Mean–Variance Portfolio Selection by Continuous-Time Reinforcement Learning: Algorithms, Regret Analysis, and Empirical Study

Yilie Huang* Yanwei Jia† Xun Yu Zhou‡

December 9, 2024

Abstract

We study continuous-time mean–variance portfolio selection in markets where stock prices are diffusion processes driven by observable factors that are also diffusion processes yet the coefficients of these processes are unknown. Based on the recently developed reinforcement learning (RL) theory for diffusion processes, we present a general data-driven RL algorithm that learns the pre-committed investment strategy directly without attempting to learn or estimate the market coefficients. For multi-stock Black–Scholes markets without factors, we further devise a baseline algorithm and prove its performance guarantee by deriving a sublinear regret bound in terms of Sharpe ratio. For performance enhancement and practical implementation, we modify the baseline algorithm into four variants, and carry out an extensive empirical study to compare their performance, in terms of a host of common metrics, with a large number of widely used portfolio allocation strategies on S&P 500 constituents. The results demonstrate that the continuous-time RL strategies are consistently among the best especially in a volatile bear market, and decisively outperform the model-based continuous-time counterparts by significant margins.

Keywords: Portfolio selection; Dynamic mean–variance analysis; Reinforcement learning; Regret bound

*Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, 10027, USA. Email: yh2971@columbia.edu.

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong, N.T. Email: yanweijia@cuhk.edu.hk

‡Department of Industrial Engineering and Operations Research & Data Science Institute, Columbia University, New York, NY, 10027, USA. Email: xz2574@columbia.edu.

1 Introduction

In this paper, we study portfolio selection (or asset allocation) in dynamically traded markets for an investor who aims to achieve mean–variance efficiency in a finite investment horizon using reinforcement learning (RL). Since Markowitz (1952) introduced the mean–variance (MV) framework for static (single-period) portfolio choice, it has become one of the central topics in both modern portfolio theory and quantitative investment practice. However, despite its profound theoretical appeal and implications, practically implementing MV efficient strategies is challenging. First, most applications of the MV analysis are still restricted to the static setting to this day in practice (Kim et al., 2021), whereas applying static strategies myopically is surely inefficient from the dynamic perspective (Kim and Omberg, 1996). Second, accurately estimating the moments of asset returns is notoriously difficult, especially for the expected returns (Merton, 1980; Luenberger, 1998). Portfolios derived from analytical/numerical solutions of the MV problems in the static setting are known to be extremely sensitive to such estimation errors (Best and Grauer, 1991a,b; Britten-Jones, 1999), and become even worse for the dynamic one. Mitigating such errors and sensitivity and achieving MV efficiency in the dynamic environment remains largely an important open question.

Recent developments in machine learning have slowly but surely changed the practice of decision-making under uncertainty in a fundamental way, and reinforcement learning-based approaches have become more popular and better accepted in many application domains. One important feature of RL is to learn optimal actions (portfolio strategies in our case) *directly* via dynamic interactions with the environment (market) in a data-driven, model-free, and online fashion, *without* estimating any parameter of a statistical/probabilistic model. In the setting of this paper, "data" are both exogenous (including asset price data and other possibly time-varying but *observable/computable*, aggregate or individual covariates that affect the means and covariances of asset returns) and *endogenous* generated by an agent's strategic interactions with the unknown market. The dynamic nature of RL aligns with the setting of dynamically traded markets and farsighted investors. More importantly, learning portfolio choices directly while bypassing model estimation provides a powerful remedy to the aforementioned drawbacks of estimation errors and sensitivity inherent in the classical MV approach.

This paper studies portfolio selection in a continuously traded market for an agent with an MV preference. The agent observes stock prices and market factors but has minimum knowledge about the market and is unable to form a precise statistical model about the law of motions as assumed in the conventional financial economics literature. The only assumption about the market environment is that the stock prices are diffusion processes driven by observable factors that are also diffusion processes. The agent does not know the coefficients of these diffusion processes and aims to solve the continuous-time MV problem based on the observable data (such as the factors, stock prices, and wealth processes under different investment strategies) *only*.

The main contributions of this paper are three-fold. First, we propose RL algorithms for this problem by applying the general continuous-time RL theory developed by Wang et al. (2020) and Jia and Zhou (2022a,b). The foundation of the algorithms is to solve moment conditions arising from certain martingale conditions. Yet, these moment conditions are profoundly different from those employed in conventional econometrics in terms of actively generating *new* data for learning. Second, when the stock prices follow a multi-dimensional Black–Scholes environment without factors, we devise a more specific RL algorithm and prove its convergence. Moreover, we show that the algorithm achieves a *sublinear* regret in terms of the Sharpe ratio. Here, "regret" is the cumulative error over a number of learning episodes between the algorithm and the "oracle" one (i.e., the theoretically optimal one under the complete knowledge of the market environment). The sublinearity ensures that the RL algorithm will achieve nearly optimal results after a sufficiently long training period, even with an unknown market. This is the first model-free regret analysis (i.e., it is not based on estimating the market parameters) for continuous-time MV portfolio choice, whose proof is premised upon a delicate analysis of diffusion processes and stochastic approximation techniques. Finally, we modify this theoretically proven efficient algorithm for performance enhancement and practical implementations by turning it into, among others, online learning in real-time and incorporating leverage constraints and rebalancing frequency. Then, we carry out a comprehensive empirical study to compare the resulting RL strategies with 15 alternative popular methods using multiple performance metrics on S&P 500 constituents for the period 2000–2020, with 1990-2000 as the burnt-in period for training. These alternatives include the market portfolio, equally weighted portfolio, sample-based estimation, factor models, Bayesian estimation, distributional robust optimization, model-based continuous-time MV, linear predictive models, and two general-purpose RL algorithms. The performance criteria cover annualized return, Sharpe ratio and its variants, maximum drawdown, and recovery time. An unequivocal conclusion from the extensive empirical study is that our RL strategies outperform the classical model-based, plug-in continuous-time counterparts in all the metrics regardless of the market conditions. The RL strategies are also consistently among the best of all the methods, especially in a volatile and downturn market. The superiority of our approach does not stem from the use of predictive factors or complex neural networks but rather from our fundamentally distinct decision-making approach: learning the optimal strategy without learning the model.

Related Literature

There are two main directions in the literature for mitigating issues with sample-based estimation for (static) MV problems. The first is to develop more efficient estimators, including Bayesian inference and shrinkage estimators (James and Stein, 1992) to reduce estimation errors in probabilistic ways. The latter has been particularly popular for portfolio selection, e.g., shrinkage estimators for mean (Jorion, 1986; Black and Litterman, 1990), covariance matrix (Ledoit and Wolf, 2003, 2017), and covariance matrix of idiosyncratic error in factor models (Fan et al., 2008, 2012). The second direction takes the robust optimization approach. The idea is, instead of pinpointing a fixed model for optimization, to consider a *family* of models (also known as the ambiguity set) that contain the true but unknown model and optimize the objective in the worst scenario among these many models. Applying to MV portfolio selection, this approach modifies the original MV preference to a max-min MV objective (Garlappi et al., 2007; Goldfarb and Iyengar, 2003). Other works along this line include portfolio weight norm regularization (DeMiguel et al., 2009a), performance-based regularization (Ban et al., 2018), and many others. Most of the related formulations, however, need to set the width/radius of the ambiguity set as an exogenous hyper-parameters. Blanchet et al. (2022) employ a distributional robust approach and propose a statistical inference way of determining this uncertain set endogenously with a performance guarantee. However, robust approaches have been developed predominantly for static optimization, which becomes amply complex and intractable when dealing with a dynamic environment. On the other hand, DeMiguel et al. (2009b), in a thorough empirical study, show that most of these approaches do not consistently outperform the naïve equally-weighted portfolio. Blanchet et al. (2022) corroborate the competitive performance of the equally-weighted portfolio but find their distributional robust portfolios achieve a higher Sharpe ratio on average. However, they stop short of experimenting with other popular metrics such as maximum drawdown and recovery time. Above all, all these studies are on static MV problems. By contrast, we investigate forward-looking and dynamically planning investment policies, while providing a more comprehensive empirical study.

Dynamic portfolio selection has been studied extensively in the conventional financial economics and financial engineering literature. However, the research typically focuses on specific models, such as Zhou and Li (2000); Lim and Zhou (2002); Basak and Chabakauri (2010); Dai et al. (2021); Wachter (2002); Liu (2007); Gennotte (1986); Cvitanić et al. (2006), among many others, by assuming the agent has complete or at least partial knowledge about the models. In the latter case when, for instance, the agent knows that the stock prices follow geometric Brownian motions but the drift and/or volatility coefficients are unknown, she employs Bayesian learning to estimate the unknown coefficients. By contrast, the RL framework distinguishes itself by considering a "model-free" paradigm; that is, the agent only has the minimum knowledge about the market (such as that stock prices are diffusion processes) and learns optimal/efficient portfolio strategies directly which is not guided by statistical principles (such as Bayesian learning).

Despite the long history of machine learning research, applications to finance only started recently in the wake of AI and FinTech boom. For example, deep neural networks have been employed to study empirical asset pricing (Lettau and Pelger, 2020; Gu et al., 2020, 2021; Bianchi et al., 2021; Guijarro-Ordonez et al., 2021; Leippold et al., 2022; Chen et al., 2024). These works focus more on building nonlinear predictive models for asset returns or constructing trading signals using firms' characteristics to learn complex patterns in historical data. These predictive machine learning models have stimulated a significant increase in AI-driven funds (Bartram et al., 2021). However, RL has been hitherto barely used by the asset management industry (Snow, 2020), largely due to its lack of intrepretibility/explainability and theoretical guarantee even under the simplest Black–Scholes environment. This paper is part of the on-going effort that aims to provide rigorous underpinnings for continuous-time RL. Wang and Zhou (2020), which is closely related to the present paper, is probably the first to propose an RL algorithm for continuous-time MV portfolio selection built on the rigorous mathematical foundation established by Wang et al. (2020) with an entropy-regularized relaxed control formulation for general continuous-time RL. However, Wang and Zhou (2020) employ the commonly used mean–square temporal–difference error as the objective to perform policy evaluation (PE), which was later pointed out by Jia and Zhou (2022a) as a *wrong* objective. Instead, Jia and Zhou (2022a) prove some martingale condition that theoretically underpins their proposed offline and online PE algorithms. The present paper is based on Jia and Zhou (2022a) for PE and the subsequent Jia and Zhou (2022b) for policy gradient. Finally, a very preliminary version (without regret analysis and with a limited empirical study under a slightly different experimental setup) of this paper appeared earlier (Huang et al., 2022).

The rest of the paper is organized as follows. In Section 2, we present the MV formulation in a continuous-time multi-stock market environment with factors, and discuss the fundamental differences between the conventional model-based plug-in paradigm and that of RL. Section 3 explains the key steps in a general RL algorithm to solve the MV problem. Section 4 presents a baseline algorithm and its theoretical guarantee on the convergence of the learned policies along with a regret analysis in terms of the Sharpe ratio. Section 5 reports and discusses the results of an extensive comparative empirical study. Finally, Section 6 concludes. Proofs and additional numerical results are included in the appendix.

2 Dynamic Mean–Variance Portfolio Choice

In this section, we describe the market environment and the objective of an MV agent in the continuous-time setting with minimum assumption, and review two paradigms – those of the conventional plug-in and the RL, respectively.

2.1 Market environment and mean–variance agents

We first describe a general continuously traded market. There are $d+1$ assets, of which the 0-th asset is risk free whose price is $S^0(t)$ with a short interest rate $r(t)$. The other d assets are risky stocks whose prices at t are denoted by $S^1(t), \cdots, S^d(t)$. In addition, there are m additional *observable* covariates $F(t) \in \mathbb{R}^m$ that are associated with the interest rate, mean and covariance of the asset returns, referred to as the (market) factors.

In general, a *model* for the financial market makes further structural assumptions about the dynamics of asset prices and factors. For example, the celebrated Black–Scholes model assumes stock prices follow geometric Brownian motions, and Heston's model (Heston, 1993) stipulates stochastic volatility as factors. However, we do not assume agents know concrete forms of the market models, other than that

$S(t)=(S^0(t),S^1(t),\cdots,S^d(t))^\top$ and $F(t)$ are Itô's diffusions.[1] As a consequence, it is extremely difficult if not impossible to apply conventional statistical methods including Bayesian learning to estimate/learn the models.

Consider such a "model-free" agent with initial wealth x_0 and a pre-specified investment horizon $T>0$. We denote the agent's portfolio choice at time t by $u(t)=(u^1(t),u^2(t),\cdots,u^d(t))^\top\in\mathbb{R}^d$, where $u^i(t)$ is the discounted dollar amount (equivalently, $u^i(t)S^0(t)$ is the nominal dollar amount) invested in the i-th risky asset at time t, $1\leqslant i\leqslant d$. Denote by $\{x^u(t):0\leqslant t\leqslant T\}$ the discounted self-financing wealth process of the agent under a portfolio process $u\equiv\{u(t):0\leqslant t\leqslant T\}$. Then the agent's discounted wealth process satisfies the wealth equation

$$\mathrm{d}x^u(t)=\sum_{i=1}^d u^i(t)\frac{\mathrm{d}S^i(t)}{S^i(t)}-\boldsymbol{e}_d^\top u(t)\frac{\mathrm{d}S^0(t)}{S^0(t)},\tag{1}$$

where $\boldsymbol{e}_d=(1,\cdots,1)^\top\in\mathbb{R}^d$ is a d-dimensional unit vector, and $\frac{\mathrm{d}S^i(t)}{S^i(t)}$ is the return of the i-th asset.[2] Note that the wealth equation (1) follows from a simple fact that the change of wealth is caused by the changes in asset prices; hence it is very general, *independent* of any model about the stock prices or factors.

Note that for a small investor (a price taker), the asset prices and market factors are exogenous that are unaffected by her actions (portfolios). By contrast, a large investor's portfolio choice can alter the price and factor processes, e.g., through temporary or permanent price impact, and other frictions from the market microstructures. Such trading frictions and microstructures are an important part of the market environment, which is not assumed to be known by the investor. In our RL setting, the only assumption about the market is that $(S(t),F(t))$ are Itô's diffusions.

Given the investment horizon T, the agent aims to find MV efficient allocations in this dynamically traded market. As the continuous-time counterpart to the Markowitz problem (Markowitz, 1952), the classical model-based continuous-time MV problem is formulated as follows. *Assuming* a model for $S(t),F(t)$ and the wealth equation (1) are known and given, to minimize the variance of the portfolio while achieving a given expected target return:

$$\begin{aligned}&\min_u \operatorname{Var}\left(x^u(T)\right)\\&\text{subject to } \mathbb{E}\left[x^u(T)\right]=z\end{aligned}\tag{2}$$

where z is the target expected terminal wealth that is pre-specified at $t=0$ by the agent as part of the agent's preference. A larger z indicates that the agent pursues higher return and hence is less risk-averse.

[1] We assume these processes are all well-defined and adapted in a given filtered probability space $\left(\Omega,\mathcal{F},\mathbb{P};(\mathcal{F}_t)_{t\geqslant0}\right)$ satisfying the usual conditions. An Itô's diffusion belongs to a wide class of Markov processes, which can be represented as the solution to a stochastic differential equation driven by a (multi-dimensional) Brownian motion. It satisfies the strong Markov property and admits an infinitesimal generator. The developed model-free methodologies apply to a much more general class of stochastic processes like jump-diffusions or Markov chains. In this paper, we restrict the framework to Itô's diffusions because they are commonly used in the finance literature. We do not consider non-Markov processes here, which can be equivalently formulated as path-dependent Markov ones where certain factors can be summary statistics of the path history (e.g. the momentum).

[2] The nominal wealth process $x^u(t)S^0(t)$ satisfies $\mathrm{d}(x^u(t)S^0(t))=\sum_{i=1}^d S^0(t)u^i(t)\frac{\mathrm{d}S^i(t)}{S^i(t)}+(x^u(t)S^0(t)-\boldsymbol{e}_d^\top S^0(t)u(t))\frac{\mathrm{d}S^0(t)}{S^0(t)}$. Applying stochastic calculus leads to (1).

As a remark, the above is not a standard stochastic control problem (as with e.g., the expected utility maximization) due to the presence of the variance term in (2). This term causes time-inconsistency so the dynamic programming principle does not apply directly. Zhou and Li (2000) introduce a method of using a Lagrange multiplier w to transform the problem into an unconstrained expected quadratic (dis)utility minimization problem:

$$\min_u \mathbb{E}[(x^u(T) - w)^2] - (w - z)^2 \tag{3}$$

and then finding a proper multiplier w to enforce the mean constraint. This problem is a standard stochastic control problem and is time consistent, whose solution gives rise to a *pre-committed* investment strategy to the original problem (2). See a recent survey He and Zhou (2022) on pre-committed strategies and other types of strategies under time-inconsistency.

2.2 The conventional plug-in paradigm

The solution to the asset allocation problem such as (2) can be found or computed when the market model is completely specified, thanks to the well-developed stochastic control methodologies. How to mathematically solve (2) and what the economic implications are have been the focus of conventional research on quantitative finance and financial economics. Conceptually, these works take the rational expectation point of view so that agents can form their belief about the market environment correctly and, hence, behave optimally.

Practically, however, agents are always limited by their knowledge about the "true" market model (let there be one). The traditional approach to asset allocation (or indeed more general decision making problems) is to first propose and estimate a specific model and then plug the estimated model parameters (such as the drift and volatility of the stock price process) into the optimal solution for the corresponding model. This is usually referred to as the *model-based* approach which combines two steps/techniques: certain algorithms to estimate the model parameters (e.g. maximum likelihood or Kalman filtering) and certain algorithms to solve the stochastic control problem (e.g. analytical or numerical solutions to the Hamilton–Jacobi–Bellman equation). The two techniques have been developed quite separately from each other.

Such a paradigm, however, suffers from problems of model mis-specification, estimation errors caused by insufficiency of data and inadequacy of statistical methods, and sensitivity of optimal solutions to model parameters. In practice, the agent observes the asset price trajectories $S^0(t), \cdots, S^d(t)$ and/or the factor process $F(t)$ but also needs to know the specific form of a model to optimize her portfolio. Even with a simple market like the Black–Scholes market, there is an intrinsic difficulty in estimating the expected return accurately given any realistic sample size (Merton, 1980), also known as the "mean-blur" problem (Luenberger, 1998). The simulation study conducted by DeMiguel et al. (2009b) shows that even if the return distribution is correctly specified and estimated with various statistical techniques, 10-year data are still not sufficient for the plug-in method to get close to the theoretically optimal benchmark.[3]

2.3 The reinforcement learning paradigm

The RL version of the problem (2) is to solve it based only on the observable data including the price/-factor processes and the agent's own wealth processes under various portfolios, without any knowledge about the market other than that the underlying processes are Itô's diffusions. Hence, it addresses the task of a model-free agent rather than one having rational expectations. Recall that the plug-in approach first estimates (explores) and then optimizes (exploits), carrying out these two steps separately and *sequentially.* By contrast, the RL approach does exploration and exploitation *simultaneously* all the time. With RL, the agent interacts with the unknown (market) environment directly by trial and error, and improves strategies by incorporating the responses of the environment to the exploration. The classical stochastic control theory (Yong and Zhou 1999; Fleming and Soner 2006) stipulates that, (under mild conditions) for a Markov system, the optimal portfolio choice can be written as a *deterministic* function of the time and the state–factor variables, that is, $u(t) = \boldsymbol{u}^*(t, x(t), F(t))$ for some measurable, deterministic function $\boldsymbol{u}^*$, also known as a *policy.* Ultimately, both plug-in and RL approaches attempt to find this function, but in profoundly different ways. The former estimates a model for the market, and then solves a potentially high dimensional Hamilton–Jacobi–Bellman (HJB) equation, whereas the latter skips the intermediary step of estimating the model parameters altogether, approximates the policy function by a suitable class of functions with finite-dimensional parameters (using polynomials, spline functions, or neural networks), and learns/updates directly these parameters through exploration. It goes without saying that estimating market coefficients and solving HJB equations are of great importance in their own right, including the provision of insights and guidance about the structure of optimal policies. But the idea of RL is premised upon the notion that the ultimate purpose, or an end-to-end solution in asset allocation, is nothing more than to learn optimal or near-optimal policies from observed or (strategically) generated data. This idea underpins the *model-free* approach. We emphasize here that a model-free approach does not mean there are no models; rather, there is an underlying *structural* model (e.g., a Markov chain, an Itô diffusion, or a jump-diffusion) for the data-generating process but we do not know the model parameters, *nor do we attempt to estimate them.* Any provable performance guarantee including regret bounds must be established upon this structural assumption *only.*

To sum up, the model-based approach specifies a class of models, pins down one of them with certain statistical criteria, and then optimizes under that model. By contrast, the RL approach specifies a class of policies and optimizes directly based on the decision criteria.

[3] More analysis on how estimation errors affect portfolio choice includes Best and Grauer (1991a); Britten-Jones (1999); Chan et al. (1999) and Chopra and Ziemba (2013).

3 Foundation of Reinforcement Learning Algorithms

A typical RL algorithm involves answering three questions: how to choose actions to strategically interact with the environment for the purpose of exploration, how to evaluate the performance of a given policy, and how to update the policy to improve its performance. We follow the framework of continuous-time RL in the work of Wang et al. (2020) and Jia and Zhou (2022a,b) to address these three questions, respectively. In applying those general results, there is an additional Lagrange multiplier w in (3) that needs to be learned in the current MV setting.

3.1 Deterministic versus stochastic policies

In the classical model-based setting, an optimal policy is a deterministic mapping from the time–state–factor triplet to an action (a portfolio choice in the context of asset allocation). However, when the market environment is unknown, the RL agent undergoes exploration by *randomizing* the policies in order to broaden the search space. Mathematically, these exploratory policies are now mappings that map time–state–factor triplets to probability (density) distributions on the action space:

$$\left\{\boldsymbol{\pi} : (t, x, F) \mapsto \boldsymbol{\pi}(\cdot|t, x, F) \in \mathcal{P}(\mathbb{R}^d)\right\},$$

where $\mathcal{P}(\mathbb{R}^d)$ denotes the set of all probability density functions on $\mathbb{R}^d$.[4] Note that the agent wealth process $x(t)$ and the factor process $F(t)$ are both observable (i.e. they are *data*) at any time t. Given a mapping $\boldsymbol{\pi}$, at each time t, a portfolio $u(t)$ is independently generated or sampled from the distribution given by $\boldsymbol{\pi}(\cdot|t, x(t), F(t))$, denoted by $u(t) \sim \boldsymbol{\pi}(\cdot|t, x(t), F(t))$. Such a rule $\boldsymbol{\pi}$ to generate portfolios is called a *stochastic policy*. Obviously, when $\boldsymbol{\pi}$ is a point mass (aka Dirac measure), it reduces to the conventional deterministic policy. Once $\boldsymbol{\pi}$ is specified, the portfolio processes to be *actually* executed could be sampled according to $u(t) \sim \boldsymbol{\pi}(\cdot|t, x(t), F(t))$ and the resulting wealth trajectories, while following (1), can be directly observed, both without requiring the knowledge of the market coefficients. We denote the wealth trajectories under a stochastic policy $\boldsymbol{\pi}$ by $x^{u^{\boldsymbol{\pi}}}$. The statistical properties of this process are described in Appendix A.1.

