eta(\mathbf{K}_{tj}^{\nu+1})'\mathbf{V}_{tj}^{\nu+1} \\
&\quad -2a_{(t+1)j}^{\nu}r_t\mathbb{E}_j[\mathbf{P}_t']\mathbf{V}_{tj}^{\nu+1} + (\mathbf{K}_{tj}^{\nu+1})'\mathbf{X}_{tj}^{\nu} \\
&\quad -(\mathbf{V}_{tj}^{\nu+1})'\mathbf{W}_{tj}^{\nu} - \eta(\mathbf{K}_{tj}^{\nu+1})'\hat{\mathbf{V}}_t^{\nu} \\
&\quad -\eta(\mathbf{V}_{tj}^{\nu+1})'\hat{\mathbf{K}}_t^{\nu} + \eta(\hat{\mathbf{K}}_t^{\nu})'\hat{\mathbf{V}}_t^{\nu} \\
&\quad -b_{(t+1)j}^{\nu}\mathbb{E}_j[\mathbf{P}_t']\mathbf{K}_{tj}^{\nu+1} + b_{(t+1)j}^{\nu}r_t.
\end{aligned}$$

Based on $\mathbf{u}_{tj}^{\nu+1}$ in (15), we then get the related implemented policy denoted by $\hat{\mathbf{u}}_t^{\nu+1}$ with coefficients similar as in (12),

$$\hat{\mathbf{u}}_t^{\nu+1}(x_t;\epsilon,\lambda_0) = -\hat{\mathbf{K}}_t^{\nu+1}x_t + \hat{\mathbf{V}}_t^{\nu+1}(2\epsilon+\lambda_0), \qquad (18)$$

where $\hat{\mathbf{K}}_t^{\nu+1} = \sum_{j\in\mathcal{J}_t}(s_{\tau j}/\sum_{\tilde{j}\in\mathcal{J}_t}s_{\tau\tilde{j}})\mathbf{K}_{tj}^{\nu+1}$, and $\hat{\mathbf{V}}_t^{\nu+1} = \sum_{j\in\mathcal{J}_t}(s_{\tau j}/\sum_{\tilde{j}\in\mathcal{J}_t}s_{\tau\tilde{j}})\mathbf{V}_{tj}^{\nu+1}$. Finally, the Lagrangian multiplier $\mathbf{w}_{tj}^{\nu+1}$ is updated through

$$\mathbf{w}_{tj}^{\nu+1}(x_t;\epsilon,\lambda_0) = \mathbf{W}_{tj}^{\nu+1}x_t + \mathbf{X}_{tj}^{\nu+1}(2\epsilon+\lambda_0), \qquad (19)$$

where coefficients satisfy $\mathbf{W}_{tj}^{\nu+1} = \mathbf{W}_{tj}^{\nu} + \eta(\hat{\mathbf{K}}_t^{\nu+1} - \mathbf{K}_{tj}^{\nu+1})$ and $\mathbf{X}_{tj}^{\nu+1} = \mathbf{X}_{tj}^{\nu} + \eta(\mathbf{V}_{tj}^{\nu+1} - \hat{\mathbf{V}}_t^{\nu+1})$.

Step 1.2.2: check the stopping criteria. Let us calculate the distance dis between two iterative results defined by coefficients $dis := \sum_{t=\tau}^{T-1}[\,|\hat{\mathbf{K}}_t^{\nu+1} - \hat{\mathbf{K}}_t^{\nu}|_2 + |\hat{\mathbf{V}}_t^{\nu+1} - \hat{\mathbf{V}}_t^{\nu}|_2 + 1/\eta^2(|\hat{\mathbf{W}}_t^{\nu+1} - \hat{\mathbf{W}}_t^{\nu}|_2 + |\hat{\mathbf{X}}_t^{\nu+1} - \hat{\mathbf{X}}_t^{\nu}|_2)\,]$. If $dis > tol$, we go back to the ***Step 1.2.1*** by replacing ν with $\nu+1$; otherwise, we stop and execute the resulted time-τ policy only, that is, the action outputted by our TL scheme at time τ based on the current wealth x_τ is

$$\mathbf{u}_\tau^{TL}(x_\tau;\epsilon,\lambda_0) = -\mathbf{K}_\tau^{TL}x_\tau + \mathbf{V}_\tau^{TL}(2\epsilon+\lambda_0), \qquad (20)$$

where $\mathbf{K}_\tau^{TL} = \hat{\mathbf{K}}_\tau^{\nu+1}$ and $\mathbf{V}_\tau^{TL} = \hat{\mathbf{V}}_\tau^{\nu+1}$ are given in (18).