As the essence of RL is to balance exploration and exploitation while the former is needed during the entire time horizon, we further add an entropy regularizer to the objective function to encourage and, indeed, enforce exploration. The entropy regularization is closely related to the soft-max approximation in the RL literature (Ziebart et al., 2008; Haarnoja et al., 2018; Wang et al., 2020), as well as the choice model and the perturbed utility (Hotz and Miller, 1993; Fudenberg et al., 2015; Feng et al., 2017) in microeconomic theory.

[4] In this paper, we restrict randomization to distributions having density functions because they are the most commonly used and compatible with the entropy regularizer to be introduced momentarily. With more involved notation, the density functions can be replaced by distribution functions.

This leads to the following entropy regularized objective function:

$$\mathbb{E}\left[\left(x^{u^{\boldsymbol{\pi}}}(T)-w\right)^2+\gamma\int_0^T \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}}(t)|t, x^{u^{\boldsymbol{\pi}}}(t), F(t))\mathrm{d}t\right]-(w-z)^2 \tag{4}$$

where $\gamma \geqslant 0$ is an exogenous parameter, called a *temperature parameter*, that specifies the weight put on exploration. Clearly, a larger γ favors more exploration and vice versa.

Let us conclude this subsection by noting some important points about stochastic policies. In optimization broadly, randomization is taken for conceptual and/or technical reasons; see Zhou (2023) for an assay on this. In RL specifically, randomization is used primarily for exploration or information collection (e.g., ϵ-greedy policy for the bandit problem; see Sutton and Barto 2018), as randomization broadens search space and enables an agent to experience counterfactuals.[5] Intuitively, by trying out different trading portfolios the agent gets to know more about the market including market impact which in turn guides her to optimize gradually.

However, if the agent is a *small* investor, the current portfolio selection problem has a *distinctive* feature in this aspect. Recall that the (discounted) wealth equation is described by (1), where $\frac{\mathrm{d}S^i(t)}{S^i(t)}$ and $\frac{\mathrm{d}S^0(t)}{S^0(t)}$ are the (instantaneous) returns of the risky and risk-free assets respectively that can be observed directly from the market *without* having to know the market coefficients. So wealth change is jointly caused by portfolio choice and price movement in a *known*, multiplicative way. However, with a small investor, the price movement is purely *exogenous* and observable regardless of what portfolios she applies. Therefore, (1) reveals all the counterfactuals under alternative portfolios without having actually to execute them.[6] To wit, *there is no informational motive for exploration/randomization for a price taker*. That said, there are important *technical* reasons to use stochastic policies for learning. In general, randomization convexifies policy spaces and facilitates differentiation. Specifically, in this paper, we will apply the policy gradient algorithms developed in Jia and Zhou (2022b) to update policies, whereas the key idea of Jia and Zhou (2022b) is to turn the policy gradient into a policy evaluation problem which works *only* for stochastic policies. Therefore, we will train our algorithms using stochastic policies and implement portfolios with deterministic policies.[7]

3.2 Policy evaluation

Policy evaluation is a crucial step in estimating/predicting the payoff function of a given policy, based on which the agent decides how to update and improve the policy. In our case, it is to estimate the expected payoff (4) for a given stochastic policy $\boldsymbol{\pi}$, a given multiplier w and a given temperature parameter γ, based

[5]For example, in the bandit problem, randomization allows the agent to play a currently sub-optimal machine that otherwise would have never been played (and therefore whose information would have remained unknown).

[6]This is a very specific feature of this specific setting (a small investor), which is not owned by most stochastic control problems including portfolio choice with large investors.

[7]In RL terms, this is a type of *off*-policy learning (Sutton and Barto 2018, Chapter 6), i.e. we use stochastic policies – called the behavior policies – to improve deterministic policies which are the target policies.

on data only. Moreover, the policy evaluation requires learning the expected payoff starting from *any* initial time–state–factor triplet and hence calls for estimating the *entire* objective function (instead of functional *values* at some given triplets). Precisely, based on the Markov property, we need to learn a function J of (t, x, F), known as the value function, where

$$J(t, x, F; \boldsymbol{\pi}; w) = \mathbb{E}\left[\left(x^{u^{\boldsymbol{\pi}}}(T) - w\right)^2 + \gamma \int_t^T \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}}(s)|s, x^{u^{\boldsymbol{\pi}}}(s), F(s))\mathrm{d}s \middle| x^{u^{\boldsymbol{\pi}}}(t) = x, F(t) = F\right] - (w - z)^2.$$

Jia and Zhou (2022a) show that the value function is characterized by two conditions. First, it satisfies a known terminal condition: $J(T, x, F; \boldsymbol{\pi}; w) = (x - w)^2 - (w - z)^2$. Second, it maintains the following process

$$J(t, x^{u^{\boldsymbol{\pi}}}(t), F(t); \boldsymbol{\pi}; w) + \gamma \int_0^t \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}}(s)|s, x^{u^{\boldsymbol{\pi}}}(s), F(s))\mathrm{d}s,$$

where $u^{\boldsymbol{\pi}}(s) \sim \boldsymbol{\pi}(\cdot|s, x^{u^{\boldsymbol{\pi}}}(s), F(s))$, to be a martingale with respect to the filtration generated by $x^{u^{\boldsymbol{\pi}}}(s), F(s)$.

As computers cannot process learning functions that are infinite-dimensional objects, in RL, one uses function approximation to approximate the value function by a class of parameterized functions $J(\cdot, \cdot, \cdot; w; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a finite-dimensional parameter. The choice of approximators may depend on the special structure of each problem or be through neural networks. Note that, for a given policy $\boldsymbol{\pi}$ and a function approximator $J(\cdot, \cdot, \cdot; w; \boldsymbol{\theta})$, both $J(t, x^{u^{\boldsymbol{\pi}}}(t), F(t); w; \boldsymbol{\theta})$ and $\log \boldsymbol{\pi}(u^{\boldsymbol{\pi}}(s)|s, x^{u^{\boldsymbol{\pi}}}(s), F(s))$ can be computed by observable samples or data. Jia and Zhou (2022a) develop several data-driven ways to learn or update $\boldsymbol{\theta}$ based on the aforementioned martingality. In this paper, we will apply one of them that is consistent with the well-known temporal–difference (TD) learning: to force $\mathrm{d}J(t, x^{u^{\boldsymbol{\pi}}}(t), F(t); w; \boldsymbol{\theta}) + \gamma \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}}(t)|t, x^{u^{\boldsymbol{\pi}}}(t), F(t))\mathrm{d}t$ to be a "martingale difference sequence" so that it is orthogonal to any adapted process. More precisely, it means

$$\mathbb{E}\left[\int_0^T \mathcal{I}(t)\left\{\mathrm{d}J(t, x^{u^{\boldsymbol{\pi}}}(t), F(t); w; \boldsymbol{\theta}) + \gamma \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}}(t)|t, x^{u^{\boldsymbol{\pi}}}(t), F(t))\mathrm{d}t\right\}\right] = 0, \tag{5}$$

for all adapted processes $\{\mathcal{I}(t) : 0 \leqslant t \leqslant T\}$, called *test functions* (or *instrumental variables* in the econometrics literature). While theoretically one needs to choose infinitely many test functions, for implementation one can take $\mathcal{I}(t) = \frac{\partial}{\partial \boldsymbol{\theta}} J(t, x^{u^{\boldsymbol{\pi}}}(t), F(t); w; \boldsymbol{\theta})$ which is a vector having the same dimension as $\boldsymbol{\theta}$. The task of estimating $\boldsymbol{\theta}$ from the system of equations (5) can thus be accomplished by the well-developed *generalized methods of moments (GMM).*

3.3 Policy gradient

Now that we have learned the value function of a given stochastic policy, the next step is to improve the policy. For that, following the general gradient-based approach in optimization, we need to estimate the gradient of the value function with respect to the policy. However, the policy itself lies in an infinite dimensional space of probability distributions; so it is infeasible to compute the derivative upon policy directly. As in policy evaluation, we parameterize policies by a finite-dimensional vector $\boldsymbol{\phi}$:

$$\boldsymbol{\pi}^{\boldsymbol{\phi}} \equiv \boldsymbol{\pi}(\cdot|t, x, F; w; \boldsymbol{\phi}).$$

Denote by $J(\cdot,\cdot,\cdot;\boldsymbol{\pi}^{\boldsymbol{\phi}};w)$ the value function under $\boldsymbol{\pi}^{\boldsymbol{\phi}}$. It now suffices to consider $\frac{\partial}{\partial \boldsymbol{\phi}} J(0, x_0, F_0; \boldsymbol{\pi}^{\boldsymbol{\phi}}; w)$, the gradient of $J(0, x_0, F_0; \boldsymbol{\pi}^{\boldsymbol{\phi}}; w)$ in $\boldsymbol{\phi}$.

Jia and Zhou (2022b) derive the policy gradient representation as follows

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\phi}} J(0, x_0, F_0; \boldsymbol{\pi}^{\boldsymbol{\phi}}; w) \\
=&\mathbb{E}\Bigg[\int_0^T \left[\frac{\partial}{\partial \boldsymbol{\phi}} \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}(t)|t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t), F(t); w; \boldsymbol{\phi}) + \mathcal{H}(t)\right] \Big[\mathrm{d}J(t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t), F(t); \boldsymbol{\pi}^{\boldsymbol{\phi}}; w) \\
&\qquad + \gamma \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}(t)|t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t), F(t); w; \boldsymbol{\phi})\mathrm{d}t\Big]\Bigg],
\end{aligned} \tag{6}$$

for all test functions $\mathcal{H}$. Compared with Jia and Zhou (2022b), here we have added the test functions $\mathcal{H}$ (by virtue of (5)) to make the approximation of policy gradient more flexible. The right-hand side terms inside the expectation in (6) can be computed using observable state samples under the policy $\boldsymbol{\pi}^{\boldsymbol{\phi}}$ and the known parametric form $\boldsymbol{\pi}^{\boldsymbol{\phi}}$, together with an estimated value function from the policy evaluation step discussed in Section 3.2, without knowing the market coefficients.

3.4 Actor–critic learning by solving moment conditions

Alternating policy evaluation and policy gradient iteratively leads to what is called an actor–critic type of learning in RL. More precisely, there are three equations that need to be satisfied by the optimal value function, the optimal policy, and the Lagrange multiplier:

$$\begin{cases}
\mathbb{E}\left[\int_0^T \mathcal{I}(t)\left\{\mathrm{d}J(t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t), F(t); w; \boldsymbol{\theta}) + \gamma \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}(t)|t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t), F(t); w; \boldsymbol{\phi})\mathrm{d}t\right\}\right] = 0, \\
\mathbb{E}\Bigg[\int_0^T \left[\frac{\partial}{\partial \boldsymbol{\phi}} \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}(t)|t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t), F(t); w; \boldsymbol{\phi}) + \mathcal{H}(t)\right] \Big[\mathrm{d}J(t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t), F(t); w; \boldsymbol{\theta}) \\
\qquad + \gamma \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}(t)|t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t), F(t); w; \boldsymbol{\phi})\mathrm{d}t\Big] = 0, \\
\mathbb{E}\left[x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(T) - z\right] = 0.
\end{cases} \tag{7}$$

The first equation in (7) follows from the martingale condition (5) by substituting the policy by its approximation $\boldsymbol{\pi}^{\phi}$, with test function $\mathcal{I}$. The second equation follows from (6), implying that the gradient of the optimal value function with respect to the parameters $\boldsymbol{\phi}$ (with test function $\mathcal{H}$) is zero, which is the usual first-order condition for optimality. In applying (6) we replace the true value function under $\boldsymbol{\pi}^{\boldsymbol{\phi}}$ with its approximator, i.e. (with a slight abuse of notion),

$$J(t,x,F;w;\boldsymbol{\theta}) \approx J(t,x,F;\boldsymbol{\pi}^{\boldsymbol{\phi}};w).$$

The last equation in (7) is due to the constraint on the expected return imposed by the original MV problem (2).

The conditions in (7) are generally called *moment conditions*. Let us explain a few major differences between these conditions and those in the classical GMM literature. First, for usual GMM, data are exogenously given or generated by exogenous distributions. By contrast, except for the exogenous market factors $\{F(t) : 0 \leqslant t \leqslant T\}$,[8] the wealth and the portfolio choice data $\{(x(t), u(t)) : 0 \leqslant t \leqslant T\}$ are generated/sampled from a stochastic policy that needs to be learned through updating $\boldsymbol{\phi}$. This is one important and distinctive aspect of RL: data are both exogenous and *endogenous* and should be acquired actively and strategically. Specifically in the current context, we need to try different policies by varying possible $\boldsymbol{\phi}$'s, repeatedly sample portfolios from them and observe the resulting wealth trajectories in order to solve (7). Second, moment conditions in conventional GMM arising in economics and finance problems are usually based on optimality or equilibrium conditions of an economic model, where data are assumed to be generated from optimal or equilibrium decisions. For example, in testing the consumption-based capital asset pricing model under rational expectation (Hansen and Singleton, 1982), it is assumed that the consumption flow is optimal with respect to the environment and the household's preference. However, in our RL problem, the agent does not know the market environment; hence the data used in applying the moment conditions (7) are not produced from the optimal policy in general. Moreover, the target for an RL agent is to learn the optimal policy; by contrast, the target of an econometrician is to learn an economic model.

Lastly, from a technical perspective, the conventional GMM relies on numerical optimization algorithms that are usually based on sample average approximation. However, in our RL setting, usually only one sample trajectory generated under the given policy is used in one iteration step to evaluate the expectation in (7); so stochastic approximation algorithms are often used instead. Though stochastic approximation algorithms have slower convergence than gradient-based algorithms, they require much less storage and computation in each iteration and are generally flexible and easy to apply; see Section 4 for more discussions. In addition, the analysis of RL algorithms also focuses on the suboptimal gap (i.e., regret) of the learning procedure because suboptimal policies are used to generate data.

4 A Provably Efficient Algorithm for the Black–Scholes Market

Establishing a model-free theoretical guarantee of the efficiency of an RL algorithm is generally extremely hard, due to complicated function approximations (e.g., with neural networks) and possible non-stationarity

[8]While the asset prices $\{S(t) : 0 \leqslant t \leqslant T\}$ are also exogenous and observable data, we do not need them in our learning.

of state processes involved. In this section, we present an RL algorithm for a frictionless, multi-stock Black–Scholes market without any factor F, i.e., the stock prices follow multi-dimensional geometric Brownian motions. We prove that a stochastic approximation type algorithm with specific actor and critic function approximations converges to a Sharpe ratio maximizing policy, and derive a sublinear regret bound in terms of the Sharpe ratio. Our algorithm is model-free in the sense that it is based on the model-free characterization of the optimal policy (7); yet the proof depends on the specific Black–Scholes market structure. We leave the question of empirical performance to the next section, where the distributions of stock returns are unknown and unverifiable.

4.1 A baseline algorithm

To recap what was introduced in Section 3.4, an RL algorithm consists of approximating the value function (critic) and policy (actor), sampling/generating data, and updating/improving the approximation. Approximation or parametrization can be, in general, constructed through neural networks or by exploiting the specific problem structure, such as with the present case. A theoretical analysis of the exploratory MV problem with the Black–Scholes environment is presented in Appendix A.2. The *theoretical* optimal value function and optimal policy given by (8) and (9) involve the unknown model parameters so they cannot be used as the final solutions. However, the specific *forms* of these functions suggest that we can apply the following approximations for the value functions and stochastic policies:

$$J(t,x;w;\boldsymbol{\theta}) = (x-w)^2 e^{-\theta_3(T-t)} + \theta_2\left(t^2 - T^2\right) + \theta_1(t-T) - (w-z)^2, \tag{8}$$

and

$$\boldsymbol{\pi}(\cdot \mid t,x;w;\boldsymbol{\phi}) = \mathcal{N}\left(\cdot \mid -\phi_1(x-w), \phi_2 e^{\phi_3(T-t)}\right), \tag{9}$$

where $(\theta_1,\theta_2,\theta_3)^\top \in \mathbb{R}^3$ and $(\phi_1,\phi_2,\phi_3) \in \mathbb{R}^d \times \mathbb{S}^d_{++} \times \mathbb{R}$ are two sets of parameters, $w \in \mathbb{R}$ is the Lagrange multiplier, and $\mathcal{N}(\cdot \mid \mu, \Sigma)$ is the multivariate normal distribution with mean vector μ and covariance matrix Σ. If we are to completely reconcile the approximated solutions (8) and (9) with the theoretical solutions (26) and (27), then these parameters should not be entirely mutually independent. However, we do not enforce their relations based on the theoretical solutions and treat them largely independently in our learning procedure for generality and flexibility. One exception is that we let $\phi_3 = \theta_3$, inspired by (26) and (27). Moreover, in our algorithm, we set $\phi_3 = \theta_3$ to be a sufficiently large constant (a hyperparameter) without updating it, because it turns out this parameter plays no role in the convergence analysis (see Theorems 1–3). Thus, we denote $\boldsymbol{\theta} = (\theta_1,\theta_2)^\top \in \mathbb{R}^2$ and $\boldsymbol{\phi} = (\phi_1,\phi_2)^\top \in \mathbb{R}^d \times \mathbb{S}^d_{++}$ which, together with w, are to be updated and learned.

The *baseline* algorithm we devise relies on the whole trajectory, meaning that in each iteration, parameters $(\boldsymbol{\theta}, \boldsymbol{\phi}, w)$ are updated after the data generated during the entire episode $[0,T]$ are used. It is a stochastic approximation algorithm based principally on the moment conditions (7), with a modification to boost the numerical efficiency. Specifically, in applying (7) we reparameterize $\boldsymbol{\phi} = (\phi_1, \phi_2)$ to $\tilde{\boldsymbol{\phi}} = (\phi_1, \phi_2^{-1})$ and turn the second equation in (7) in terms of the gradient in ϕ_2 to

$$\begin{aligned} 0 =& \mathbb{E}\bigg[\int_0^T \bigg[\frac{\partial}{\partial \phi_2^{-1}} \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}(t)|t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t); w; \boldsymbol{\phi}) + \mathcal{H}(t) \bigg] \Big[\mathrm{d}J(t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t), w; \boldsymbol{\theta}) \\ & + \gamma \log \boldsymbol{\pi}(u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}(t)|t, x^{u^{\boldsymbol{\pi}^{\boldsymbol{\phi}}}}(t); w; \boldsymbol{\phi}) \mathrm{d}t \Big] \bigg]. \end{aligned} \tag{10}$$

The above follows from the chain rule and the fact that the extra term $\frac{\partial \phi_2^{-1}}{\partial \phi_2}$, resulting from the chain rule, is a deterministic, time-invariant constant, and hence can be removed. Thus, our stochastic approximation algorithm for the component ϕ_2 will be based on (10), the gradient in ϕ_2^{-1}, instead of ϕ_2 as in the original conditions (7). This trick of using the inverse covariance matrix will prove instrumental in the proof of our convergence results.

We use subscript n to represent the n-th iteration. For example, $\phi_{1,n}$ is the value of the parameter ϕ_1 in its n-th iteration. At the first iteration $n = 1$, we initialize $\boldsymbol{\theta}_1 = (\theta_{1,1}, \theta_{2,1})^\top$, $\boldsymbol{\phi}_1 = (\phi_{1,1}, \phi_{2,1})^\top$ and w_1 to be some constants. At the $(n+1)$-th iteration, with the current parameters $(\boldsymbol{\theta}_n, \boldsymbol{\phi}_n, w_n)$, we use the policy $\boldsymbol{\pi}(\cdot \mid t, x; w_n; \boldsymbol{\phi}_n)$ determined by (9) to generate the portfolio–wealth process $\{(u_n(t), x_n(t)) : 0 \leqslant t \leqslant T\}$, where x_n satisfies (1) under $u = u_n$ with $u_n(t) \sim \boldsymbol{\pi}(\cdot \mid t, x_n(t); w_n; \boldsymbol{\phi}_n)$.

By choosing two specific test functions $\mathcal{I}(t) = \frac{\partial}{\partial \boldsymbol{\theta}} J(t, x(t); w; \boldsymbol{\theta})$ and $\mathcal{H}(t) = 0$, the learnable parameters are then updated by the following rules:

$$\boldsymbol{\theta}_{n+1} \leftarrow \Pi_{K_{\theta,n}} \bigg(\boldsymbol{\theta}_n + a_n \int_0^T \frac{\partial J}{\partial \theta} (t, x_n(t); w_n; \boldsymbol{\theta}_n) \left[\mathrm{d}J\left(t, x_n(t); w_n; \boldsymbol{\theta}_n\right) + \gamma \log \boldsymbol{\pi}(u_n(t)|t, x_n(t); w_n; \boldsymbol{\phi}_n) \mathrm{d}t \right] \bigg), \tag{11}$$

$$\phi_{1,n+1} \leftarrow \Pi_{K_{1,n}} \Big(\phi_{1,n} - a_n Z_{1,n}(T) \Big), \tag{12}$$

$$\phi_{2,n+1} \leftarrow \Pi_{K_{2,n}} \Big(\phi_{2,n} + a_n Z_{2,n}(T) \Big), \tag{13}$$

$$w_{n+1} \leftarrow \Pi_{K_{w,n}} \Big(w_n - a_{w,n}(x_n(T) - z) \Big), \tag{14}$$

where

$$Z_{1,n}(t) = \int_0^t \bigg\{ \frac{\partial}{\partial \phi_1} \log \boldsymbol{\pi}\left(u_n(s) \mid s, x_n(s); w_n; \boldsymbol{\phi}_n\right) \left[\mathrm{d}J\left(s, x_n(s); w_n; \boldsymbol{\theta}_n\right) + \gamma \log \boldsymbol{\pi}(u_n(s)|s, x_n(s); w_n; \boldsymbol{\phi}_n) \mathrm{d}s \right] \bigg\}, \tag{15}$$

$$Z_{2,n}(t) = \int_0^t \bigg\{ \frac{\partial}{\partial \phi_2^{-1}} \log \boldsymbol{\pi}\left(u_n(s) \mid s, x_n(s); w_n, \boldsymbol{\phi}_n\right) \left[\mathrm{d}J\left(s, x_n(s); w_n; \boldsymbol{\theta}_n\right) + \gamma \log \boldsymbol{\pi}(u_n(s)|s, x_n(s); w_n; \boldsymbol{\phi}_n) \mathrm{d}s \right] \bigg\}, \tag{16}$$

and $\Pi_K(z) := \arg\min_{y \in K} |y - z|^2$ is the projection of a point z onto the subset K. The subsets involved in the above are:

$$K_{\theta,n} = \left\{(\theta_1,\theta_2) \in \mathbb{R}^2 \Big| |\theta_1| \leqslant c_{\theta_1}, |\theta_2| \leqslant c_{\theta_2}\right\},\ K_{1,n} = \left\{\phi_1 \in \mathbb{R}^d \Big| |\phi_1| \leqslant c_{1,n}\right\},$$
$$K_{2,n} = \left\{\phi_2 \in \mathbb{S}^d_{++} \Big| |\phi_2| \leqslant c_{2,n}, \phi_2 - \frac{1}{b_n} I \in \mathbb{S}^d_{++}\right\},\ K_{w,n} = \left\{w \in \mathbb{R} \Big| |w| \leqslant c_{w,n}\right\}.$$

In this procedure, the constants a_n, $a_{w,n}$, c_{θ_1}, c_{θ_2}, $c_{1,n}$, $c_{2,n}$, $c_{w,n}$ and b_n are hyperparameters that can be set according to Theorem 1 below. Note that the second equation in (7) represents the gradient with respect to $\boldsymbol{\phi}$ to *minimize* the variance, and hence each iteration should move in the opposite direction of the gradient. This is why there is a negative sign in (12) that updates ϕ_1. However, in (13), the increment $Z_{2,n}(T)$ is with respect to the gradient in ϕ_2^{-1}, which is *decreasing* in ϕ_2; so the sign in (13) is changed back from negative to positive.

The updating rules on $\boldsymbol{\phi}$ and w described above are (nonlinear) stochastic approximation algorithms (cf. Chau and Fu 2014). However, we need to adapt the general stochastic approximation theory to our case in order to avoid encountering extreme states and having unbounded errors. This is achieved by introducing certain projections onto bounded sets in the learning process, a technique pioneered by Andradóttir (1995). Note that these bounded sets do not require any prior knowledge about the market environment to specify, and they expand to span the whole space as the number of iterations grows. Therefore, our algorithm still remains model-free.

We now present the baseline algorithm as Algorithm 1, and put the analysis leading to it in Appendix B. Note that while the algorithm is derived from a continuous-time analysis, for the final computer implementation it needs to be discretized. The time discretization in Algorithm 1 plays a dual role: the discrete times indicate both the moments for rebalancing and applying the new portfolio, and they are also used to approximate the integrals involved in the solutions.

Algorithm 1 CTRL Baseline Algorithm

Initialize $\boldsymbol{\theta}, \boldsymbol{\phi}$ and w.
for `iter = 1 to N` **do**
  Initialize $k = 0$, time $t = t_k = 0$, wealth $x(t_k) = x_0$.
  Generate wealth trajectory by policy (9) following dynamics (1).
  Collect the whole trajectory $\{(t_k, x(t_k), u(t_k))\}_{k \geqslant 0}$.
  Update $\boldsymbol{\theta}$ using (30).
  Update $\boldsymbol{\phi}$ using (31) and (32).
  Update w by (14).
end for

The following theorem, whose proof (for a more general version of the theorem) is relegated to Appendix G.1, presents the convergence and convergence rates of the parameters updated according to Algorithm 1.

Theorem 1. *Assume that the stock prices follow a multi-dimensional geometric Brownian motion with constant return and volatility rates, and the risk-free rate is a constant. In Algorithm 1, let the parameters*

c_{θ_1}, c_{θ_2} be some positive constants, and a_n, $a_{w,n}$, $c_{1,n}$, $c_{2,n}$, $c_{w,n}$ and b_n be set as follows:

(i) $a_n = a_{w,n} = \dfrac{\alpha}{n+\beta}$, *for some constants $\alpha > 0$ and $\beta > 0$;*

(ii) $b_n = 1 \vee (\log\log n)^{\frac{1}{8}}, c_{1,n} = 1 \vee (\log\log n)^{\frac{1}{8}}, c_{2,n} = 1 \vee (\log\log n)^{\frac{1}{8}}, c_{w,n} = 1 \vee (\log\log n)^{\frac{1}{16}}$.

Then

(a) As $n \to \infty$, $\phi_{1,n}$ and $\phi_{2,n}$ almost surely converge to the true values $\phi_1^* = (\sigma\sigma^\top)^{-1}(\mu - r)$ and $\phi_2^* = \frac{\gamma}{2}(\sigma\sigma^\top)^{-1}$ respectively, and w_n almost surely converges to the true value $w^* = \frac{ze^{(\mu-r)^\top(\sigma\sigma^\top)^{-1}(\mu-r)T} - x_0}{e^{(\mu-r)^\top(\sigma\sigma^\top)^{-1}(\mu-r)T} - 1}$.

(b) For any n, $\mathbb{E}[|\phi_{1,n+1} - \phi_1^*|^2] \leqslant C\frac{(\log n)^p(\log\log n)}{n}$, where C and p are positive constants independent of n.

Note the convergence rate of $\phi_{1,n}$ is of the order $\frac{(\log n)^p(\log\log n)}{n}$, which nearly matches the typical optimal convergence rate of stochastic approximation algorithms (e.g. Broadie et al. 2011) and differs only by a factor $(\log n)^p(\log\log n)$ which is very small relative to n. Moreover, Assumptions (i)-(ii) in Theorem 1 are not necessary for the first statement about the almost sure convergence of $\phi_{1,n}$, $\phi_{2,n}$ and w_n. More general assumptions are presented in (63) and (72), in Appendix G.1.

4.2 Stochastic training and deterministic execution

We first present the following result.

Theorem 2. *Consider two policies $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$ in the same form as (9), given by $\boldsymbol{\pi} = \mathcal{N}\Big(-\phi_1(x-w), C(t)\Big)$, $\hat{\boldsymbol{\pi}} = \mathcal{N}\Big(-\phi_1(x-w), \hat{C}(t)\Big)$, where $C(\cdot), \hat{C}(\cdot) \in \mathbb{S}^d_{++}$ are two deterministic functions satisfying $C(t) - \hat{C}(t) \in \mathbb{S}^d_+$ for all $t \in [0,T]$, along with their respective wealth trajectories $\{x^{u^{\boldsymbol{\pi}}}(t) : 0 \leqslant t \leqslant T\}$ and $\{x^{u^{\hat{\boldsymbol{\pi}}}}(t) : 0 \leqslant t \leqslant T\}$. Then $\hat{\boldsymbol{\pi}}$ mean–variance dominates $\boldsymbol{\pi}$; i.e. $\mathbb{E}[x^{u^{\boldsymbol{\pi}}}(T)] = \mathbb{E}[x^{u^{\hat{\boldsymbol{\pi}}}}(T)]$ and $\mathrm{Var}\left(x^{u^{\boldsymbol{\pi}}}(T)\right) \geqslant \mathrm{Var}\left(x^{u^{\hat{\boldsymbol{\pi}}}}(T)\right)$.*

A proof of this theorem is delayed to Appendix G.2. The theorem indicates that, for the same entropy-regularized MV problem (with the same temperature parameter), even though the two policies with the same mean generate the same expected terminal wealth, the one with a lower level of exploration has a more stable result in terms of the variance of the terminal wealth. So more exploration is worse off from the MV perspective. In particular, we only need to use *deterministic* policies for *actual* execution of a portfolio (instead of using a random sampler from the learned optimal stochastic policy). Note that this feature is specific to the current setting of the problem (i.e. a frictionless market with a small investor), and is not necessarily true in general where actual actions also need to be sampled from stochastic policies in order to broaden search space and observe the responses to actions from the environment. As discussed at the end of Section 3.1, the dynamics (1) with a small investor dictate that the investor would know the consequence of executing any portfolio even if she does not actually execute it. This is in sharp contrast to, say, a bandit problem, in which an agent has no knowledge about counterfactuals. Thus, intuitively, in the MV problem one does not need to do exploration *per se* for the purpose of trial and error; she could do it *on paper*. However, again as explained before, stochastic policies are necessary for computing the policy gradient due to technical reasons; so we will use them for training (i.e. for updating the parameters). We do this by randomly generating portfolio processes from the current stochastic policy and simulating the corresponding (counterfactual) wealth trajectories based on (1).

Denote a deterministic policy for execution of the original (non-exploratory) wealth equation (1) by

$$\boldsymbol{u}(t, x; w; \boldsymbol{\phi}) = -\phi_1(x - w), \tag{17}$$

which is a degenerate stochastic policy with a Dirac distribution and coincides with the mean of the policy $\boldsymbol{\pi}^{\boldsymbol{\phi}}$ defined in (9). The Sharpe ratio of the terminal wealth of (1) under this policy is defined as

$$\frac{\mathbb{E}\left[x^{\boldsymbol{u}}(T)/x^{\boldsymbol{u}}(0)\right] - 1}{\sqrt{\operatorname{Var}(x^{\boldsymbol{u}}(T)/x^{\boldsymbol{u}}(0))}}, \tag{18}$$

which depends only on ϕ_1, and is denoted by $\mathrm{SR}(\phi_1)$.

Theorem 3. *Under the same setting of Theorem 1, we have*

$$\mathbb{E}\left[\sum_{n=1}^{N}(\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_{1,n}))\right] \leqslant C + C\sqrt{N(\log N)^p \log\log N}, \quad \forall N,$$

where $C > 0$ is a constant independent of N, and p is the same constants appearing in Theorem 1.

A proof of Theorem 3 is given in Appendix G.3. The result reveals the importance of the parameter ϕ_1. Indeed, the theoretical value of the vector $\phi_1^* = -(\sigma\sigma^\top)^{-1}(\mu - r)$ (see (27)) constitutes the proportions allocated to the risky assets and hence the composition of an MV efficient mutual fund. This composition in turn determines the Sharpe ratio of the resulting portfolio, noting that any MV efficient portfolio has the same Sharpe ratio. Theorem 3 stipulates that, in terms of the Sharpe ratio, the cumulative gap between the iterates of our algorithm and the "oracle" (i.e. the theoretically optimal portfolio should all the market parameters be known) up to the Nth iteration is of the order of $\sqrt{N(\log N)^p \log\log N}$. The sublinearity of this gap yields that in the long run, the algorithm performs almost optimally.

5 Empirical Performance and Comparisons

To assess the efficacy of our CTRL algorithm, we carry out an empirical study to juxtapose its performance with other well-established asset allocation strategies using standard/popular metrics such as Sharpe ratio and maximum drawdown.

The dataset employed in this study was sourced from the Wharton Research Data Services (WRDS). The asset universe includes the S&P 500 constituents that remained continuously listed from 1990 to 2020 and had daily trading data available in WRDS. From this set, we compile a pool of the first 300 stocks, organized alphabetically according to their tickers. These stocks' returns are calculated based on dividend-adjusted prices. For each experiment, we choose 10 stocks randomly from this pool and backtest them with various portfolio strategies from 2000 to 2020, while utilizing the period 1990-1999 for pretraining. We repeat these experiments 100 times to derive statistical properties.

Moreover, for algorithms requiring a specified target return, including our own and other MV based strategies, we set the target annualized return at 15%, corresponding to $z = 1.15$ in our model. This figure aligns with the approximate annualized return of the S&P 500 during the pretraining period 1990–2000. For static models with monthly rebalancing, the target annual return is translated into a monthly target return $\mu^* = (1.15)^{\frac{1}{12}} - 1 \approx 1.17\%$. For simplicity, we assume the risk-free interest rate (r) to be zero. The initial wealth is normalized to be \$1 for all the experiments. Furthermore, since the alternative asset allocation methods such as the buy-and-hold market index and sample-based minimum variance inherently require full investment in the risky assets (i.e. no risk-free allocation), we ensure comparability across all methods by mandating that all available funds are allocated to stocks only in our experiments.

We briefly summarize the methods and performance criteria as follows.

The proposed CTRL algorithm While Algorithm 1 has been proved to have good theoretical properties including sublinear convergence, modifications are needed for practical implementation. Here, we show four variants of our RL portfolio choices with history-dependent incremental updating:

- **vCTRL:** No pretraining, borrowing allowed, daily rebalancing.
- **pCTRL:** Pretrained, borrowing allowed, daily rebalancing.
- **c-mCTRL:** Pretrained, borrowing not allowed, monthly rebalancing.
- **c-dCTRL:** Pretrained, borrowing not allowed, daily rebalancing.

Alternative allocation strategies We compare our CTRL strategies with 15 existing alternative portfolio allocation strategies (or benchmarks) broadly studied/employed in theory/practice, including buy-and-hold market index (S&P500), equal weight (ew), sample-based single-period mean-variance (mv), sample-based minimum variance (min_v), James–Stein shrinkage estimator (js), Ledoit–Wolf shrinkage estimator (lw), Black–Litterman model (bl), Fama–French three factor model (ff), risk parity (rp), distributionally robust mean–variance (drmv), sample-based monthly-rebalancing continuous-time mean–variance (mctmv), sample-based daily-rebalancing continuous-time mean–variance (dctmv), predictive mean-variance (pmv), deep deterministic policy gradient (ddpg), and proximal policy optimization (ppo). The details of alternative strategies are described in Appendix C.

Performance criteria We use a comprehensive set of performance metrics including annualized return, volatility, Sharpe ratio, Sortino ratio, Calmar ratio, maximum drawdown (MDD), and recovery time (RT).

The exact definitions of these metrics are given in Appendix D.

We now present a detailed analysis of the backtesting results.

5.1 Average wealth trajectories

First, we compare the *average* wealth trajectories under the different strategies over 100 independent experiments, each (except the S&P 500 index) with 10 randomly selected S&P 500 constituents.[9] They are depicted in Figure 1. These trajectories average out outliers and provide "first impressions" of the respective strategies. The four strategies reaching the highest average final wealth are the four CTRL variants. In particular, vCTRL and pCTRL both significantly outperform the other methods in terms of terminal wealth. These two strategies are closest to the theoretically optimal continuous-time RL strategy over a long time horizon because they have no leverage constraint and are rebalanced daily. These two corresponding wealth trajectories are remarkably similar, with the pCTRL strategy exhibiting lower volatility and slightly higher returns compared to vCTRL. This suggests that pretraining can enhance algorithmic performance.

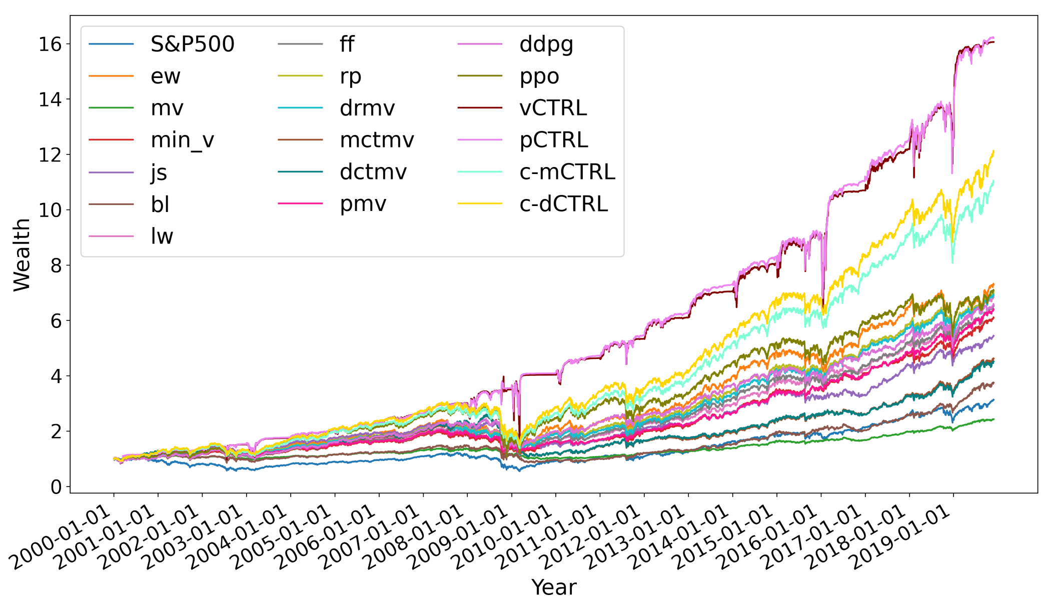

Figure 1: **Average wealth trajectories under 4 CTRL algorithms and 15 alternative methods over 100 independent experiments each with 10 randomly selected stocks (except the S&P 500 index) from 2000 to 2020.**

[9]In our experiments, whenever a sample wealth trajectory hits zero the remainder of the trajectory is set to be zero.

5.2 Comparative performance analysis

While Figure 1 offers a bird's-eye view of the performance comparison of various allocation methods, a more detailed evaluation, using the criteria outlined in Appendix D, is necessary for a comprehensive understanding. Table 1 reports those performance criteria, all averaged over 100 independent experiments, each with 10 randomly selected S&P 500 constituents (except for the S&P 500 index), for the period from 2000 to 2020.

First of all, the 4 CTRL strategies attain the highest average annualized returns, commensurate with Figure 1. However, in terms of the risk-adjusted return – the Sharpe ratio – c-dCTRL and c-mCTRL (the two strategies with the no-borrowing constraint) beat vCTRL and pCTRL (the two without leverage constraint) by big margins, due to the latter two's (understandably) substantial volatilities. Indeed, the former two have the greatest Sharpe ratios among all the strategies under comparison. Between these two, on the other hand, c-dCTRL (daily rebalanced) has a higher, if only slightly, Sharpe than c-mCTRL (monthly rebalanced). While this is expected in theory due to the underlying continuous-time setting, in practice the monthly rebalanced c-mCTRL may be preferred for saving transaction costs. Other high Sharpe strategies include ew, lw, rp, drmv, and pmv. The outperformance of the naïve equally-weighted strategy (ew) is documented in the literature (e.g. DeMiguel et al., 2009b). On the other hand, under the Sortino ratio and Calmar ratio, pCTRL and vCTRL rank higher. This is because both strategies exhibit very high returns, and the Sortino ratio, in particular, does not penalize upside volatility, allowing the unconstrained CTRL strategies to benefit.

Second, the maximum drawdowns (MDDs) of vCTRL and pCTRL are among the largest and considerably greater than the index. This is again due to the unconstrained leverage leading to unconstrained risk exposure. By contrast, c-dCTRL and c-mCTRL have comparable MDDs to the overall market. Remarkably, the 4 CTRL strategies have the shortest recovery times (RTs), averaging around 400 days, compared with the S&P 500 index of 869 days.

Last but probably most importantly, we observe the notably low or negative annualized returns and Sharpe ratios of mv, mctmv, and dctmv. These are all derived by the classical model-based, plug-in approach, the first a (rolling horizon) static model and the other two dynamic MV models. They all need to estimate the model parameters first before optimizing. The inherent difficulty in estimating those parameters (especially the mean) and the high sensitivity of the optimal solutions with respect to the estimations have caused poor performances, as discussed earlier. In particular, the dynamic models are worse than the static counterpart, resulting in even bankruptcy in some instances, due to the *cumulative* estimation errors in a dynamic environment. By contrast, the 4 CTRL strategies mitigate this problem by bypassing the model parameter estimation altogether, which is the fundamental reason for their outstanding performances.

	Return	Volatility	Sharpe	Sortino	Calmar	MDD	RT
S&P500	5.90% (0.00%)	0.19 (0.0)	0.311 (0.0)	0.494 (0.0)	0.107 (0.0)	0.552 (0.0)	869 (0)
ew	10.28% (0.16%)	0.211 (0.002)	0.496 (0.011)	0.807 (0.018)	0.188 (0.005)	0.565 (0.009)	547 (27)
mv	4.06% (0.24%)	0.149 (0.002)	0.29 (0.018)	0.466 (0.03)	0.114 (0.009)	0.438 (0.013)	1371 (70)
min_v	8.86% (0.28%)	0.187 (0.002)	0.488 (0.018)	0.79 (0.03)	0.186 (0.008)	0.513 (0.011)	870 (36)
js	6.40% (0.67%)	0.354 (0.026)	0.27 (0.024)	0.44 (0.039)	0.123 (0.012)	0.694 (0.022)	1435 (69)
bl	5.83% (0.38%)	0.293 (0.029)	0.285 (0.019)	0.46 (0.031)	0.12 (0.009)	0.634 (0.044)	1417 (67)
lw	9.54% (0.28%)	0.194 (0.002)	0.501 (0.016)	0.812 (0.027)	0.207 (0.008)	0.488 (0.009)	842 (42)
ff	9.43% (0.24%)	0.202 (0.002)	0.476 (0.014)	0.769 (0.023)	0.196 (0.007)	0.506 (0.009)	711 (38)
rp	10.02% (0.17%)	0.192 (0.002)	0.529 (0.012)	0.856 (0.02)	0.193 (0.006)	0.54 (0.009)	653 (24)
drmv	9.89% (0.19%)	0.189 (0.002)	0.532 (0.013)	0.86 (0.022)	0.193 (0.006)	0.534 (0.009)	705 (26)
mctmv	-2.22% (2.94%)	0.315 (0.017)	0.12 (0.056)	0.237 (0.076)	0.023 (0.032)	0.699 (0.017)	1505 (63)
dctmv	2.03% (2.13%)	0.276 (0.008)	0.185 (0.045)	0.32 (0.063)	0.063 (0.023)	0.687 (0.014)	1556 (62)
pmv	9.15% (0.29%)	0.182 (0.002)	0.511 (0.017)	0.832 (0.029)	0.199 (0.008)	0.487 (0.009)	887 (39)
ddpg	9.61% (0.52%)	0.423 (0.028)	0.297 (0.02)	0.503 (0.033)	0.153 (0.01)	0.714 (0.019)	1284 (75)
ppo	9.71% (0.52%)	0.457 (0.038)	0.344 (0.022)	0.57 (0.036)	0.152 (0.01)	0.77 (0.026)	1320 (90)
vCTRL	15.00% (0.06%)	0.551 (0.014)	0.275 (0.02)	0.636 (0.033)	0.215 (0.01)	0.701 (0.026)	402 (31)
pCTRL	15.01% (0.02%)	0.458 (0.013)	0.328 (0.021)	0.77 (0.028)	0.243 (0.009)	0.634 (0.021)	420 (38)
c-mCTRL	12.52% (0.19%)	0.22 (0.002)	0.567 (0.012)	0.905 (0.019)	0.209 (0.005)	0.581 (0.007)	409 (20)
c-dCTRL	13.05% (0.19%)	0.219 (0.003)	0.574 (0.012)	0.937 (0.02)	0.216 (0.006)	0.568 (0.01)	365 (16)

Table 1: **Comparison of out-of-sample performance of different allocation methods from 2000 to 2020.** We report return, volatility, Sharpe ratio, Sortino ratio, Calmar ratio, maximum drawdown (MDD) and recovery time (RT), all annualized, over 100 independent experiments each with 10 randomly selected stocks (except S&P 500 index). For each cell, the upper number is the average (over the 100 experiments) while the lower one with parentheses is the standard deviation.