Step 2: Learning and Rolling Horizon

If $\tau = T-1$, TL process is done; otherwise let us move forward to the time $\tau+1$. We update the information set $I_{\tau+1}$. In other words, we observe the new assets prices and calculate the realized total return $\mathbf{e}_{\tau+1}$ by (2). We then update the wealth level $x_{\tau+1}$ by (4) and proportions belief $\mathfrak{p}_{\tau+1}$ by Algorithm 1 based on $I_{\tau+1}$ to achieve the learning process of the investor, and it is then utilized to calculate the new scenario probabilities $s_{(\tau+1)j}$'s of a new permutation scenario tree with remaining time length $(T-\tau-1)$. Armed with the learning results, we are now going to solve the truncated problem at time $\tau+1$, that is, we go back to the **Step 1** and replace τ there by $\tau+1$ and take into account the new posterior belief $\mathfrak{p}_{\tau+1}$ on the proportions and proceed thereafter. □

Algorithm 1 Expectation maximization algorithm on estimating mixing proportions $\mathfrak{p}_{\tau+1}$ at time $\tau+1$

Input: Historical total returns $\mathbf{R}_t$ till time $\tau+1$, component distributions D_m, $m=1,\ldots,M$, and the tolerance level ξ
Output: Posterior mixing proportions $\mathfrak{p}_{\tau+1}$

1: Let $\kappa=0$ and assign an initial value (for example, former proportions $\mathfrak{p}_\tau$) to $\mathfrak{q}^\kappa := (q_1^\kappa,\ldots,q_M^\kappa)'$.
2: Calculate for every $t=1,\ldots,\tau+1$ and every m,

$$\gamma_m(\mathbf{R}_t|\mathfrak{q}^\kappa) := \frac{q_m^\kappa D_m(\mathbf{R}_t)}{\sum_{\tilde{m}=1}^M q_{\tilde{m}}^\kappa D_{\tilde{m}}(\mathbf{R}_t)}.$$

3: Set $q_m^{\kappa+1} = \frac{1}{\tau+1}\sum_{t=1}^{\tau+1}\gamma_m(\mathbf{R}_t|\mathfrak{q}^\kappa)$ for each m.
4: **if** $|\mathfrak{q}^{\kappa+1} - \mathfrak{q}^\kappa|_2 < \xi$ **then**
5: Set $\mathfrak{p}_{\tau+1} = \mathfrak{q}^{\kappa+1}$, and return $\mathfrak{p}_{\tau+1}$.
6: **else**
7: $\kappa \leftarrow \kappa+1$, and go back to line 2
8: **end if**

In the above we show our TL policy framework step by step on solving the ambiguous MV portfolio selection problem, where we could see a close interaction between the time-decomposed approach DP and the scenario-decomposed method PHA. Note that the convex nature of $\big(L_j(\epsilon,\lambda_j)\big)$ for each scenario j guarantees the convergence of the TL method.

4 Experimental Study

In this section, we illustrate how to calculate those key coefficients that appears in our TL policy through a concrete example, and compare with another non-learning (NL) policy on their gaps from the full-knowledge (FK) policy, in order to reveal the significance of learning in an ambiguous financial market. This experimental work is done in MATLAB by the Monte Carlo simulations of returns from a Gaussian mixture model (GMM) as the underlying mixture distribution.

Example 1 (Gaussian mixture model)**.** *Consider a market with three risky assets ($N=3$) and one risk-free asset, and an investment plan of forty periods ($T=40$). The total return of risky assets $\mathbf{e}_\tau$, $\tau=0,1,\ldots,T-1$, is assumed to i.i.d. follow a GMM $\sum_{m=1}^M p_m^*\mathcal{N}(\mu_m,\Sigma_m)$ with three components ($M=3$), where each $\mathcal{N}(\mu_m,\Sigma_m)$ represents a multivariate Gaussian distribution with distinct means μ_1 = (1.162, 1.246, 1.228)', μ_2 = (1,1,1)', and μ_3 = (0.962, 0.846, 0.828)' that could reflect different market modes (bull, neutral, and bear) but a same covariance for simplicity,*

$$\Sigma_m = \begin{bmatrix} 0.0146 & 0.0187 & 0.0145 \\ 0.0187 & 0.0854 & 0.0104 \\ 0.0145 & 0.0104 & 0.0289 \end{bmatrix},\ m=1,2,3.$$