5.3 Bull and bear markets

The previously reported results are drawn from a long period of 20 years consisting of a number of bull and bear market cycles. We now examine the performances over a bull period and a bear one respectively. It just so happened that the first period 2000-2010 was overall a bear market, during which there were the 2001 dot com bubble and the 2008 financial crisis, and S&P 500 had a *negative* annualized return of 0.9%. The second period 2010-2020, meanwhile, had a rarely seen long bull run during which S&P 500 returned an annual average of 13.1%. Tables 2 and 3 report the comparison results for these two periods respectively.

In the bear period 2000-2010, the 4 CTRL strategies now significantly outperform all the others including ew in terms of annualized return. Moreover, c-dCTRL and c-mCTRL achieve the highest Sharpe ratios of 0.407 and 0.381, respectively, surpassing substantially the next runner-up recorded at 0.319 by ew. The CTRL strategies also top the charts in both Sortino and Calmar ratios, even though they are also among the higher ones in MDD. As for RT, S&P 500 had never returned to the previous peak from the bottom before 2010, while our CTRL strategies render much shorter recovery periods, with three of them outperforming ew and all four notably surpassing the other methods.

During the 2010–2020 bull market, slightly more than half of the strategies, including the 4 CTRL strategies, outperformed the market return of 13.1%, with the CTRL strategies achieving the highest returns, exceeding 15%. In terms of Sharpe ratio, eight of the strategies outperform the market, and c-dCTRL and c-mCTRL along with min_v, drmv, and pmv are the top five.[10] Our CTRL strategies are mediocre in Calmar ratio due to the average MDDs, but they rank among the fastest in terms of RTs. Overall, during this long bull run, almost all the strategies are doing well and CTRLs are still among the best in terms of Sharpe ratio and annualized return. This close match indicates that it is harder to outperform in a bull market, and a strategy needs to outperform especially during bear periods in order to excel in the long run. It in turn calls for more robust performances, something CTRL can provide as evident from this empirical study.

One of the most important conclusions from this empirical study is that the model-free continuous-time RL strategies *decisively* outperform the classical mode-based, plug-in continuous-time counterparts (i.e. mctmv and dctmv) in all the metrics and regardless of the market conditions, corroborating the key benefit of bypassing model parameter estimation. The CTRL strategies are also consistently among the best in a host of widely studied and practiced portfolio strategies, especially during volatile and downturn periods. Of the CTRL variants, we recommend c-mCTRL, the pretrained, no-borrowing and monthly rebalanced strategy, for its good balance between performance, practicality and robustness.

[10] Among these five strategies with close Sharpe ratios, a pairwise comparison using the Wilcoxon rank test does not show statistical significance; see Appendix F.1.

	Return	Volatility	Sharpe	Sortino	Calmar	MDD	RT
S&P500	-0.90%	0.224	-0.041	-0.066	-0.017	0.552	N/A
	(0.00%)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(N/A)
ew	7.69%	0.244	0.319	0.524	0.142	0.565	547
	(0.26%)	(0.002)	(0.012)	(0.019)	(0.006)	(0.009)	(27)
mv	-0.17%	0.17	0.011	0.024	0.017	0.43	1337
	(0.36%)	(0.003)	(0.021)	(0.033)	(0.009)	(0.013)	(71)
min_v	4.07%	0.219	0.197	0.321	0.093	0.513	870
	(0.40%)	(0.003)	(0.019)	(0.031)	(0.009)	(0.011)	(36)
js	2.51%	0.366	0.118	0.195	0.064	0.631	1408
	(0.84%)	(0.023)	(0.022)	(0.035)	(0.013)	(0.018)	(71)
bl	-1.54%	0.298	-0.024	-0.03	-0.005	0.576	1477
	(0.60%)	(0.026)	(0.021)	(0.033)	(0.01)	(0.016)	(65)
lw	5.22%	0.228	0.236	0.384	0.117	0.487	842
	(0.36%)	(0.002)	(0.016)	(0.026)	(0.008)	(0.009)	(42)
ff	5.39%	0.239	0.232	0.378	0.114	0.505	711
	(0.33%)	(0.003)	(0.014)	(0.023)	(0.007)	(0.009)	(38)
rp	6.67%	0.226	0.301	0.49	0.131	0.54	653
	(0.27%)	(0.002)	(0.013)	(0.022)	(0.006)	(0.009)	(24)
drmv	6.18%	0.222	0.284	0.462	0.124	0.534	705
	(0.31%)	(0.002)	(0.015)	(0.024)	(0.007)	(0.009)	(26)
mctmv	-6.87%	0.349	-0.06	-0.054	-0.048	0.685	1505
	(2.82%)	(0.02)	(0.05)	(0.067)	(0.03)	(0.016)	(63)
dctmv	-2.89%	0.307	-0.011	0.009	-0.012	0.672	1556
	(2.06%)	(0.008)	(0.04)	(0.056)	(0.022)	(0.014)	(62)
pmv	4.52%	0.21	0.221	0.361	0.102	0.486	887
	(0.36%)	(0.002)	(0.017)	(0.028)	(0.008)	(0.009)	(39)
ddpg	8.43%	0.471	0.224	0.381	0.141	0.691	1290
	(0.72%)	(0.024)	(0.02)	(0.034)	(0.012)	(0.016)	(73)
ppo	8.80%	0.452	0.269	0.453	0.149	0.717	1295
	(0.86%)	(0.023)	(0.023)	(0.039)	(0.014)	(0.018)	(81)
vCTRL	14.99%	0.711	0.235	0.551	0.226	0.697	595
	(0.03%)	(0.019)	(0.017)	(0.03)	(0.011)	(0.02)	(33)
pCTRL	15.00%	0.601	0.28	0.666	0.26	0.632	423
	(0.02%)	(0.017)	(0.019)	(0.027)	(0.01)	(0.019)	(24)
c-mCTRL	9.92%	0.263	0.381	0.639	0.174	0.581	409
	(0.24%)	(0.005)	(0.012)	(0.019)	(0.006)	(0.007)	(20)
c-dCTRL	10.77%	0.261	0.407	0.658	0.183	0.568	365
	(0.25%)	(0.004)	(0.012)	(0.02)	(0.007)	(0.01)	(16)

Table 2: **Comparison of out-of-sample performance of different allocation methods from 2000 to 2010.** We report return, volatility, Sharpe ratio, Sortino ratio, Calmar ratio, maximum drawdown (MDD) and recovery time (RT), all annualized, over 100 independent experiments each with 10 randomly selected stocks (except S&P 500 index). For each cell, the upper number is the average (over the 100 experiments) while the lower one with parentheses is the standard deviation.

	Return	Volatility	Sharpe	Sortino	Calmar	MDD	RT
S&P500	13.10%	0.147	0.887	1.388	0.675	0.193	75
	(0.00%)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0)
ew	13.01%	0.171	0.78	1.263	0.55	0.253	210
	(0.28%)	(0.002)	(0.022)	(0.036)	(0.02)	(0.005)	(22)
mv	8.57%	0.122	0.713	1.187	0.502	0.198	332
	(0.32%)	(0.001)	(0.027)	(0.046)	(0.027)	(0.007)	(41)
min_v	13.96%	0.146	0.974	1.592	0.768	0.198	175
	(0.30%)	(0.002)	(0.025)	(0.042)	(0.028)	(0.006)	(22)
js	11.40%	0.27	0.565	0.939	0.431	0.471	804
	(0.91%)	(0.013)	(0.046)	(0.077)	(0.042)	(0.027)	(82)
bl	14.09%	0.185	0.78	1.29	0.563	0.277	380
	(0.45%)	(0.003)	(0.026)	(0.043)	(0.026)	(0.008)	(39)
lw	14.13%	0.152	0.945	1.555	0.764	0.204	215
	(0.40%)	(0.001)	(0.03)	(0.051)	(0.034)	(0.005)	(29)
ff	13.73%	0.156	0.897	1.456	0.66	0.223	215
	(0.35%)	(0.001)	(0.026)	(0.044)	(0.025)	(0.005)	(27)
rp	13.54%	0.151	0.913	1.476	0.683	0.211	181
	(0.25%)	(0.002)	(0.022)	(0.036)	(0.022)	(0.005)	(22)
drmv	13.81%	0.148	0.951	1.541	0.712	0.206	188
	(0.26%)	(0.002)	(0.022)	(0.038)	(0.023)	(0.004)	(24)
mctmv	11.91%	0.229	0.582	0.969	0.422	0.368	683
	(0.84%)	(0.008)	(0.034)	(0.057)	(0.029)	(0.016)	(58)
dctmv	11.89%	0.223	0.575	0.954	0.419	0.363	645
	(0.68%)	(0.006)	(0.033)	(0.056)	(0.029)	(0.015)	(60)
pmv	14.09%	0.147	0.965	1.594	0.762	0.203	275
	(0.39%)	(0.001)	(0.029)	(0.05)	(0.032)	(0.006)	(35)
ddpg	10.91%	0.265	0.516	0.872	0.38	0.438	731
	(0.84%)	(0.012)	(0.039)	(0.067)	(0.037)	(0.019)	(59)
ppo	11.95%	0.25	0.574	0.944	0.375	0.423	720
	(0.61%)	(0.011)	(0.032)	(0.053)	(0.024)	(0.02)	(65)
vCTRL	15.01%	0.245	0.671	1.184	0.466	0.386	33
	(0.10%)	(0.006)	(0.028)	(0.042)	(0.022)	(0.013)	(3)
pCTRL	15.02%	0.211	0.754	1.298	0.511	0.371	87
	(0.03%)	(0.005)	(0.03)	(0.046)	(0.025)	(0.012)	(28)
c-mCTRL	15.23%	0.163	0.946	1.534	0.691	0.229	159
	(0.27%)	(0.002)	(0.02)	(0.032)	(0.018)	(0.004)	(12)
c-dCTRL	15.42%	0.164	0.959	1.557	0.697	0.23	151
	(0.27%)	(0.002)	(0.02)	(0.033)	(0.019)	(0.004)	(11)

Table 3: **Comparison of out-of-sample performance of different allocation methods from 2010 to 2020.** We report return, volatility, Sharpe ratio, Sortino ratio, Calmar ratio, maximum drawdown (MDD) and recovery time (RT), all annualized, over 100 independent experiments each with 10 randomly selected stocks (except S&P 500 index). For each cell, the upper number is the average (over the 100 experiments) while the lower one with parentheses is the standard deviation.

6 Conclusions

This paper presents a general data-driven RL algorithm to solve continuous-time MV portfolio selection in markets described by observable Itô's diffusion processes without knowing their coefficients/parameters or attempting to estimate them. The general algorithm specializes to a more specific baseline algorithm for the Black–Scholes market environment, and we prove its theoretical performance guarantee including a sublinear regret, and then modify it into several variants for further performance enhancement and implementation practicality. Through a thorough comparative empirical study, we demonstrate the performance and robustness of the proposed CTRL strategies. This paper distinguishes itself from most existing works on applying RL to portfolio optimization in that its algorithms are based on a rigorous and explainable mathematical underpinning (relaxed control and martingality) established in Wang et al. (2020) and Jia and Zhou (2022a,b). Moreover, it is the first to derive a *model-free* sublinear regret bound for dynamic MV problems to our best knowledge.

One of the most notable insights derived from this work is the decisive outperformance of the explore-*and*-exploit approach of RL over the traditional estimate-*then*-plug-in counterparts in a dynamic market. This superiority is not because of the availability of "big data", as our baseline algorithm depends only on the stock price data (instead of thousands of factor data, which can be incorporated into our framework to further enhance the performance); rather it is due to a fundamentally different decision-making approach, namely, to learn the optimal policy without learning the model.

Despite the recent upsurge of interest in continuous-time RL, its study is still in its infancy, not to mention that its applications to financial decision-making are particularly a largely uncharted territory. In the MV setting, important open questions include performance guarantees of modified online algorithms, improvement of regret bound, off-policy learning, and large investors whose actions impact the asset prices (so counterfactuals become unobservable by mere "paper portfolios").

References

Andradóttir, S. (1995). A stochastic approximation algorithm with varying bounds. *Operations Research*, 43(6):1037–1048.

Bali, T. G., Engle, R. F., and Murray, S. (2016). *Empirical asset pricing: The cross section of stock returns.* John Wiley & Sons.

Ban, G.-Y., El Karoui, N., and Lim, A. E. (2018). Machine learning and portfolio optimization. *Management Science*, 64(3):1136–1154.

Bartram, S. M., Branke, J., De Rossi, G., and Motahari, M. (2021). Machine learning for active portfolio management. *The Journal of Financial Data Science*, 3(3):9–30.

Basak, S. and Chabakauri, G. (2010). Dynamic mean-variance asset allocation. *The Review of Financial Studies*, 23(8):2970–3016.

Best, M. J. and Grauer, R. R. (1991a). On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results. *The Review of Financial Studies*, 4(2):315–342.

Best, M. J. and Grauer, R. R. (1991b). Sensitivity analysis for mean-variance portfolio problems. *Management Science*, 37(8):980–989.

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2):1046–1089.

Black, F. and Litterman, R. (1990). Asset allocation: Combining investor views with market equilibrium. *Goldman Sachs Fixed Income Research*, 115.

Blanchet, J., Chen, L., and Zhou, X. Y. (2022). Distributionally robust mean-variance portfolio selection with Wasserstein distances. *Management Science*, 68(9):6382–6410.

Britten-Jones, M. (1999). The sampling error in estimates of mean-variance efficient portfolio weights. *Journal of Finance*, 54(2):655–671.

Broadie, M., Cicek, D., and Zeevi, A. (2011). General bounds and finite-time improvement for the Kiefer-Wolfowitz stochastic approximation algorithm. *Operations Research*, 59(5):1211–1224.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.

Chan, L. K., Karceski, J., and Lakonishok, J. (1999). On portfolio optimization: Forecasting covariances and choosing the risk model. *The Review of Financial Studies*, 12(5):937–974.

Chau, M. and Fu, M. C. (2014). An overview of stochastic approximation. *Handbook of Simulation Optimization*, pages 149–178.

Chen, L., Pelger, M., and Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2):714–750.

Chopra, V. K. and Ziemba, W. T. (2013). The effect of errors in means, variances, and covariances on optimal portfolio choice. In *Handbook of the Fundamentals of Financial Decision Making: Part I*, pages 365–373. World Scientific.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, 66(4):1047–1108.

Cvitanić, J., Lazrak, A., Martellini, L., and Zapatero, F. (2006). Dynamic portfolio choice with parameter uncertainty and the economic value of analysts' recommendations. *The Review of Financial Studies*, 19(4):1113–1156.

Dai, M., Jin, H., Kou, S., and Xu, Y. (2021). A dynamic mean-variance analysis for log returns. *Management Science*, 67(2):1093–1108.

DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009a). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.

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

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606.

Feng, G., Li, X., and Wang, Z. (2017). On the relation between several discrete choice models. *Operations Research*, 65(6):1516–1525.

Fleming, W. H. and Soner, H. M. (2006). *Controlled Markov Processes and Viscosity Solutions*, volume 25. Springer Science & Business Media.

Fudenberg, D., Iijima, R., and Strzalecki, T. (2015). Stochastic choice and revealed perturbed utility. *Econometrica*, 83(6):2371–2409.

Garlappi, L., Uppal, R., and Wang, T. (2007). Portfolio selection with parameter and model uncertainty: A multi-prior approach. *The Review of Financial Studies*, 20(1):41–81.

Gennotte, G. (1986). Optimal portfolio choice under incomplete information. *The Journal of Finance*, 41(3):733–746.

Goldfarb, D. and Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Gu, S., Kelly, B., and Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450.

Guijarro-Ordonez, J., Pelger, M., and Zanotti, G. (2021). Deep learning statistical arbitrage. *arXiv preprint arXiv:2106.04028.*

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR.

Hansen, L. P. and Singleton, K. J. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, pages 1269–1286.

He, X. D. and Zhou, X. Y. (2022). Who are I: Time inconsistency and intrapersonal conflict and reconciliation. In *Stochastic Analysis, Filtering, and Stochastic Optimization: A Commemorative Volume to Honor Mark HA Davis's Contributions*, pages 177–208. Springer.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343.

Hotz, V. J. and Miller, R. A. (1993). Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, 60(3):497–529.

Huang, Y., Jia, Y., and Zhou, X. Y. (2022). Achieving mean–variance efficiency by continuous-time reinforcement learning. In *Proceedings of the Third ACM International Conference on AI in Finance*, pages 377–385.

James, W. and Stein, C. (1992). Estimation with quadratic loss. In *Breakthroughs in Statistics: Foundations and Basic Theory*, pages 443–460. Springer.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance*, 45(3):881–898.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91.

Jia, Y. and Zhou, X. Y. (2022a). Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23(154):1–55.

Jia, Y. and Zhou, X. Y. (2022b). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(154):1–55.

Jorion, P. (1986). Bayes-Stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, pages 279–292.

Kim, J. H., Lee, Y., Kim, W. C., and Fabozzi, F. J. (2021). Mean–variance optimization for asset allocation. *The Journal of Portfolio Management*, 47(5):24–40.

Kim, T. S. and Omberg, E. (1996). Dynamic nonmyopic portfolio behavior. *The Review of Financial Studies*, 9(1):141–161.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Ledoit, O. and Wolf, M. (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *The Review of Financial Studies*, 30(12):4349–4388.

Lehmann, B. N. (1990). Fads, martingales, and market efficiency. *The Quarterly Journal of Economics*, 105(1):1–28.

Leippold, M., Wang, Q., and Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics*, 145(2):64–82.

Lettau, M. and Pelger, M. (2020). Factors that fit the time series and cross-section of stock returns. *The Review of Financial Studies*, 33(5):2274–2325.

Lewellen, J. et al. (2015). The cross-section of expected stock returns. *Critical Finance Review*, 4(1):1–44.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., et al. (2016). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Lim, A. E. and Zhou, X. Y. (2002). Mean-variance portfolio selection with random parameters in a complete market. *Mathematics of Operations Research*, 27(1):101–120.

Liu, J. (2007). Portfolio selection in stochastic environments. *The Review of Financial Studies*, 20(1):1–39.

Luenberger, D. G. (1998). *Investment Science*. Oxford University Press.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4):323–361.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science*, 9(2):277–293.

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

Snow, D. (2020). Machine learning in asset management—part 2: Portfolio construction—weight optimization. *The Journal of Financial Data Science*, 2(2):17–24.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Wachter, J. A. (2002). Portfolio and consumption decisions under mean-reverting returns: An exact solution for complete markets. *Journal of Financial and Quantitative Analysis*, 37(1):63–91.

Wang, H., Zariphopoulou, T., and Zhou, X. Y. (2020). Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34.

Wang, H. and Zhou, X. Y. (2020). Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 30(4):1273–1308.

Yong, J. and Zhou, X. Y. (1999). *Stochastic Controls: Hamiltonian Systems and HJB Equations*. New York, NY: Spinger.

Zhou, X. Y. (2023). Curse of optimality, and how to break it? In *Capponi, A., Lehalle, C.-A. (eds) Machine Learning and Data Sciences for Financial Markets*, pages 354–368. Cambridge University Press.

Zhou, X. Y. and Li, D. (2000). Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Applied Mathematics and Optimization*, 42(1):19–33.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA.

A Formulation and Solutions of Exploratory Mean–Variance Problem

A.1 Exploratory state dynamics under stochastic policies

We now present the precise formulation of the market environment, i.e., the asset price dynamics appearing in (1), as well as the exploratory wealth dynamics under stochastic policies.

Recall that $S^0(t)$ is the price of the risk-free asset, $S^i(t)$ the price of the i-th risky asset, $F(t)$ represents the values of the observable covariates/factors, and $u(t)$ is the portfolio choice vector, all at time t. We assume S^0 satisfies

$$\mathrm{d}S^0(t) = r(t, u(t), F(t))S^0(t)\mathrm{d}t, \tag{19}$$

and S^i follows

$$\mathrm{d}S^i(t) = S^i(t)\left[\mu^i(t, u(t), F(t))\mathrm{d}t + \sum_{j=1}^{m} \sigma^{ij}(t, u(t), F(t))\mathrm{d}W^j(t)\right], \quad i = 1, 2, \cdots, d, \tag{20}$$

where $r(t, u(t), F(t))$ is the short rate, $\mu(t, u(t), F(t)) := (\mu^1(t, u(t), F(t)), \mu^2(t, u(t), F(t)), \cdots,$ $\mu^d(t, u(t), F(t)))^\top \in \mathbb{R}^d$ and $\sigma(t, u(t), F(t)) := (\sigma^{ij}(t, u(t), F(t)))_{1\leqslant i\leqslant d, 1\leqslant j\leqslant m} \in \mathbb{R}^{d\times m}$ are respectively the instantaneous expectation and volatility of the risky asset returns at t, and W is an m-dimensional standard Brownian motion. We define $\Sigma(t, u(t), F(t)) := \sigma(t, u(t), F(t))\sigma(t, u(t), F(t))^\top \in \mathbb{R}^{d\times d}$ and assume it satisfies $\Sigma(t, u(t), F(t)) - \alpha I \in \mathbb{S}^d_+$ for all $t \geqslant 0$ with probability 1 for some constant $\alpha > 0$. We further assume that the above mentioned processes $\{r(t, u(t), F(t)), \mu(t, u(t), F(t)), \sigma(t, u(t), F(t)) : 0 \leqslant t \leqslant T\}$ are all well-defined and adapted in a given filtered probability space $\left(\Omega, \mathcal{F}, \mathbb{P}; (\mathcal{F}_t)_{t\geqslant 0}\right)$ satisfying the usual conditions. Moreover, the factor process F follows

$$\mathrm{d}F(t) = \iota(t, u(t), F(t))\mathrm{d}t + \nu(t, u(t), F(t))\mathrm{d}W(t). \tag{21}$$

All the coefficients in (19)–(21) depend on portfolio u to capture the most general scenario that a larger investor's actions may impact the values of assets and factors. They are independent of u when we consider a small investor.

The wealth equation (1) now specializes to

$$\mathrm{d}x^u(t) = (\mu(t, u(t), F(t)) - r(t, u(t), F(t))\boldsymbol{e}_d)^\top u(t)\mathrm{d}t + u(t)^\top \sigma(t, u(t), F(t))\mathrm{d}W(t), \quad 0 \leqslant t \leqslant T; \quad x_0^u = x_0. \tag{22}$$

Under a stochastic policy $\boldsymbol{\pi}$, its "dynamic of wealth" now describes the average of the (infinitely many) wealth processes under portfolios repeatedly sampled from $\boldsymbol{\pi}$; hence is different from (1). Applying the notion of relaxed stochastic control, Wang et al. (2020) derive the following "exploratory" dynamic:

$$
\begin{aligned}
\mathrm{d}x^{\boldsymbol{\pi}}(t) = & \int_{\mathbb{R}^d} [\mu(t,u,F(t)) - r(t,u,F(t))\boldsymbol{e}_d]^\top u\boldsymbol{\pi}(u|t,x^{\boldsymbol{\pi}}(t),F(t))\mathrm{d}u \\
& + \sqrt{\int_{\mathbb{R}^d} u^\top \Sigma(t,u,F(t)) u\boldsymbol{\pi}(u|t,x^{\boldsymbol{\pi}}(t),F(t))\mathrm{d}u}\mathrm{d}W(t).
\end{aligned}
\tag{23}
$$

Here, Brownian motion W may be different from the one in (20), although we use the same notation for simplicity. We emphasize that the averaged wealth process $x^{\boldsymbol{\pi}}$ is not observable (i.e. it is not part of the data) and (23) is used mainly for the *theoretical* analysis of the learning algorithms.

A.2 Exploratory mean–variance formulation and solutions in the Black–Scholes environment

We re-state the exploratory MV problem in a frictionless, multi-stock Black–Scholes market, without any factors F. The exploratory wealth equation is

$$
\mathrm{d}x^{\boldsymbol{\pi}}(t) = (\mu - r\boldsymbol{e}_d)^\top \int_{\mathbb{R}^d} u\boldsymbol{\pi}(u|t,x^{\boldsymbol{\pi}}(t))\mathrm{d}t + \sqrt{\int_{\mathbb{R}^d} u^\top \Sigma u\boldsymbol{\pi}(u|t,x^{\boldsymbol{\pi}}(t))\mathrm{d}u}\mathrm{d}W(t), \tag{24}
$$

while the goal is to find the stochastic policy $\boldsymbol{\pi}$ that minimizes the entropy regularized value function

$$
\mathbb{E}\left[(x^{\boldsymbol{\pi}}(T) - w)^2 + \gamma \int_0^T \int_{\mathbb{R}^d} \boldsymbol{\pi}(u|t,x^{\boldsymbol{\pi}}(t)) \log \boldsymbol{\pi}(u|t,x^{\boldsymbol{\pi}}(t))\mathrm{d}u\mathrm{d}t\right] - (w-z)^2. \tag{25}
$$

This problem has been solved by Wang and Zhou (2020) for the case of one stock, which can be extended readily to the multi-stock case. The optimal value function is

$$
\begin{aligned}
V^*(t,x;w) = & (x-w)^2 e^{-(\mu-r)^\top(\sigma\sigma^\top)^{-1}(\mu-r)(T-t)} \\
& + \frac{\gamma d}{4}(\mu-r)^\top(\sigma\sigma^\top)^{-1}(\mu-r)(T^2-t^2) \\
& - \frac{\gamma d}{2}\left((\mu-r)^\top(\sigma\sigma^\top)^{-1}(\mu-r)T - \frac{(T-t)}{d}\log\frac{\det(\sigma\sigma^\top)}{\pi\gamma}\right) - (w-z)^2, \\
& (t,x,w) \in [0,T]\times\mathbb{R}\times\mathbb{R},
\end{aligned}
\tag{26}
$$

the optimal policy is

$$
\begin{aligned}
\boldsymbol{\pi}^*(u \mid t,x,w) = \mathcal{N}\left(u \mid -(\sigma\sigma^\top)^{-1}(\mu-r)(x-w), (\sigma\sigma^\top)^{-1}\frac{\gamma}{2}e^{(\mu-r)^\top(\sigma\sigma^\top)^{-1}(\mu-r)(T-t)}\right) \\
(u,t,x,w) \in \mathbb{R}^d \times [0,T]\times\mathbb{R}\times\mathbb{R},
\end{aligned}
\tag{27}
$$

and the corresponding Lagrange multiplier is

$$w^* = \frac{ze^{(\mu-r)^\top(\sigma\sigma^\top)^{-1}(\mu-r)T} - x_0}{e^{(\mu-r)^\top(\sigma\sigma^\top)^{-1}(\mu-r)T} - 1}. \tag{28}$$

Once again, these analytical expressions are not used to compute the solutions (because the problem primitives are unknown); rather they are employed to *parameterize* the policies and value functions for learning.

B Details of baseline algorithm

To begin, consider the moment conditions in (7) under the absence of the factor $F(t)$ and with the stochastic policy $\boldsymbol{\pi}^{\boldsymbol{\phi}}$ parameterized as (9). The moment conditions can be reformulated as:

$$\begin{cases} \mathbb{E}\left[\int_0^T \frac{\partial J}{\partial \theta}(t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\theta}) \left[\mathrm{d}J(t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\theta}) + \gamma \hat{p}(t, \boldsymbol{\phi})\mathrm{d}t\right]\right] = 0, \\ \mathbb{E}\Bigg[\int_0^T \left[\frac{\partial}{\partial \boldsymbol{\phi}} \log \boldsymbol{\pi}(u(t)|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi})\right] \Big[\mathrm{d}J(t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\theta}) + \gamma \hat{p}(t, \boldsymbol{\phi})\mathrm{d}t\Big] + \gamma \frac{\partial \hat{p}}{\partial \boldsymbol{\phi}}(t, \boldsymbol{\phi})\Bigg] = 0, \\ \mathbb{E}\left[x^{\boldsymbol{\pi}}(T) - z\right] = 0, \end{cases} \tag{29}$$

where $\hat{p}(t, \boldsymbol{\phi})$ represents the the differential entropy of the policy $\boldsymbol{\pi}(\cdot \mid t, x; w; \boldsymbol{\phi})$, which can be explicitly calculated as

$$\hat{p}(t, \boldsymbol{\phi}) = -\frac{d}{2}\log(2\pi e) + \frac{1}{2}\log(\det \phi_2^{-1}) - \frac{d}{2}\phi_3(T-t),$$

revealing its independence of x, w and ϕ_1. To see the equivalence between conditions (7) and (29), we note that $\mathbb{E}[\log \boldsymbol{\pi}(u(t)|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi})] = \hat{p}(t, \boldsymbol{\phi})$, and hence,

$$\begin{aligned} \frac{\partial \hat{p}}{\partial \boldsymbol{\phi}}(t, \boldsymbol{\phi}) =& \mathbb{E}\left[\int_{\mathbb{R}^d} \log \boldsymbol{\pi}(u|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi}) \frac{\partial}{\partial \boldsymbol{\phi}} \boldsymbol{\pi}(u|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi})\mathrm{d}u\right] + \mathbb{E}\left[\int_{\mathbb{R}^d} \frac{\partial}{\partial \boldsymbol{\phi}} \boldsymbol{\pi}(u|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi})\mathrm{d}u\right] \\ =& \mathbb{E}\left[\int_{\mathbb{R}^d} \log \boldsymbol{\pi}(u|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi}) \frac{\partial}{\partial \boldsymbol{\phi}} \boldsymbol{\pi}(u|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi})\mathrm{d}u\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathbb{E}\bigg[\frac{\partial}{\partial \boldsymbol{\phi}} \log \boldsymbol{\pi}(u(t)|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi}) \log \boldsymbol{\pi}(u(t)|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi})\bigg] \\ =& \mathbb{E}\left[\int_{\mathbb{R}^d} \log \boldsymbol{\pi}(u|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi}) \frac{\partial}{\partial \boldsymbol{\phi}} \boldsymbol{\pi}(u|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi})\mathrm{d}u\right] \\ =& \frac{\partial \hat{p}}{\partial \boldsymbol{\phi}}(t, \boldsymbol{\phi}) = \mathbb{E}\bigg[\frac{\partial}{\partial \boldsymbol{\phi}} \log \boldsymbol{\pi}(u(t)|t, x^{\boldsymbol{\pi}}(t); w; \boldsymbol{\phi})\bigg]\hat{p}(t, \boldsymbol{\phi}) + \frac{\partial \hat{p}}{\partial \boldsymbol{\phi}}(t, \boldsymbol{\phi}). \end{aligned}$$

To design the baseline algorithm, we first calculate the relevant gradients given the parameterization in (8) and (9). Indeed, we have

$$\frac{\partial J}{\partial \theta_1}(t, x; w; \boldsymbol{\theta}) = t - T,\ \frac{\partial J}{\partial \theta_2}(t, x; w; \boldsymbol{\theta}) = t^2 - T^2,$$

$$\frac{\partial \hat{p}}{\partial \phi_1}(t,\boldsymbol{\phi}) = 0, \ \frac{\partial \hat{p}}{\partial \phi_2^{-1}}(t,\boldsymbol{\phi}) = \frac{\phi_2}{2},$$

and

$$\begin{aligned}
\frac{\partial \log \boldsymbol{\pi}(u \mid t,x;w;\boldsymbol{\phi})}{\partial \phi_1} &= -e^{-\phi_3(T-t)}\left[(x-w)\phi_2^{-1}u + (x-w)^2\phi_2^{-1}\phi_1\right], \\
\frac{\partial \log \boldsymbol{\pi}(u \mid t,x;w;\boldsymbol{\phi})}{\partial \phi_2^{-1}} &= \frac{1}{2}\phi_2 - \frac{1}{2}e^{-\phi_3(T-t)}(u+\phi_1(x-w))(u+\phi_1(x-w))^\top.
\end{aligned}$$

Recall the (theoretical) updating rules for $\boldsymbol{\theta}$, ϕ_1, ϕ_2 in (11)–(13) involve integrals. For actual implementation we use discretized summations to approximate those integrals: we discretize $[0,T]$ into small time intervals with an equal length of Δt. Then the updating rules are modified to

$$\boldsymbol{\theta} \leftarrow \Pi_{K_{\boldsymbol{\theta},n}}\left(\boldsymbol{\theta} + a_n \sum_{k=0}^{\lfloor \frac{T}{\Delta t}\rfloor - 1} \frac{\partial J}{\partial \boldsymbol{\theta}}(t_k, x_{t_k}; w; \boldsymbol{\theta})\left[J\left(t_{k+1}, x_{t_{k+1}}; w; \boldsymbol{\theta}\right) - J\left(t_k, x_{t_k}; w; \boldsymbol{\theta}\right) + \gamma \hat{p}(t_k, \boldsymbol{\phi})\Delta t\right]\right), \tag{30}$$

$$\begin{aligned}
\phi_1 \leftarrow \Pi_{K_{1,n}}\Bigg(\phi_1 - a_n \sum_{k=0}^{\lfloor \frac{T}{\Delta t}\rfloor - 1} \Bigg\{ &\frac{\partial \log \boldsymbol{\pi}}{\partial \phi_1}(u_{t_k} \mid t_k, x_{t_k}; w; \boldsymbol{\phi})\Big[J\left(t_{k+1}, x_{t_{k+1}}; w; \boldsymbol{\theta}\right) - J\left(t_k, x_{t_k}; w; \boldsymbol{\theta}\right) \\
&+ \gamma \hat{p}(t_k, \boldsymbol{\phi})\Delta t\Big] + \gamma \frac{\partial \hat{p}}{\partial \phi_1}(t_k, \boldsymbol{\phi})\Delta t\Bigg\}\Bigg),
\end{aligned} \tag{31}$$

$$\begin{aligned}
\phi_2 \leftarrow \Pi_{K_{2,n}}\Bigg(\phi_2 + a_n \sum_{k=0}^{\lfloor \frac{T}{\Delta t}\rfloor - 1} \Bigg\{ &\frac{\partial \log \boldsymbol{\pi}}{\partial \phi_2^{-1}}(u_{t_k} \mid t_k, x_{t_k}; w; \boldsymbol{\phi})\Big[J\left(t_{k+1}, x_{t_{k+1}}; w; \boldsymbol{\theta}\right) - J\left(t_k, x_{t_k}; w; \boldsymbol{\theta}\right) \\
&+ \gamma \hat{p}(t_k, \boldsymbol{\phi})\Delta t\Big] + \gamma \frac{\partial \hat{p}}{\partial \phi_2^{-1}}(t_k, \boldsymbol{\phi})\Delta t\Bigg\}\Bigg).
\end{aligned} \tag{32}$$

For each iteration, the algorithm starts with time 0 and initial wealth x_0. At each discretized timestep t, $t = 0, \Delta t, 2\Delta t, ..., \lfloor \frac{T}{\Delta t} \rfloor - 1$, it samples an action $u(t)$ from the Gaussian policy in (9), and calculates the new wealth at next timestep based on the current wealth, action and the assets price movement. At the final timestep $\lfloor \frac{T}{\Delta t} \rfloor$, the algorithm then uses the whole wealth trajectory to update parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

C Alternative Asset Allocation Methods

In this section, we briefly describe the other portfolio selection methods to be compared with our CTRL strategies. They include computation-free approaches, risk-based strategies, other RL methods and, predominantly, those based on both static and dynamic MV frameworks with different statistical techniques to estimate the mean and covariance matrix of asset returns. To dynamically implement all the static models, we rebalance monthly with the following month for single-period optimization, taking a rolling window of the immediately prior 10 years for estimating the model parameters.

For all the methods involved, we define $R(t) = (R^1(t), \cdots, R^d(t))^\top$ to be the vector of monthly excess returns of the d assets in the t-th month, and $w(t) = (w^1(t), \cdots, w^d(t))$ the portfolio in the t-th month, where $w^i(t)$ is the fraction of total wealth allocated to the i-th asset at t, $1 \leqslant t \leqslant 240$. The various portfolio choice methods are used to determine these weights.

For a fair comparison, we add the constraint that only investment in the risky assets is allowed in all the allocation methods. That is, the risky weights sum up to be 1: $w(t)^\top \boldsymbol{e}_d = 1$.

C.1 Buy-and-hold market index (S&P 500)

The S&P 500 index is capitalization weighted with dynamically adjusted constituents (`https://www.spglobal.com/spdji/en/index-family/equity/us-equity/us-market-cap/#overview`). It serves as a natural barometer of the overall market performance, a proxy of the market portfolio, and a benchmark many funds compare against. The buy-and-hold strategy of the S&P 500 index does not require any computation – its return over any given period is calculated based on the index's values on the first and last days of the period.

C.2 Equally weighted allocation (ew)

Another straightforward allocation method is the equally weighted allocation where $w^i(t) = \frac{1}{d}$ for $1 \leqslant i \leqslant d$. This strategy does not depend on any data, nor does it require any statistical estimation. Despite its simplicity and disregard of information, DeMiguel et al. (2009b) find that it exhibits admirable performance and remarkable robustness. Indeed, none of the 14 alternative allocation methods they tested consistently outperformed the equally weighted portfolio on real market data. As such, we take it as another important benchmark for comparison in our study.

C.3 Sample-based (single-period) mean–variance (mv)

Many portfolio selection methods are based on the one-period MV problem (Markowitz, 1952):

$$\begin{array}{ll} \min\limits_{w} & w^\top \Sigma w \\ \text{subject to} & w^\top \mu \geqslant \mu^*, \ w^\top \boldsymbol{e}_d = 1, \end{array} \tag{33}$$

where μ and Σ are the mean vector and covariance matrix of asset excess returns respectively, and μ^* is the investor's target expected return. The budget constraint $w^\top \boldsymbol{e}_d = 1$ ensures that the agent invests only in the risky assets. The solution to this problem can be found explicitly as

$$w^* = \frac{(\boldsymbol{e}_d^\top \Sigma^{-1} \boldsymbol{e}_d)\mu^* - \mu^\top \Sigma^{-1} \boldsymbol{e}_d}{(\mu^\top \Sigma^{-1} \mu)(\boldsymbol{e}_d^\top \Sigma^{-1} \boldsymbol{e}_d) - (\mu^\top \Sigma^{-1} \boldsymbol{e}_d)^2} \Sigma^{-1} \mu + \frac{-(\mu^\top \Sigma^{-1} \boldsymbol{e}_d)\mu^* + \mu^\top \Sigma^{-1} \mu}{(\mu^\top \Sigma^{-1} \mu)(\boldsymbol{e}_d^\top \Sigma^{-1} \boldsymbol{e}_d) - (\mu^\top \Sigma^{-1} \boldsymbol{e}_d)^2} \Sigma^{-1} \boldsymbol{e}_d. \tag{34}$$

Various plug-in methods are differentiated by the ways to estimate the unknown mean and covariance. Among them, the sample-based method estimates them using sample mean and covariance based on the most recent 120-month data:

$$\hat{\mu}(t) \equiv (\hat{\mu}_1(t), \cdots, \hat{\mu}_d(t))^\top = \frac{1}{M}\sum_{\tau=1}^{M} R_{t-\tau}, \quad \hat{\Sigma}(t) \equiv (\hat{\Sigma}_{ij}(t))_{d\times d} = \frac{1}{M-1}\sum_{\tau=1}^{M}(R_{t-\tau} - \hat{\mu})(R_{t-\tau} - \hat{\mu})^\top, \tag{35}$$

and then plugs them into the formula (34) to compute the portfolio weights.

C.4 Sample-based minimum variance (min_v)

A minimum variance portfolio achieves the lowest variance with a set of risky assets, without setting any expected return target. Mathematically, the minimum variance portfolio is obtained by solving

$$\begin{aligned} \min_{w} \quad & w^\top \Sigma w \\ \text{subject to} \quad & w^\top \boldsymbol{e}_d = 1. \end{aligned}$$

The solution is

$$w^* = \frac{1}{\boldsymbol{e}_d^\top \Sigma^{-1} \boldsymbol{e}_d} \Sigma^{-1} \boldsymbol{e}_d. \tag{36}$$

An advantage of the minimum variance portfolio is that it does not involve the mean returns of the stocks, which are significantly harder to estimate to a workable accuracy compared with the covariances. In our experiments we plug the sample covariance in (35) into (36) to obtain the minimum variance portfolio.

C.5 James–Stein shrinkage estimator for mean (js)

Jorion (1986) proposes a James–Stein type of shrinkage estimator (James and Stein, 1992) to shrink the estimates for the mean returns towards those of the sample-based minimum variance portfolio:

$$\tilde{\mu}(t) = \frac{\hat{\mu}(t)^\top \hat{\Sigma}(t)^{-1} \boldsymbol{e}_d}{\boldsymbol{e}_d^\top \hat{\Sigma}(t)^{-1} \boldsymbol{e}_d} \boldsymbol{e}_d,$$

where $\hat{\mu}, \hat{\Sigma}$ are the sample estimators given by (35). The James–Stein shrinkage estimator for the mean is then

$$\hat{\mu}^{\text{js}}(t) = (1 - \alpha(t))\hat{\mu}(t) + \alpha(t)\tilde{\mu}(t),$$

where $\alpha(t) = \frac{d+2}{d+2+(M-d-2)(\hat{\mu}(t)-\tilde{\mu}(t))^\top \hat{\Sigma}(t)^{-1}(\hat{\mu}(t)-\tilde{\mu}(t))}$, to be plugged into the solution of the MV problem, (34).

C.6 Ledoit–Wolf shrinkage estimator for covariance matrix (lw)

Ledoit and Wolf (2003) propose a shrinkage estimator for the covariance matrix. It starts with the single-index model for stock returns (Sharpe, 1963) at the t-th month:

$$R^i(t) = a_i + b_i R^m(t) + \varepsilon^i(t), \quad i = 1, 2, \cdots, d,$$

where $R^m(t)$ is the excess return of the market and $\varepsilon^i(t)$, $i = 1, 2, \cdots, d$, are the residuals that are uncorrelated to the market and to one another. Then the sample estimator of the covariance matrix of this model is:

$$\hat{F}(t) \equiv (\hat{F}_{ij}(t))_{d\times d} = bb^\top \hat{\sigma}_m^2(t) + \hat{D}(t)$$

where $\hat{\sigma}_m^2(t)$ is the sample variance of the market return, b is the vector of the slopes b_i and $\hat{D}(t)$ is the diagonal matrix containing the residual sample variances estimates. Denote by $\hat{\mu}_m(t)$ the sample mean of the market, and by $\hat{\sigma}_{im}(t)$ be the sample covariance between stock i and the market, both at time t.