It is reasonable to do so, since people in practice find that the means of stock returns fluctuate significantly when the market mode changes whereas the variances do not. The true proportions are set to be p_1^* = 0.5, p_2^* = 0.2 and p_3^* = 0.3 but are unknown to the investor. The total risk-free rate is r_τ = 1.04 for all τ. Suppose the investor's initial wealth is scaled to be x_0 = 1 and the target expected final wealth is ϵ = 2.

We analyse the above example from the following aspects.

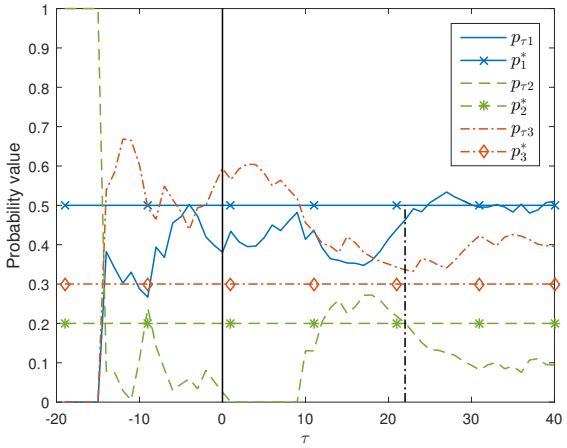

Figure 2: Proportions learning in both pre-investing ($\tau = -19$ to 0) and rolling-horizon ($\tau = 0$ to 40) periods

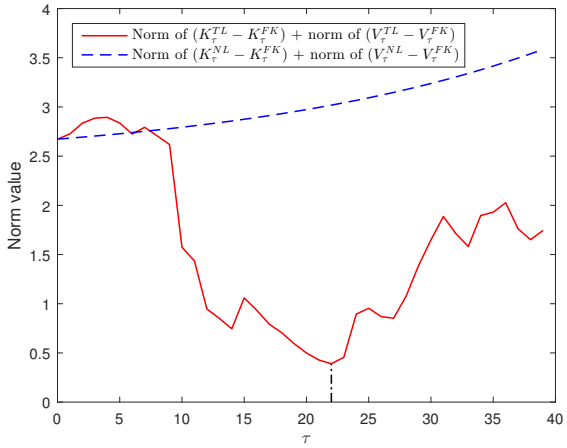

Figure 3: Norms of differences of coefficients in the two-level (non-learning) policy and the full-knowledge policy

Prior belief formation and proportions learning. Since assets' returns can be directly collected from the market even without any real investment, a suitable action for the investor in an ambiguous environment would be waiting and observing for a while before really entering into it. We adopt a *pre-investing* learning period with length $T/2 = 20$ for the investor to form the prior belief $\mathfrak{p}_0$ on the proportions before starting to do the investment at $\tau = 0$. In our experiment, this is done by sampling $T/2$ returns (corresponding to the time from $\tau = -(T/2-1)$ to $\tau = 0$) based on the *real* GMM and the prior belief is estimated by EM method (Algorithm 1) as well using these samples. This leads to $\mathfrak{p}_0 = (0.3808, 0.0251, 0.5941)'$. Another set of T samples are also generated for the rolling-horizon learning period in our TL framework after time 0. In fact at any time τ, the *whole* historical observations with size of τ+$T/2$ will be used to update the proportions. Figure 2 exhibits these learning outcomes, where each horizontal line represents the level of true proportion p_m^* while the waved curves represent the value movements of each $p_{\tau m}$ w.r.t. time τ based on one series of returns' samples in our experiment. The vertical dotted line points out the best estimation at $\tau = 22$. Note that the learning by EM in general will finally converge to their true values as long as T is larger enough.