Set $k_{ij}(t) = \frac{p_{ij} - r_{ij}}{c_{ij}(t)}$ to be the shrinkage estimator, where

$$p_{ij} = \frac{1}{M}\sum_{t=1}^{M}\left\{\left(R^i(t) - \hat{\mu}_i(t)\right)\left(R^j(t) - \hat{\mu}_j(t)\right) - \hat{\Sigma}_{ij}(t)\right\}^2, \;\; c_{ij}(t) = \left(\hat{F}_{ij}(t) - \hat{\Sigma}_{ij}(t)\right)^2,$$

and $r_{ii} = p_{ii}$ while $r_{ij} = \frac{\sum_{t=1}^{M} r_{ij}(t)}{M}$ for $i \neq j$ where

$$\begin{aligned} r_{ij}(t) =& \frac{\hat{\sigma}_{jm}(t)\hat{\sigma}_m(t)\left(R^i(t) - \hat{\mu}_i(t)\right) + \hat{\sigma}_{im}(t)\hat{\sigma}_m(t)\left(R^j(t) - \hat{\mu}_j(t)\right) - \hat{\sigma}_{im}(t)\hat{\sigma}_{jm}(t)\left(R^m(t) - \hat{\mu}_m(t)\right)}{\hat{\sigma}_m^2(t)} \\ & \times \left(R^m(t) - \hat{\mu}_m(t)\right)\left(R^i(t) - \hat{\mu}_i(t)\right)\left(R^j(t) - \hat{\mu}_j(t)\right) - \hat{F}_{ij}\hat{\Sigma}_{ij}(t). \end{aligned}$$

Then the Ledoit–Wolf shrinkage estimator for the covariance matrix is $\hat{\Sigma}^{lw}(t) \equiv (\hat{\Sigma}^{lw}_{ij}(t))$ where

$$\hat{\Sigma}^{lw}_{ij}(t) = \frac{k_{ij}(t)}{M}\hat{F}_{ij}(t) + \left(1 - \frac{k_{ij}(t)}{M}\right)\hat{\Sigma}_{ij}(t). \tag{37}$$

This estimator, along with any estimated mean returns, is to be plugged into the solution of the MV problem, (34).

C.7 Black–Litterman model (bl)

Premised upon the CAPM (Sharpe, 1964), Black and Litterman (1990) propose to use the market portfolio to infer mean returns of individual stocks. More precisely, at time t, we take the sample covariance matrix $\hat{\Sigma}^{all}(t)$ of all the 300 stocks in our stock universe, and compute the corresponding market portfolio (of these 300 stocks) $w^{all}(t)$ based on their market capitalizations. Then the implied stock mean return vector $\mu^{all}(t)$ and the market portfolio have the relation:

$$\mu^{all}(t) = \gamma(t)\hat{\Sigma}^{all}(t)w^{all}(t)$$

for some risk-adjusted coefficient $\gamma(t)$. This parameter is estimated using $\hat{\gamma}(t) = \frac{\hat{\mu}_m(t)}{\hat{\sigma}_m^2(t)}$, where $\hat{\mu}_m(t)$ and $\hat{\sigma}_m^2(t)$ are the sample mean and variance of the S&P 500 index respectively at t.

Then we extract the corresponding entries in $\mu^{all}(t)$ as our estimated expected returns for the d selected stocks, denoted by $\hat{\mu}^{bl}(t) \in \mathbb{R}^d$, and feed them along with any estimate of the covariance matrix into the solution (34).

C.8 Fama–French three factor model (ff)

The celebrated Fame–French three factor model (Fama and French, 1993) provides a decomposition for asset returns in the following form:

$$R(t) = \boldsymbol{\alpha} + \boldsymbol{B}F(t) + \boldsymbol{\epsilon}(t),$$

where $F(t) \in \mathbb{R}^3$ is a vector of mean-zero factors ("MKT", "SMB", and "HML"; see `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`) and $\boldsymbol{\epsilon}(t)$ consists of i.i.d. idiosyncratic noise terms for the stocks. Then the model parameters can be estimated by running linear regression on each individual stock against the centered factor values. Specifically, we first centralize the factors by

$$\tilde{F}(s) = \left(MKT(s) - \frac{1}{M}\sum_{\tau=1}^{M} MKT(t-\tau), SMB(s) - \frac{1}{M}\sum_{\tau=1}^{M} SMB(t-\tau), HML(s) - \frac{1}{M}\sum_{\tau=1}^{M} HML(t-\tau) \right)^\top,$$

for $s = t-M, \cdots, t-1$, where $MKT(s), SMB(s), HML(s)$ are the observed factor values at time s. Then we use the least square to estimate the linear regression:

$$R^i(s) = \boldsymbol{\alpha}^i + \boldsymbol{B}[i,]\tilde{F}(s) + \boldsymbol{\epsilon}^i(s),$$

for each individual asset i, where $\boldsymbol{B}[i,]$ stands for the i-th row of the matrix $\boldsymbol{B}$.

This procedure produces estimates $\hat{\boldsymbol{\alpha}}^i, \hat{\boldsymbol{B}}[i,]$, and the residual $\hat{\boldsymbol{\epsilon}}^i(s)$ for each $1 \leqslant i \leqslant d$ and each time instant $t-M \leqslant s \leqslant t-1$. The first two items lead to the estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{B}}$. Moreover, we obtain the sample covariance matrix of the factors by $\hat{\Sigma}_F(t) = \frac{1}{M-1}\sum_{\tau=1}^{M} \tilde{F}(t-\tau)\tilde{F}(t-\tau)^\top$, as well as a diagonal residual matrix $\hat{\Sigma}_\epsilon(t) = diag\{\sum_{\tau=1}^{M} \hat{\boldsymbol{\epsilon}}^1(t-\tau)^2, \cdots, \sum_{\tau=1}^{M} \hat{\boldsymbol{\epsilon}}^d(t-\tau)^2\}$. Finally, we set

$$\hat{\mu}^{\text{ff}}(t) = \hat{\boldsymbol{\alpha}}, \ \hat{\boldsymbol{\Sigma}}^{\text{ff}}(t) = \hat{\boldsymbol{B}}\hat{\Sigma}_F(t)\hat{\boldsymbol{B}}^\top + \hat{\Sigma}_\epsilon(t)$$

to be plugged into the solution (34).

C.9 Risk parity (rp)

Risk parity is a volatility based portfolio allocation strategy that equalizes risk contribution of individual stocks to the whole portfolio. Mathematically, the volatility (standard deviation) of a portfolio

$w = (w_1, \cdots, w_d)^\top$ is

$$C(w) = \sqrt{w^\top \Sigma w} = \sum_{i=1}^{d} C_i(w)$$

where

$$C_i(w) = w_i \frac{\partial C(w)}{\partial w_i} = \frac{w_i (\Sigma w)_i}{\sqrt{w^\top \Sigma w}}$$

is the risk contribution of the asset i. A risk parity portfolio w requires $C_i(w) = \frac{C(w)}{d}$, which can in turn be determined by the following system of equations:

$$w_i = \frac{C(w)^2}{(\Sigma w)_i d}, \quad i = 1, \cdots, d.$$

Alternatively, it can be derived by solving the following optimization problem

$$\begin{array}{ll} \min\limits_{w} & \sum_{i=1}^{d} \left[w_i - \frac{C(w)^2}{(\Sigma w)_i d} \right]^2 \\ \text{subject to} & w^\top \boldsymbol{e}_d = 1. \end{array}$$

C.10 Distributionally robust mean–variance (drmv)

Blanchet et al. (2022) develop a distributionally robust approach to address the model uncertainty issue for (static) MV allocation, with a data-driven technique to determine the size of the uncertainty set endogenously. They formulate the distributional robust version of (33) as

$$\begin{array}{ll} \min\limits_{w} \max\limits_{\mathbb{P} \in U_\delta(\hat{\mathbb{P}})} & \mathrm{Var}_{\mathbb{P}}(w^\top \boldsymbol{R}) \\ \text{subject to} & \min\limits_{\mathbb{P} \in U_\delta(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{P}}[w^\top \boldsymbol{R}] \geqslant \mu^*, \ w^\top \boldsymbol{e}_d = 1, \end{array}$$

where $U_\delta(\hat{\mathbb{P}})$ is a ball in the space of probability measures centered at the empirical probability $\hat{\mathbb{P}}$ with the radius δ in Wasserstein distance of L^q norm and order 2, where $1 \leqslant q \leqslant \infty$.

Blanchet et al. (2022) prove that this problem is equivalent to a non-robust, regularized convex optimization problem:

$$\begin{array}{ll} \min\limits_{w} & \sqrt{w^\top \hat{\Sigma} w} + \sqrt{\delta} ||w||_p \\ \text{subject to} & \hat{\mu}^\top w - \sqrt{\delta} ||w||_p \geqslant \mu^*, \ w^\top \boldsymbol{e}_d = 1, \end{array}$$

where $\frac{1}{p} + \frac{1}{q} = 1$, $||w||_p$ is the L^p norm of w, and $(\hat{\mu}, \hat{\Sigma})$ are the sample mean and covariance matrix of the stocks. In our implementation, we take $p = q = 2$, and use the immediate past 10 years of data to obtain the sample mean and covariance. The choice of δ is determined by the menu detailed in Blanchet et al. (2022).

C.11 Sample-based continuous-time mean–variance (ctmv)

All the methods described so far in this section are for static models, while implemented dynamically on a rolling horizon basis. Our CTRL algorithms are inherently for dynamic optimization; so we also include the classical model-based continuous-time MV method (Zhou and Li, 2000) for comparison purpose. As with all the plug-in approaches, this method does not explore nor update policy parameters. Instead, it estimates the mean vector $\mu(t)$ and covariance matrix $\sigma(t)$ using the 10-year historical stock data immediately prior to t, and then plug into the following formula for optimal policy (Zhou and Li, 2000):

$$\boldsymbol{u}^*(t,x;w^*) = -\Sigma(t)^{-1}\left(\mu(t)-r(t)\right)(x-w^*) \tag{38}$$

where

$$w^* = \frac{ze^{(\mu(t)-r(t))^\top\Sigma(t)^{-1}(\mu(t)-r(t))T} - x_0}{e^{(\mu(t)-r(t))^\top\Sigma(t)^{-1}(\mu(t)-r(t))T} - 1}. \tag{39}$$

In our experiments, we choose $T = 1$ (year). To fully compare this method against ours, we experiment with its two versions: the monthly rebalancing scheme (mctmv) and the daily rebalancing one (dctmv). All the model coefficients are re-estimated at a rebalaning time point using the most recent 10 years of stock price data. As the model in Zhou and Li (2000) allows risk-free allocation, to ensure a fair comparison with other methods, we normalize the portfolio (38) to a full risky allocation as $\hat{u}^*(t,x;w^*) := \frac{u^*(t,x;w^*)}{\sum_{i=1}^d u^{*i}(t,x;w^*)}x(t)$.

C.12 Predictive mean–variance (pmv)

None of the classical methods described above employs any predictive models for expected returns. However, the empirical asset pricing literature highlights the predictive power of certain factors in forecasting stock returns (Lewellen et al., 2015). Among the hundreds of factors documented in the literature (Cochrane, 2011), we focus on two of the most prominent ones as advocated by Bali et al. (2016): the short-term reversal factor (Jegadeesh, 1990; Lehmann, 1990) and the medium-term momentum factor (Jegadeesh and Titman, 1993).

The short-term reversal factor is among the strongest and most straightforward in empirical asset pricing (Bali et al., 2016). It is based on the empirical observation that top performers in a given month tend to underperform in the following month, while underperforming stocks often rebound. Tthis reversal factor is defined as, simply,

$$F^i_{rev}(t) = R^i(t),$$

for stock i in month t.

The medium-term momentum factor is based on investors' often delayed responses and overreactions to information. The momentum of a stock i in month t is defined as the return of the stock during the 11-month period from months $t-11$ to $t-1$:

$$F^i_{mom}(t) = \prod_{s=t-11}^{t-1} (1 + R^i(s)) - 1.$$

We then employ a predictive linear regression model to estimate expected stock returns:

$$R^i(t+1) = \alpha + \beta^i_{rev} F^i_{rev}(t) + \beta^i_{mom} F^i_{mom}(t) + \epsilon^i(t+1),$$

for $i = 1, 2, \ldots, d$. The coefficients α, β^i_{rev}, and β^i_{mom} are estimated using the method of least squares over a 10-year rolling window of historical data, resulting in parameter estimates $\hat{\alpha}$, $\hat{\beta}_{rev}$, and $\hat{\beta}_{mom}$.

To enhance accuracy of predictions and ensure alignment with established economic theory and empirical evidence, we impose constraints on the estimated coefficients based on their anticipated economic behaviors, as suggested by Campbell and Thompson (2008). Specifically, the momentum coefficient β^i_{mom} is anticipated to be positive, indicating a persistence in returns, while the short-term reversal coefficient β^i_{rev} is expected to be negative, capturing the mean-reversion. Therefore, we modify $\tilde{\beta}_{mom} := \max\{\hat{\beta}_{mom}, 0\}$ and $\tilde{\beta}_{rev} := \min\{\hat{\beta}_{rev}, 0\}$. This adjustment results in the final predictive model:

$$R^i(t+1) = \hat{\alpha} + \tilde{\beta}^i_{rev} F^i_{rev}(t) + \tilde{\beta}^i_{mom} F^i_{mom}(t). \tag{40}$$

Finally, the predicted returns obtained from (40) are plugged into the MV solution in (34).

C.13 Two existing reinforcement learning algorithms

In this subsection, we introduce two existing state-of-the-art RL algorithms: deep deterministic policy gradient and proximal policy optimization which we will use to compare with our CTRL algorithm.

Deep deterministic policy gradient (DDPG) DDPG is a cutting-edge actor–critic algorithm designed for problems with continuous action spaces. It integrates the benefits of both DPG (deterministic policy gradient) and DQN (deep Q-network); see e.g. Lillicrap et al. (2016); Mnih et al. (2015). DDPG has the following important features:

- **Architecture.** It employs two separate networks: the actor network that proposes an action given the current state, and the critic network that evaluates the proposed action by estimating the value function. The two networks are trained simultaneously.
- **Exploration strategy.** It carries out exploration by adding noises to action output, often using Ornstein–Uhlenbeck processes, to facilitate efficient exploration of the action space.
- **Experience replay.** It utilizes experience replay, where a replay buffer stores past states, actions, and rewards. This technique improves sample efficiency and breaks correlations between consecutive

learning steps.

- **Advantages for financial applications.** DDPG's ability to handle high-dimensional and continuous action spaces makes it particularly well-suited for financial applications including dynamic portfolio choice.

Proximal policy optimization (PPO) PPO has emerged as a popular choice in RL for its balance between performance and ease of implementation. It modifies traditional policy gradient approaches for improved stability and efficiency (Schulman et al., 2017). PPO has the following important features:

- **Objective function.** It introduces a clipped objective function that limits the size of policy updates. This approach reduces the likelihood of destructive large policy updates, ensuring more stable training.
- **Policy update.** It uses a policy update rule that keeps the new policy not too far away from the old one (hence the term "proximal").
- **Advantage estimation.** It often employs generalized advantage estimation (GAE) for calculating the advantage function, which helps reduce the variance of policy gradient estimates while retaining a bias.
- **Advantages for financial applications.** PPO's robustness and adaptability to various environments make it suitable for modeling complex financial systems, including optimal portfolio strategies over a range of market conditions.

D Details of Performance Metrics

Annualized return, volatility, and Sharpe ratio We use r_p to denote the return of the constructed portfolio. The annualized mean return rate $\mu_p = \mathbb{E}[r_p]$ and annualized volatility (standard deviation)

$$\sigma_p = \sqrt{\mathbb{E}[(r_p - \mu_p)^2]}$$

are two fundamental measures of portfolio performance. The (annualized) Sharpe ratio is defined as

$$\text{Sharpe Ratio} = \frac{\mu_p - r}{\sigma_p},$$

which is a widely-used risk-adjusted return measure.

Sortino ratio The Sharpe ratio equally penalizes upside and downside volatilities, while investors often take upside volatility positively. The Sortino ratio addresses this by focusing on downside risk. It is defined as:

$$\text{Sortino Ratio} = \frac{\mu_p - r}{\sigma_{downside}},$$

where $\sigma_{\text{downside}} = \sqrt{\mathbb{E}[(r_p - \mu_p)^-]^2}$ is the downside semi-deviation. The Sortino ratio offers a more nuanced evaluation of risk-adjusted return.

Maximum drawdown (MDD), Calmar ratio, and recovery time (RT) Maximum drawdown (MDD) is another key metric for downside risk. It measures the loss from the peak to the trough during a given period, relative to the peak value, and is defined as:

$$\text{MDD} = \frac{\text{Trough Value} - \text{Peak Value}}{\text{Peak Value}}.$$

The Calmar ratio provides a risk-adjusted return measure but uses MDD as the risk denominator:

$$\text{Calmar Ratio} = \frac{\mu_p - r}{\text{MDD}}.$$

Lastly, recovery time is the time (in days), *within a given testing period*, spent by a portfolio to rebound from its lowest point back to its previous peak. In our empirical results presented in Tables 1, 2, and 3, if a strategy's wealth trajectory (among the 100 independent trajectories) does not fully recover, we use the highest observed RT value from the other trajectories as a substitute, in order to calculate the average RT.

E Results Based on Simulated Data

This section presents the results of a simulation study on the vCTRL strategy. Recall that this strategy is the least modified and closest to the one generated by the baseline Algorithm 1 for which there are theoretically proved performance guarantees, and meanwhile one knows the "ground truth" in a simulation (as opposed to an empirical study). As such, the purpose of this study is to demonstrate that the convergence of the related parameters, its speed and the regret bound closely match the theoretical results.

Our experiment simulates a two-stock market environment, with each stock's price following a geometric Brownian motion. The model parameters are set as follows: drift vector $\mu = (0.2, 0.3)^\top$, marginal volatilities 0.3 and 0.4 with a correlation coefficient of 0.1, risk-free rate $r = 0.02$, initial wealth $x_0 = 1$, investment horizon $T = 1$, target expected terminal wealth $z = 1.4$, and temperature parameter $\gamma = 0.1$. The time discretization is set to be $\Delta t = 0.004$, and the total number of episodes is 10^5.

Algorithm 1 is initialized with $\boldsymbol{\theta} = (0,0)^\top$, $\phi_1 = (0,0)^\top$, $\phi_2 = I$, and $w = 1.5$. A total of 1000 independent simulation runs are conducted independently. As we know the oracle values of ϕ_1, ϕ_2, and w, we can compute the mean-squared errors (MSEs) of these learned parameters against number of episodes, both in log scale. Figures 2, 3, and 4 indicate that the learned parameters ϕ_1, ϕ_2, and w all converge, and converge rapidly after certain numbers of episodes. Moreover, by Theorem 1, the theoretical convergence rate of $\phi_{1,n}$ is $\frac{(\log n)^p \log\log n}{n}$ under the configuration specified in Remark 1. On a log scale, this corresponds to a slope close to -1, because $\log \frac{(\log n)^p \log\log n}{n} = -\log n + p \log\log n + \log\log\log n$, of which the first term dominates when n is large. Figure 2 shows that the fitted slope of the log average error against log number of episodes for ϕ_1 is -1.09, closely approximating the theoretical one. While theoretical convergence bounds for $\phi_2 = I$ and $w = 1.5$ are not yet available, Figures 3 and 4 show fitted slopes of -0.91 and -0.97 respectively, yielding convergence rates of these two parameters of close to $1/n$.

On the other hand, Theorem 3 stipulates that the theoretical regret bound of Algorithm 1 is $\sqrt{N(\log N)^p \log\log N}$ under the setting of Remark 1. On a log scale, this corresponds approximately to a slope close to 0.5, because $\log\sqrt{N(\log N)^p \log\log N} = \frac{1}{2}(\log N + p\log\log N + \log\log\log N)$. Figure 5 shows that the fitted slope of regret is 0.520 (on log scale), again very close to the theoretical one.

Figure 2: **Error of parameter** ϕ_1 The solid curves and the upper and lower boundaries of the shaded regions represent the average, 2.5% and 97.5% percentile of the error over 1000 independent simulation runs, respectively. The slope for ϕ_1 by least squares regression is -1.09. The vertical and horizontal axes are on natural log-scale.

F Additional Empirical Analysis

F.1 Pairwise test for statistical significance

We carry out pairwise tests employing the Wilcoxon rank test, a non-parametric alternative to the paired t-test, to evaluate the statistical significance of the comparisons between different investment strategies. Note that the Sharpe ratio typically does not follow a normal distribution. This rules out the paired t-test that relies on the normality assumption. By contrast, the Wilcoxon test, which assesses the median differences between pairs, does not require any specific distribution. Its another advantage is the resilience to outliers. Financial data, such as returns, often exhibit skewness or heavy tails, and the Wilcoxon test, focusing on ranks rather than absolute values, is less affected by extreme values.

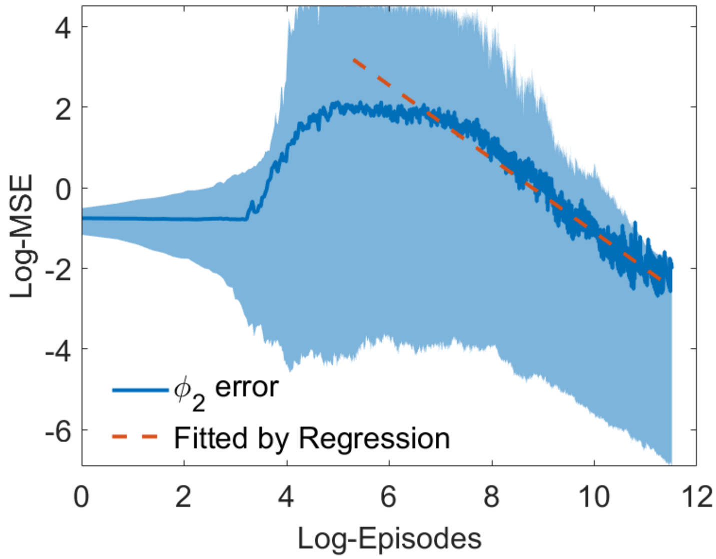

Figure 3: **Error of parameter** ϕ_2 The solid curves and the upper and lower boundaries of the shaded regions represent the average, 2.5% and 97.5% percentile of the error over 1000 independent simulation runs, respectively. The slope for ϕ_2 by least squares regression is -0.91. The vertical and horizontal axes are on natural log-scale.

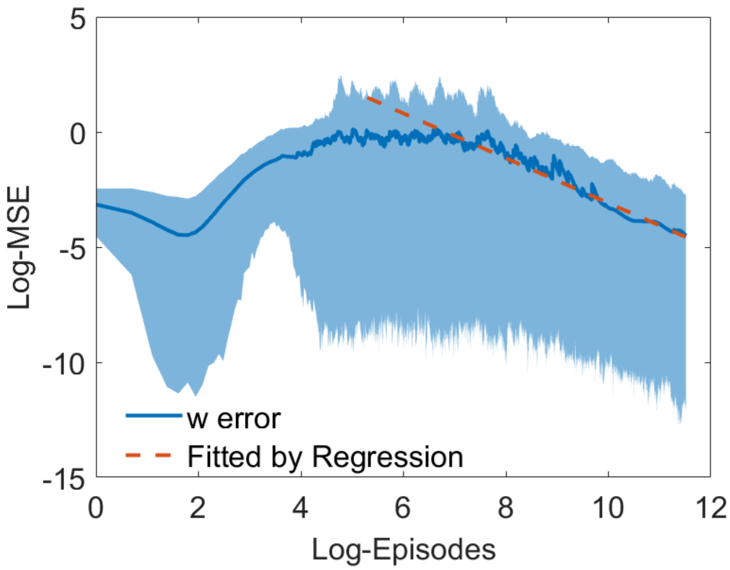

Figure 4: **Error of parameter** w The solid curves and the upper and lower boundaries of the shaded regions represent the average, 2.5% and 97.5% percentile of the error over 1000 independent simulation runs, respectively. The slope for w by least squares regression is -0.97. The vertical and horizontal axes are on natural log-scale.