Comparison among different policies. According to the market data given in Example 1 and the prior belief obtained above, we get $\lambda_0 = -0.0079$. After assigning η and tol, we calculate $\mathbf{K}_\tau^{TL}$ and $\mathbf{V}_\tau^{TL}$ in (20) for $\tau = 0, 1, \ldots, T-1$, the two core coefficients in our TL policy $\mathbf{u}_\tau^{TL}$. In order to evaluate the TL scheme, another two kinds of policies are also obtained: the FK policy as a benchmark and the NL policy. More precisely, $\mathbf{K}_\tau^{FK}$ and $\mathbf{V}_\tau^{FK}$ can be computed via $\mathbf{K}_{\tau j*}$ and $\mathbf{V}_{\tau j*}$ in (7) for a special scenario j^*, under which $\mathbb{E}_{j^*}[\mathbf{e}_t]$ and $\mathbb{E}_{j^*}[\mathbf{e}_t\mathbf{e}_t']$ are determined by real p_m^*'s. By replacing j^* with j_0 that depends only on $\mathfrak{p}_0$, we get $\mathbf{K}_\tau^{NL}$ and $\mathbf{V}_\tau^{NL}$ of the NL policy. Note that the FK policy cannot really be implemented in the ambiguous market and the NL policy treats the prior proportions as the "true" ones (thus no learning behaviors occur ever since). We finally demonstrate the improvement on the policy derivation with learning. This is done by calculating 2-norm of differences of key coefficients among three types of policies above and the results are listed in Figure 3. We see that our TL framework benefits from learning in general as it produces in the ambiguous environment a better policy that is closer to the benchmark, while the NL case keeps accumulating errors as time goes by. On the other hand, however, the TL policy is also easily affected by the learning outcome, since its closest distance to the FK policy occurs at the same time when the proportions estimation is the best (when $\tau = 22$ with $\mathfrak{p}_{22} = (0.4627, 0.2014, 0.3360)'$), but after that the performance becomes worse due to the fluctuation on learning. This phenomenon is also detected when we change the pre-learning time length or the investment time horizon.

5 Conclusion

In this paper, we deal with an ambiguous mean-variance portfolio selection problem where the returns of risky assets follow a mixture distribution, and each component distribution is completely given while their proportions are not. Although this problem can be formulated as usual, demanding on the successive learning of the unknown proportions in the solution process makes it intractable by the conventional methods. We propose a two-level (TL) scheme from reinforcement learning prospective to derive suboptimal policies but with learning feature from a type of amended problems. Our TL approach combines the exact dynamic programming adopted in the lower level to deal with the uncertainty from the random returns given a certain series of base distributions along the future times, and a scenario-decomposed method progressive hedging algorithm applied in the upper layer to handle the uncertainty from the ambiguous proportions, and learning is achieved in a rolling-horizon way. Although several experiments have demonstrated the effectiveness of our TL algorithm, it could still be improved in the future by incorporating more active learning aspects, in the sense that all possible future learning results could be considered in advance when solving the approximated primal problem at time 0.

References

[Bertsekas, 2019] Dimitri P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2019.

[Cui *et al.*, 2014] Xiangyu Cui, Jianjun Gao, Xun Li, and Duan Li. Optimal multi-period mean–variance policy under no-shorting constraint. *European Journal of Operational Research*, 234(2):459–468, 2014.

[Li and Ng, 2000] Duan Li and Wan-Lung Ng. Optimal dynamic portfolio selection: Multiperiod mean-variance formulation. *Mathematical Finance*, 10(3):387–406, 2000.

[Li *et al.*, 1998] Duan Li, Tsz-Fung Chan, and Wan-Lung Ng. Safety-first dynamic portfolio selection. *Dynamics of Continuous, Discrete and Impulsive Systems Series B: Applications and Algorithms*, 4(4):585–600, 12 1998.

[Markowitz, 1952] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[McLachlan and Peel, 2004] Geoffrey J McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.

[Rockafellar and Wets, 1991] R Tyrrell Rockafellar and Roger J-B Wets. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16(1):119–147, 1991.

[Strub *et al.*, 2019] Moris S Strub, Duan Li, Xiangyu Cui, and Jianjun Gao. Discrete-time mean-CVaR portfolio selection and time-consistency induced term structure of the CVaR. *Journal of Economic Dynamics and Control*, 108:103751, 2019.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[Zhu *et al.*, 2014] Shushang Zhu, Minjie Fan, and Duan Li. Portfolio management with robustness in both prediction and decision: A mixture model based learning approach. *Journal of Economic Dynamics and Control*, 48:1–25, 2014.