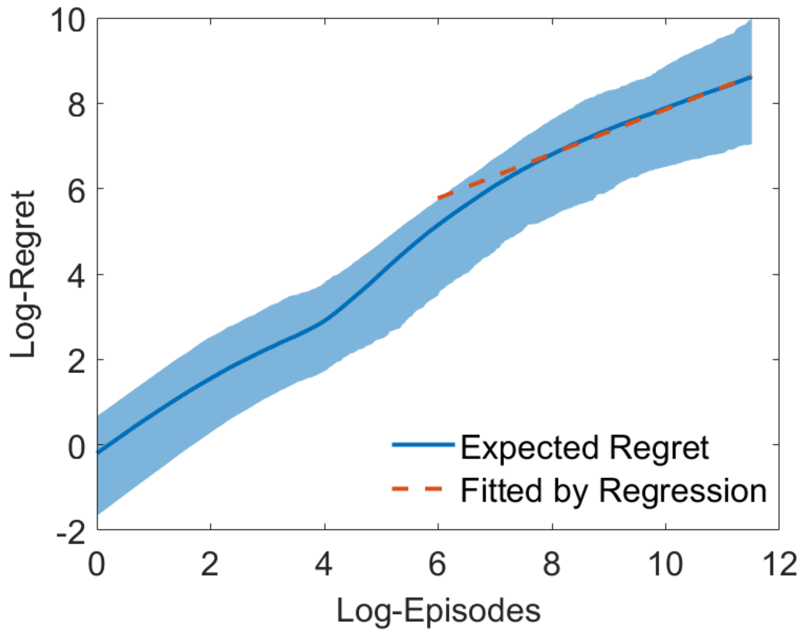

Figure 5: **Cumulative regret rate in number of episodes.** The solid blue curve and the upper and lower boundary of the shaded region represent the mean, 2.5% and 97.5% percentile of the regret over 1000 independent simulation runs, respectively. The red dashed line is the fitted value by linearly regressing the log average regret against the logarithm of the number of episodes starting from the 200th episode. The fitted slope by least squares regression is 0.520. The vertical and horizontal axes are on natural log-scale.

As our purpose is to identify superior performance between pairs, we calculate the p-value using a one-tailed Wilcoxon rank test for each pair of portfolio strategies. This approach allows us to ascertain whether one method is statistically significantly better than the other. As in the main empirical analysis, we conduct pairwise comparisons in three different time periods: the entire period from 2000 to 2020, the bear period from 2000 to 2010, and the bull period from 2010 to 2020. This segmentation provides insights into various strategies' adaptability and performance under varying market conditions. The results are presented in Tables 4, 5, and 6 respectively.

During the entire 20-year period (2000-2020), we observe standout performances from c-dCTRL, c-mCTRL, drmv, and rp. In particular, c-dCTRL outperforms, with a 99% confidence level, all the other methods. c-mCTRL exhibits a 95% likelihood of outperforming rp and a 90% likelihood of exceeding drmv in Sharpe ratio.

During the bear period (2000-2010), the leaders are c-dCTRL, c-mCTRL, ew, and rp. The performance differences are more pronounced, with c-dCTRL decisively being the top performer, followed by c-mCTRL. The statistical confidence in c-dCTRL's superiority exceeds 99%, confirming its outperformance and robustness in a volatile and unfavorable market environment.

As for the bull period (2010-2020), the leading ones are c-dCTRL, c-mCTRL, along with min_v, drmv, pmv. However, the pairwise p-values barely indicate significant differences in Sharpe ratio among these top-performers, and hence there is no decisive conclusion that one method dominates another. This in turn suggests that in a bull market, performance differences of the different strategies are insignificant, reconciling with the empirical finding reported in the main text.

Summarizing, the pairwise test reaffirms that the two RL strategies developed in this paper, c-dCTRL and c-mCTRL, outperform significantly, especially during the long aggregate period (2000-2020) and the bear phase (2000-2010). In particular, c-dCTRL demonstrates unparalleled superiority and proves to be the most robust and adaptable strategy. The c-mCTRL strategy also demonstrates a strong performance, ranking as a close second in the bear period. During the bull phase (2010-2020), on the other hand, most strategies including c-dCTRL and c-mCTRL perform well and the differences are rather small. Overall, c-dCTRL and c-mCTRL appear to be the all-round winners.

F.2 Different target returns

In the previously reported experiments, we fix the mean target (annualized) return to be 15% throughout, for all the MV related strategies. The 15% is roughly the annualized return of S&P 500 in the pretraining period 1990-2000. Theoretically, any expected target return (i.e. any value of z) in the MV model (2) is achievable so long as portfolios are unconstrained (Zhou and Li, 2000). In the RL setting, it is interesting to see if the learned policies can indeed attain any prescribed expected return. To empirically check this, we focus on the vCTRL strategy which is closest to being unconstrained. Conducting 100 independent experiments, we calculate the mean of their average annualized returns along with its 99% confidence intervals

Table 4: **p-values of out-of-sample performance of different investment strategies from 2000 to 2020.** The value at the entry of ith row and jth column, $i \neq j$, represents the p-value of the null hypothesis that the Sharpe ratio of strategy j is greater than that of strategy i.

	S&P 500	ew	mv	min_v	js	bl	lw	ff	rp	drmv	mctmv	dctmv	pmv	ddpg	ppo	v CTRL	p CTRL	c-m CTRL	c-d CTRL
S&P500	NaN	0.00	0.86	0.00	0.93	0.89	0.00	0.00	0.00	0.00	1.00	0.99	0.00	0.79	0.09	0.44	0.00	0.00	0.00
ew	1.00	NaN	1.00	0.69	1.00	1.00	0.40	0.87	0.00	0.00	1.00	1.00	0.25	1.00	1.00	1.00	1.00	0.00	0.00
mv	0.14	0.00	NaN	0.00	0.70	0.65	0.00	0.00	0.00	0.00	0.98	0.96	0.00	0.60	0.05	0.07	0.00	0.00	0.00
min_v	1.00	0.31	1.00	NaN	1.00	1.00	0.31	0.75	0.00	0.00	1.00	1.00	0.25	1.00	1.00	1.00	1.00	0.00	0.00
js	0.07	0.00	0.30	0.00	NaN	0.29	0.00	0.00	0.00	0.00	0.98	0.91	0.00	0.40	0.02	0.30	0.00	0.00	0.00
bl	0.11	0.00	0.35	0.00	0.71	NaN	0.00	0.00	0.00	0.00	0.97	0.95	0.00	0.46	0.07	0.04	0.00	0.00	0.00
lw	1.00	0.60	1.00	0.69	1.00	1.00	NaN	1.00	0.12	0.10	1.00	1.00	0.33	1.00	1.00	1.00	1.00	0.00	0.00
ff	1.00	0.13	1.00	0.25	1.00	1.00	0.00	NaN	0.00	0.00	1.00	1.00	0.01	1.00	1.00	1.00	1.00	0.00	0.00
rp	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00	NaN	0.19	1.00	1.00	0.83	1.00	1.00	1.00	1.00	0.00	0.00
drmv	1.00	1.00	1.00	1.00	1.00	1.00	0.90	1.00	0.81	NaN	1.00	1.00	0.83	1.00	1.00	1.00	1.00	0.00	0.00
mctmv	0.00	0.00	0.02	0.00	0.02	0.03	0.00	0.00	0.00	0.00	NaN	0.52	0.00	0.05	0.00	0.01	0.00	0.00	0.00
dctmv	0.01	0.00	0.04	0.00	0.09	0.05	0.00	0.00	0.00	0.00	0.48	NaN	0.00	0.04	0.01	0.04	0.00	0.00	0.00
pmv	1.00	0.75	1.00	0.75	1.00	1.00	0.67	0.99	0.17	0.17	1.00	1.00	NaN	1.00	1.00	1.00	1.00	0.00	0.00
ddpg	0.21	0.00	0.40	0.00	0.60	0.54	0.00	0.00	0.00	0.00	0.95	0.96	0.00	NaN	0.34	0.47	0.02	0.00	0.00
ppo	0.91	0.00	0.95	0.00	0.98	0.93	0.00	0.00	0.00	0.00	1.00	0.99	0.00	0.66	NaN	0.86	0.12	0.00	0.00
vCTRL	0.56	0.00	0.93	0.00	0.70	0.96	0.00	0.00	0.00	0.00	0.99	0.96	0.00	0.53	0.14	NaN	0.00	0.00	0.00
pCTRL	1.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.98	0.88	1.00	NaN	0.00	0.00
c-mCTRL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	NaN	0.00
c-dCTRL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	NaN

for a set of target return parameters ranging from 5% to 50% with an increment of 5%.

Table 5: **p-values of out-of-sample performance of different investment strategies from 2000 to 2010.** The value at the entry of ith row and jth column, $i \neq j$, represents the p-value of the null hypothesis that the Sharpe ratio of strategy j is greater than that of strategy i.

	S&P 500	ew	mv	min_v	js	bl	lw	ff	rp	drmv	mctmv	dctmv	pmv	ddpg	ppo	v CTRL	p CTRL	c-m CTRL	c-d CTRL
S&P500	NaN	0.00	0.01	0.00	0.00	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ew	1.00	NaN	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.80	0.18	0.00	0.00
mv	0.99	0.00	NaN	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.42	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00
min_v	1.00	0.00	1.00	NaN	1.00	1.00	0.04	0.05	0.00	0.00	1.00	1.00	0.25	0.29	0.02	0.00	0.00	0.00	0.00
js	1.00	0.00	1.00	0.00	NaN	1.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
bl	0.79	0.00	0.01	0.00	0.00	NaN	0.00	0.00	0.00	0.00	0.13	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lw	1.00	0.00	1.00	0.96	1.00	1.00	NaN	0.85	0.00	0.01	1.00	1.00	0.86	0.85	0.16	0.03	0.00	0.00	0.00
ff	1.00	0.00	1.00	0.95	1.00	1.00	0.15	NaN	0.00	0.00	1.00	1.00	0.77	0.77	0.13	0.01	0.00	0.00	0.00
rp	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NaN	1.00	1.00	1.00	1.00	1.00	0.84	0.35	0.09	0.00	0.00
drmv	1.00	0.00	1.00	1.00	1.00	1.00	0.99	1.00	0.00	NaN	1.00	1.00	1.00	0.99	0.68	0.10	0.03	0.00	0.00
mctmv	1.00	0.00	0.58	0.00	0.00	0.87	0.00	0.00	0.00	0.00	NaN	0.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dctmv	1.00	0.00	0.77	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.23	NaN	0.00	0.00	0.00	0.00	0.00	0.00	0.00
pmv	1.00	0.00	1.00	0.75	1.00	1.00	0.14	0.23	0.00	0.00	1.00	1.00	NaN	0.78	0.15	0.01	0.00	0.00	0.00
ddpg	1.00	0.00	1.00	0.71	0.99	1.00	0.15	0.23	0.00	0.01	1.00	1.00	0.22	NaN	0.13	0.06	0.00	0.00	0.00
ppo	1.00	0.04	1.00	0.98	1.00	1.00	0.84	0.87	0.16	0.32	1.00	1.00	0.85	0.87	NaN	0.35	0.01	0.00	0.00
vCTRL	1.00	0.20	1.00	1.00	1.00	1.00	0.97	0.99	0.65	0.90	1.00	1.00	0.99	0.94	0.65	NaN	0.00	0.00	0.00
pCTRL	1.00	0.82	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.97	1.00	1.00	1.00	1.00	0.99	1.00	NaN	0.00	0.00
c-mCTRL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	NaN	0.00
c-dCTRL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	NaN

The results, detailed in Table 7, demonstrate that the actual realized annual returns fall within the

Table 6: **p-values of out-of-sample performance of different investment strategies from 2010 to 2020.** The value at the entry of ith row and jth column, $i \neq j$, represents the p-value of the null hypothesis that the Sharpe ratio of strategy j is greater than that of strategy i.

	S&P 500	ew	mv	min_v	js	bl	lw	ff	rp	drmv	mctmv	dctmv	pmv	ddpg	ppo	v CTRL	p CTRL	c-m CTRL	c-d CTRL
S&P500	NaN	1.00	1.00	0.00	1.00	1.00	0.04	0.32	0.08	0.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00
ew	0.00	NaN	1.00	0.00	1.00	0.54	0.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	1.00	0.99	0.75	0.00	0.00
mv	0.00	0.00	NaN	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	1.00	0.56	0.12	0.00	0.00
min_v	1.00	1.00	1.00	NaN	1.00	1.00	0.77	0.99	1.00	0.94	1.00	1.00	0.63	1.00	1.00	1.00	1.00	0.87	0.72
js	0.00	0.00	0.00	0.00	NaN	0.00	0.00	0.00	0.00	0.00	0.37	0.36	0.00	0.69	0.49	0.00	0.00	0.00	0.00
bl	0.00	0.46	1.00	0.00	1.00	NaN	0.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	1.00	0.99	0.69	0.00	0.00
lw	0.96	1.00	1.00	0.23	1.00	1.00	NaN	1.00	0.84	0.46	1.00	1.00	0.11	1.00	1.00	1.00	1.00	0.57	0.42
ff	0.68	1.00	1.00	0.01	1.00	1.00	0.00	NaN	0.37	0.07	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00
rp	0.92	1.00	1.00	0.00	1.00	1.00	0.16	0.63	NaN	0.00	1.00	1.00	0.08	1.00	1.00	1.00	1.00	0.12	0.05
drmv	1.00	1.00	1.00	0.06	1.00	1.00	0.54	0.93	1.00	NaN	1.00	1.00	0.36	1.00	1.00	1.00	1.00	0.56	0.39
mctmv	0.00	0.00	0.00	0.00	0.63	0.00	0.00	0.00	0.00	0.00	NaN	0.83	0.00	0.82	0.57	0.00	0.00	0.00	0.00
dctmv	0.00	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.00	0.00	0.17	NaN	0.00	0.77	0.48	0.00	0.00	0.00	0.00
pmv	1.00	1.00	1.00	0.37	1.00	1.00	0.89	1.00	0.92	0.64	1.00	1.00	NaN	1.00	1.00	1.00	1.00	0.83	0.66
ddpg	0.00	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.18	0.23	0.00	NaN	0.15	0.00	0.00	0.00	0.00
ppo	0.00	0.00	0.00	0.00	0.51	0.00	0.00	0.00	0.00	0.00	0.43	0.52	0.00	0.85	NaN	0.01	0.00	0.00	0.00
vCTRL	0.00	0.01	0.44	0.00	1.00	0.01	0.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	0.99	NaN	0.00	0.00	0.00
pCTRL	0.00	0.25	0.88	0.00	1.00	0.31	0.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	1.00	1.00	NaN	0.00	0.00
c-mCTRL	1.00	1.00	1.00	0.13	1.00	1.00	0.43	1.00	0.88	0.44	1.00	1.00	0.17	1.00	1.00	1.00	1.00	NaN	0.00
c-dCTRL	1.00	1.00	1.00	0.28	1.00	1.00	0.58	1.00	0.95	0.61	1.00	1.00	0.34	1.00	1.00	1.00	1.00	1.00	NaN

99% confidence intervals of the respective target returns. Note in particular the tight confidence intervals and small standard errors in these results. This ability of reaching a desired target return (even at 50%, a level that may be considered excessively ambitious under typical market conditions) attests the remarkable learning ability of CTRL strategies.

Target Return	Mean Realized Return	Standard Error	Confidence Interval
0.05	0.05002	7.43909e-05	[0.05, 0.05004]
0.10	0.10008	0.000311455	[0.1, 0.10016]
0.15	0.15015	0.000645542	[0.14998, 0.15032]
0.20	0.20026	0.0011176	[0.19997, 0.20054]
0.25	0.25041	0.00193245	[0.24991, 0.25091]
0.30	0.30053	0.00239679	[0.29992, 0.30115]
0.35	0.35074	0.0045035	[0.34957, 0.3519]
0.40	0.40096	0.0057096	[0.39948, 0.40243]
0.45	0.45115	0.00664192	[0.44943, 0.45286]
0.50	0.50128	0.00563606	[0.49983, 0.50274]

Table 7: **Performance of vCTRL with different target returns.** Mean Realized Return and Standard Error represent the average and standard error of realized annual returns over 100 independent experiments. Confidence Interval is calculated with 99% confidence.

G Proofs of Statements

G.1 Proof of Theorem 1

To start, assuming there are no extra constraints (such as the leverage constraint) and portfolios are self-financing, the agent's discounted wealth process (1) in the n-th iteration satisfies the wealth equation

$$\mathrm{d}x_n(t) = (\mu - r\boldsymbol{e}_d)^\top u_n(t)\mathrm{d}t + u_n(t)^\top \sigma \mathrm{d}W_n(t), \qquad 0 \leqslant t \leqslant T; \;\; x_n(0) = x_0, \tag{41}$$

where W_n is a Brownian motion in the n-th iteration, and (with a slight abuse of notation) $u_n(t) = u_n(t, x_n(t))$ while $u_n(t,x) \sim \mathcal{N}\left(-\phi_{1,n}(x-w_n), \phi_{2,n}e^{\phi_3(T-t)}\right)$ independent of W_n.

Recall that $\theta_3 = \phi_3$ is fixed and not updated in our algorithm, and that $Z_{1,n}(T)$ and $Z_{2,n}(T)$ are defined in (15) and (16). Denote by $\xi_n = (\xi_{1,n}, \xi_{2,n})^\top$ the "noise" parts of these random variables, namely,

$$\xi_{1,n+1} = Z_{1,n}(T) - h_1(\phi_{1,n}, \phi_{2,n}, w_n) \;\; \text{where } h_1(\phi_{1,n}, \phi_{2,n}, w_n) = \mathbb{E}\left[Z_{1,n}(T) \middle| \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, w_n\right],$$

$$\xi_{2,n+1} = Z_{2,n}(T) - h_2(\phi_{1,n}, \phi_{2,n}, w_n) \;\; \text{where } h_2(\phi_{1,n}, \phi_{2,n}, w_n) = \mathbb{E}\left[Z_{2,n}(T) \middle| \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, w_n\right].$$

Similarly, define $\xi_{w,n} \in \mathbb{R}$ as the noise counterpart in updating w:

$$\xi_{w,n+1} = x_n(T) - z - h_w(\phi_{1,n}, \phi_{2,n}, w_n) \;\; \text{where } h_w(\phi_{1,n}, \phi_{2,n}, w_n) = \mathbb{E}\left[x_n(T) - z \middle| \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, w_n\right].$$

Then the updating rules for $\boldsymbol{\phi}$ and w can be rewritten as

$$\begin{aligned}
\phi_{1,n+1} &= \Pi_{K_{1,n+1}}\left(\phi_{1,n} - a_n[h_1(\phi_{1,n}, \phi_{2,n}, w_n; \phi_3) + \xi_{1,n+1}]\right), \\
\phi_{2,n+1} &= \Pi_{K_{2,n+1}}\left(\phi_{2,n} + a_n[h_2(\phi_{1,n}, \phi_{2,n}, w_n; \phi_3) + \xi_{2,n+1}]\right), \\
w_{n+1} &= \Pi_{K_{w,n+1}}\left(w_n - a_{w,n}[h_w(\phi_{1,n}, \phi_{2,n}, w_n; \phi_3) + \xi_{w,n+1}]\right).
\end{aligned} \tag{42}$$

The proof of Theorem 1 will be carried out through several steps. It will apply some general stochastic approximation results including those in Andradóttir (1995) and Broadie et al. (2011). However, we need to verify several assumptions for our specific problem and to overcome difficulties arising from those that are not satisfied by our problem.

Moment estimates.

First we establish the moment expressions and estimates for the wealth trajectory under the policy (9).

Lemma 1. *Let $\{\tilde{x}^{\boldsymbol{\phi}}(t) : 0 \leqslant t \leqslant T\}$ be the wealth trajectory under the policy* (9)*. Then we have*

$$\begin{aligned}
\mathbb{E}[\tilde{x}^{\boldsymbol{\phi}}(t) - w] =& (x_0 - w)e^{-(\mu-r)^\top \phi_1 t}, \\
\mathbb{E}[(\tilde{x}^{\boldsymbol{\phi}}(t) - w)^2] =& \left[(x_0 - w)^2 + \frac{\langle \Sigma, \phi_2 \rangle e^{\phi_3 T}}{-2(\mu - r)^\top \phi_1 + \langle \Sigma, \phi_1 \phi_1^\top \rangle + \phi_3}\right] e^{(-2(\mu-r)^\top \phi_1 + \langle \Sigma, \phi_1 \phi_1^\top \rangle)t} \\
&- \frac{\langle \Sigma, \phi_2 \rangle e^{\phi_3 (T-t)}}{-2(\mu - r)^\top \phi_1 + \langle \Sigma, \phi_1 \phi_1^\top \rangle + \phi_3}.
\end{aligned} \tag{43}$$

Moreover, there exists a constant $C > 0$ that only depends on μ, r, x_0, T and Σ such that we have

$$\begin{aligned}
&\mathbb{E}[(\tilde{x}^{\boldsymbol{\phi}}(t) - w)^2] \leqslant C(1 + |w|^2 + |\phi_2|) \exp\left(C|\phi_1|^2 t\right), \\
&\mathrm{Var}\left(\tilde{x}^{\boldsymbol{\phi}}(t)\right) \leqslant C(1 + |w|^2 + |\phi_2|) \exp\left(C|\phi_1|^2 t\right), \\
&\mathbb{E}[(\tilde{x}^{\boldsymbol{\phi}}(t) - w)^4] \leqslant C(1 + |w|^4 + |\phi_2|^2) \exp\left(C|\phi_1|^4 t\right), \\
&\mathbb{E}[(\tilde{x}^{\boldsymbol{\phi}}(t) - w)^8] \leqslant C(1 + |w|^8 + |\phi_2|^4) \exp\left(C|\phi_1|^8 t\right).
\end{aligned} \tag{44}$$

Proof. Denote $\hat{x}(t) = \tilde{x}^{\boldsymbol{\phi}}(t) - w$. It follows from (24) that

$$\hat{x}(t) = x_0 - w + \int_0^t -(\mu - r)^\top \phi_1 \hat{x}(s)\mathrm{d}s + \int_0^t \sqrt{\hat{x}(s)^2 \phi_1^\top \Sigma \phi_1 + \langle \Sigma, \phi_2 e^{\phi_3 (T-s)} \rangle}\mathrm{d}W(s). \tag{45}$$

Taking expectation on both sides and solving the resulting ODE, we obtain the first equation of (43).

Apply Itô's lemma to $\hat{x}^2(t)$ in (45) and take expectation on both sides to obtain

$$\mathbb{E}[\hat{x}^2(t)] = (x_0 - w)^2 + \mathbb{E}\int_0^t [-2(\mu - r)^\top \hat{x}^2(s) + \langle \Sigma, \phi_1 \phi_1^\top \hat{x}^2(s) + \phi_2 e^{\phi_3 (T-s)} \rangle]\mathrm{d}s.$$

Solving the above ODE in $\mathbb{E}[\hat{x}^2(\cdot)]$ we obtain the second equation of (43).

Next, we take the eighth power and then apply expectation on both sides of (45). By Hölder's inequality, we have

$$
\begin{aligned}
\mathbb{E}[\hat{x}(t)^8] =&\mathbb{E}\left[\left(x_0 - w + \int_0^t -(\mu - r)^\top \phi_1 \hat{x}(s)\mathrm{d}s + \int_0^t \sqrt{\hat{x}(s)^2 \phi_1 \Sigma \phi_1 + \langle \Sigma, \phi_2 e^{\phi_3(T-s)} \rangle}\mathrm{d}W(s)\right)^8\right] \\
\leqslant& C|x_0 - w|^8 + C\mathbb{E}\left[\left(\int_0^t -(\mu - r)^\top \phi_1 \hat{x}(s)\mathrm{d}s\right)^8\right] + C\mathbb{E}\left[\left(\int_0^t \sqrt{\hat{x}(s)^2 \phi_1 \Sigma \phi_1 + \langle \Sigma, \phi_2 e^{\phi_3(T-s)} \rangle}\mathrm{d}W(s)\right)^8\right] \\
\leqslant& C|x_0 - w|^8 + C((\mu - r)^\top \phi_1)^8 \mathbb{E}\left[(\int_0^t \hat{x}(s)\mathrm{d}s)^8\right] + C\mathbb{E}\left[(\int_0^t \hat{x}(s)^2 \phi_1 \Sigma \phi_1 + \langle \Sigma, \phi_2 e^{\phi_3(T-s)} \rangle \mathrm{d}s)^4\right] \\
\leqslant& C|x_0 - w|^8 + C|\phi_1|^8 \mathbb{E}\left[\int_0^t \hat{x}(s)^8 \mathrm{d}s\right] + C\mathbb{E}\left[\int_0^t \hat{x}(s)^8 |\phi_1|^8 + |\phi_2|^4 \mathrm{d}s\right].
\end{aligned} \tag{46}
$$

Gronwall's inequality thus leads to the fourth inequality of (44). The similar argument can be applied to prove the remaining inequalities of (44) for the moment estimate of the second and fourth orders. In particular, $\mathrm{Var}\left(\tilde{x}^{\boldsymbol{\phi}}(t)\right) \leqslant \mathbb{E}[(\tilde{x}^{\boldsymbol{\phi}}(t) - w)^2]$. □

Next, we estimate the variances of the increments $Z_{1,n}(T)$ and $Z_{2,n}(T)$ defined in (15) and (16) respectively.

Lemma 2. *There exists a constant $C > 0$ such that*

$$
\begin{aligned}
\mathrm{Var}\left(Z_{1,n}(T)\middle|\boldsymbol{\theta}_n, \boldsymbol{\phi}_n, w_n\right) &\leqslant C\left(1 + |w_n|^{16} + |\phi_{1,n}|^8 + |\phi_{2,n}|^8 + |b_n|^8\right) e^{C|\phi_{1,n}|^8}. \\
\mathrm{Var}\left(Z_{2,n}(T)\middle|\boldsymbol{\theta}_n, \boldsymbol{\phi}_n, w_n\right) &\leqslant C\left(1 + |w_n|^{16} + |\phi_{1,n}|^8 + |\phi_{2,n}|^8\right) e^{C|\phi_{1,n}|^8}.
\end{aligned} \tag{47}
$$

Proof. We first derive the dynamics of $\{(Z_{1,n}(t), Z_{2,n}(t)) : 0 \leqslant t \leqslant T\}$. Applying Itô's lemma we obtain

$$
\begin{aligned}
dJ(t, x_n(t); w_n; \boldsymbol{\theta}_n) =& \bigg(\frac{\partial J(t, x_n(t); w_n; \boldsymbol{\theta}_n)}{\partial t} + (\mu - re_d)^\top u_n(t) \frac{\partial J(t, x_n(t); w_n; \boldsymbol{\theta}_n)}{\partial x} \\
&+ \frac{u_n(t)^\top \Sigma u_n(t)}{2} \frac{\partial^2 J(t, x_n(t); w_n; \boldsymbol{\theta}_n)}{\partial x^2}\bigg) dt + u_n(t)^\top \sigma \frac{\partial J(t, x_n(t); w_n; \boldsymbol{\theta}_n)}{\partial x} dW_n(t).
\end{aligned} \tag{48}
$$

Noting the explicit forms (8) and (9), we deduce from (15) and (16) that

$$
\begin{aligned}
&\mathrm{d}Z_{1,n}(t)\\
=&\frac{\partial \log \pi(u_n(t)|t, x_n(t); w_n; \phi_n)}{\partial \phi_1}\Bigg[\Bigg(\theta_3(x_n(t)-w_n)^2 e^{-\theta_3(T-t)} + 2(x_n(t)-w_n)e^{-\theta_3(T-t)}(\mu - r e_d)^\top u_n(t) +\\
&e^{-\theta_3(T-t)} u_n(t)^\top \sigma\sigma^\top u_n(t) + 2\theta_{2,n} t + \theta_{1,n}\Bigg)\mathrm{d}t + 2(x_n(t)-w_n)u_n(t)^\top \sigma e^{-\theta_3(T-t)}\mathrm{d}W(t) + \gamma p^\phi(t)\mathrm{d}t\Bigg] + \gamma \frac{\partial p^\phi(t)}{\partial \phi_1}\mathrm{d}t\\
=&-e^{-\phi_3(T-t)}\phi_{2,n}^{-1}\left[(x_n(t)-w_n)u_n(t) + (x_n(t)-w_n)^2\phi_{1,n}\right]\\
&\times\Bigg[\Bigg(\theta_3(x_n(t)-w_n)^2 e^{-\theta_3(T-t)} + 2(x_n(t)-w_n)e^{-\theta_3(T-t)}(\mu - r e_d)^\top u_n(t) + e^{-\theta_3(T-t)} u_n(t)^\top \sigma\sigma^\top u_n(t) + 2\theta_{2,n} t + \theta_{1,n}\Bigg)\mathrm{d}t\\
&+ 2(x_n(t)-w_n)u_n(t)^\top \sigma e^{-\theta_3(T-t)}\mathrm{d}W(t) + \gamma(-\frac{d}{2}\log(2\pi e) + \frac{1}{2}\log(\det(\phi_{2,n}^{-1}))\mathrm{d}t - \frac{d}{2}\phi_3(T-t))\mathrm{d}t\Bigg]\\
=&-e^{-\phi_3(T-t)}\phi_{2,n}^{-1}\left[(x_n(t)-w_n)u_n(t) + (x_n(t)-w_n)^2\phi_{1,n}\right]\\
&\times\Bigg[\Bigg(\theta_3(x_n(t)-w_n)^2 e^{-\theta_3(T-t)} + 2(x_n(t)-w_n)e^{-\theta_3(T-t)}(\mu - r e_d)^\top u_n(t) + e^{-\theta_3(T-t)} u_n(t)^\top \sigma\sigma^\top u_n(t) + 2\theta_{2,n} t + \theta_{1,n}\Bigg)\\
&+ \gamma(-\frac{d}{2}\log(2\pi e) + \frac{1}{2}\log(\det(\phi_{2,n}^{-1})) - \frac{d}{2}\phi_3(T-t))\Bigg]\mathrm{d}t\\
&- 2e^{-\phi_3(T-t)}\phi_{2,n}^{-1}\left[(x_n(t)-w_n)u_n(t) + (x_n(t)-w_n)^2\phi_{1,n}\right](x_n(t)-w_n)u_n(t)^\top \sigma e^{-\theta_3(T-t)}\mathrm{d}W_n(t)\\
\triangleq& Z_{1,n}^{(1)}(t)\mathrm{d}t + Z_{1,n}^{(2)}(t)\mathrm{d}W_n(t),
\end{aligned}
\tag{49}
$$

and

$$
\begin{aligned}
&\mathrm{d}Z_{2,n}(t)\\
=&\frac{\partial \log \pi(u_n(t)|t, x_n(t); w_n; \phi_n)}{\partial \phi_2^{-1}}\Bigg[\Bigg(\theta_3(x_n(t)-w_n)^2 e^{-\theta_3(T-t)} + 2(x_n(t)-w_n)e^{-\theta_3(T-t)}(\mu - r e_d)^\top u_n(t)\\
&+ e^{-\theta_3(T-t)} u_n(t)^\top \sigma\sigma^\top u_n(t) + 2\theta_{2,n} t + \theta_{1,n}\Bigg)\mathrm{d}t + 2(x_n(t)-w_n)u_n(t)^\top \sigma e^{-\theta_3(T-t)}\mathrm{d}W(t) + \gamma p^\phi(t)\mathrm{d}t\Bigg] + \gamma \frac{\partial p^\phi(t)}{\partial \phi_2^{-1}}\mathrm{d}t\\
=&\left[\frac{1}{2}\phi_{2,n} - \frac{1}{2}e^{-\phi_3(T-t)}(u_n(t) + \phi_{1,n}(x_n(t)-w_n))(u_n(t) + \phi_{1,n}(x_n(t)-w_n))^\top\right]\\
&\times\Bigg[\Bigg(\theta_3(x_n(t)-w_n)^2 e^{-\theta_3(T-t)} + 2(x_n(t)-w_n)e^{-\theta_3(T-t)}(\mu - r e_d)^\top u_n(t) + e^{-\theta_3(T-t)} u_n(t)^\top \sigma\sigma^\top u_n(t) + 2\theta_{2,n} t + \theta_{1,n}\Bigg)\mathrm{d}t\\
&+ 2(x_n(t)-w_n)u_n(t)^\top \sigma e^{-\theta_3(T-t)}\mathrm{d}W(t) + \gamma(-\frac{d}{2}\log(2\pi e) + \frac{1}{2}\log(\det(\phi_{2,n}^{-1}))\mathrm{d}t - \frac{d}{2}\phi_3(T-t))\mathrm{d}t\Bigg] + \gamma\frac{\phi_{2,n}}{2}\mathrm{d}t\\
=&\left[\frac{1}{2}\phi_{2,n} - \frac{1}{2}e^{-\phi_3(T-t)}(u_n(t) + \phi_{1,n}(x_n(t)-w_n))(u_n(t) + \phi_{1,n}(x_n(t)-w_n))^\top\right]\\
&\times\Bigg\{\Bigg[\Bigg(\theta_3(x_n(t)-w_n)^2 e^{-\theta_3(T-t)} + 2(x_n(t)-w_n)e^{-\theta_3(T-t)}(\mu - r e_d)^\top u_n(t) + e^{-\theta_3(T-t)} u_n(t)^\top \sigma\sigma^\top u_n(t) + 2\theta_{2,n} t + \theta_{1,n}\Bigg)\\
&+ \gamma(-\frac{d}{2}\log(2\pi e) + \frac{1}{2}\log(\det(\phi_{2,n}^{-1})) - \frac{d}{2}\phi_3(T-t))\Bigg] + \gamma\frac{\phi_{2,n}}{2}\Bigg\}\mathrm{d}t\\
&+ 2\left\{\frac{1}{2}\phi_{2,n} - \frac{1}{2}e^{-\phi_3(T-t)}[u_n(t) + \phi_{1,n}(x_n(t)-w_n)][u_n(t) + \phi_{1,n}(x_n(t)-w_n)]^\top\right\})(x_n(t)-w_n)u_n(t)^\top \sigma e^{-\theta_3(T-t)}\mathrm{d}W_n(t)\\
\triangleq& Z_{2,n}^{(1)}(t)\mathrm{d}t + Z_{2,n}^{(2)}(t)\mathrm{d}W_n(t).
\end{aligned}
\tag{50}
$$

Noting that $u_n(t) \equiv u_n(t, x_n(t))$ while $u_n(t,x) \sim \mathcal{N}\left(-\phi_{1,n}(x - w_n), \phi_{2,n} e^{\phi_3(T-t)}\right)$, we can easily upper bound $|Z_{1,n}^{(1)}|^2$ and $|Z_{1,n}^{(2)}|^2$ by

$$\begin{aligned}\mathbb{E}[|Z_{1,n}^{(1)}(t)|^2 | \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, w_n, x_n(t)] \leqslant & C\Bigg[1 + (x_n(t) - w_n)^4 |\phi_{2,n}^{-1}|^2 + (x_n(t) - w_n)^8 |\phi_{1,n}|^2 |\phi_{2,n}^{-1}|^2 \\ & + (x_n(t) - w_n)^8 + (x_n(t) - w_n)^8 |\phi_{1,n}|^4 + (\log\det(\phi_{2,n}^{-1}))^4\Bigg],\end{aligned}$$

$$\begin{aligned}\mathbb{E}[|Z_{1,n}^{(2)}(t)|^2 | \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, w_n, x_n(t)] \leqslant & C\Bigg[1 + (x_n(t) - w_n)^4 |\phi_{2,n}^{-1}|^2 + (x_n(t) - w_n)^8 |\phi_{1,n}|^2 |\phi_{2,n}^{-1}|^2 \\ & + (x_n(t) - w_n)^4 |\phi_{1,n}|^4 + (x_n(t) - w_n)^4 |\phi_{2,n}|^2\Bigg].\end{aligned}$$

By virtue of the projection $\phi_{2,n} \geq \frac{1}{b_n} I$, or $|\phi_{2,n}^{-1}| \leqslant b_n$, we conclude from Lemma 1 that

$$\begin{aligned}& \mathbb{E}[(|Z_{1,n}^{(1)}(t)|^2 + |Z_{1,n}^{(2)}(t)|^2) | \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, w_n] \\ \leqslant & C\Bigg[1 + (1 + |w_n|^4 + |\phi_{2,n}|^2) \exp\left(C|\phi_{1,n}|^4 t\right)(b_n^2 + |\phi_{2,n}|^2 + |\phi_{1,n}|^4) \\ & + (1 + |w_n|^8 + |\phi_{2,n}|^4) \exp(C|\phi_{1,n}|^8 t)(1 + b_n^4 + |\phi_{1,n}|^4)) + (d \log b_n)^4\Bigg],\end{aligned}$$

leading to the first inequality of (47). The second inequality of (47) can be proved similarly.

□

Explicit expressions of mean increments

Next, we derive the analytical forms of the functions h_1, h_2, h_w, which are the means of the increments in the algorithms approximating ϕ_1, ϕ_2 and ϕ_w respectively.

To start, note that $\{Z_{1,n}^{(2)}(t) : 0 \leqslant t \leqslant T\}$ and $\{Z_{2,n}^{(2)}(t) : 0 \leqslant t \leqslant T\}$ are both square integrable based on the proof of Lemma 2 along with the moment estimates in Lemma 1. Thus, when we integrate (49) and (50) and take expectation, the Itô integrals vanish. Denote

$$\begin{aligned}& A_n = (\mu - r)^\top \phi_{1,n}, \quad B_n = \langle \sigma\sigma^\top, \phi_{1,n}\phi_{1,n}^\top \rangle, \quad E_n = \langle \sigma\sigma^\top, \phi_{2,n} e^{\phi_3(T-t)} \rangle, \\ & G = e^{-\theta_3(T-t)}, \quad H_n = -\frac{d}{2}\log(2\pi e) + \frac{1}{2}\log(\det(\phi_{2,n}^{-1})), \quad P_n = 2\theta_{2,n} t - \frac{\gamma d}{2}\phi_3(T-t) + \theta_{1,n} + \gamma H_n.\end{aligned}$$

Then it follows from (49) and (50) that

$$
\begin{aligned}
\mathrm{d}\mathbb{E}[Z_{1,n}(t)] =& \mathbb{E}\left\{ \frac{\partial \log \pi\left(u_n(t) \mid t, x_n(t); w_n; \boldsymbol{\phi}_n\right)}{\partial \phi_1} \left[\mathrm{d}J\left(t, x_n(t); w_n; \boldsymbol{\theta}_n\right) + \gamma \hat{p}\left(t, x_n(t), \boldsymbol{\phi}_n\right) \mathrm{d}t\right] + \gamma \frac{\partial \hat{p}}{\partial \phi_1}\left(t, x_n(t), \boldsymbol{\phi}_n\right) \mathrm{d}t \right\} \\
=& \mathbb{E}\left\{ -G[(x_n(t) - w_n)\phi_{2,n}^{-1} u_n(t) + (x_n(t) - w_n)^2 \phi_{2,n}^{-1} \phi_{1,n}] \right. \\
& \left. \times [\theta_3 G(x_n(t) - w_n)^2 + 2G(x_n(t) - w)(\mu - r)^\top u_n(t) + G(x_n(t) - w) u_n(t) \langle \sigma\sigma^\top, u_n(t) u_n(t)^\top \rangle + P_n] \mathrm{d}t \right\} \\
=& \mathbb{E}\left\{ \{-G\phi_{2,n}^{-1}[\theta_3 G(x_n(t) - w_n)^3 u_n(t) + 2G(x_n(t) - w_n)^2 u_n(t) u_n(t)^\top (\mu - r) \right. \\
& + G(x_n(t) - w_n) u_n(t) \langle \Sigma, u_n(t) u_n(t)^\top \rangle + P_n (x_n(t) - w_n) u_n(t)]\} \mathrm{d}t \\
& + \{-G\phi_{2,n}^{-1}\phi_{1,n}[\theta_3 G(x_n(t) - w_n)^4 + 2G(x_n(t) - w_n)^3 (\mu - r)^\top u_n(t) \\
& \left. + G(x_n(t) - w_n)^2 \langle \sigma\sigma^\top, u_n(t) u_n(t)^\top \rangle + P_n (x_n(t) - w_n)^2]\} \mathrm{d}t \right\} \\
=& \mathbb{E}(x_n(t) - w_n)^2 [2G(-(\mu - r) + \sigma\sigma^\top \phi_{1,n})] \mathrm{d}t,
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{d}\mathbb{E}[Z_{2,n}(t)] =& \mathbb{E}\left\{ \frac{\partial \log \pi\left(u_n(t) \mid t, x_n(t); w_n, \boldsymbol{\phi}_n\right)}{\partial \phi_2^{-1}} \left[\mathrm{d}J\left(t, x_n(t); w_n; \boldsymbol{\theta}_n\right) + \gamma \hat{p}\left(t, x_n(t), \boldsymbol{\phi}_n\right) \mathrm{d}t\right] + \gamma \frac{\partial \hat{p}}{\partial \phi_2^{-1}}\left(t, x_n(t), \phi\right) \mathrm{d}t \right\} \\
=& \mathbb{E}\left\{ [\frac{1}{2}\phi_{2,n} - \frac{1}{2} G(u_n(t) u_n(t)^\top + u_n(t) \phi_{1,n}^\top (x_n(t) - w_n) + \phi_{1,n} u_n(t)^\top (x_n(t) - w_n) + \phi_{1,n}\phi_{1,n}^\top (x_n(t) - w_n)^2)] \right. \\
& \left. \times [\theta_3 G(x_n(t) - w_n)^2 + 2G(x_n(t) - w)(\mu - r)^\top u_n(t) + G\langle \sigma\sigma^\top, u_n(t) u_n(t)^\top \rangle + P_n] \mathrm{d}t + \gamma \frac{\phi_{2,n}}{2} \mathrm{d}t \right\} \\
=& \frac{1}{2}\phi_{2,n}[(\theta_3 - 2A_n + B_n) G \mathbb{E}((x_n(t) - w_n)^2) + GE_n + P_n + \gamma] \mathrm{d}t \\
& - \frac{1}{2} G \mathbb{E}\left\{ \theta_3 G(x_n(t) - w_n)^2 u_n(t) u_n(t)^\top + 2G(x_n(t) - w_n) u_n(t) u_n(t)^\top (\mu - r)^\top u_n(t) \right. \\
& + G u_n(t) u_n(t)^\top \langle \sigma\sigma^\top, u_n(t) u_n(t)^\top \rangle + P_n u_n(t) u_n(t)^\top \\
& + \theta_3 G(x_n(t) - w_n)^3 u_n(t) \phi_{1,n}^\top + 2G(x_n(t) - w_n)^2 u_n(t) \phi_{1,n}^\top (\mu - r)^\top u_n(t) \\
& + G(x_n(t) - w_n) u_n(t) \phi_{1,n}^\top \langle \sigma\sigma^\top, u_n(t) u_n(t)^\top \rangle + P_n (x_n(t) - w_n) u_n(t) \phi_{1,n}^\top \\
& + \theta_3 G(x_n(t) - w_n)^3 \phi_{1,n} u_n(t)^\top + 2G(x_n(t) - w_n)^2 \phi_{1,n} u_n(t)^\top (\mu - r)^\top u_n(t) \\
& + G(x_n(t) - w_n) \phi_{1,n} u_n(t)^\top \langle \sigma\sigma^\top, u_n(t) u_n(t)^\top \rangle + P_n (x_n(t) - w_n) \phi_{1,n} u_n(t)^\top \\
& + \theta_3 G(x_n(t) - w_n)^4 \phi_{1,n}\phi_{1,n}^\top + 2G(x_n(t) - w_n)^3 \phi_{1,n}\phi_{1,n}^\top (\mu - r)^\top u_n(t) \\
& \left. + G(x_n(t) - w_n)^2 \phi_{1,n}\phi_{1,n}^\top \langle \sigma\sigma^\top, u_n(t) u_n(t)^\top \rangle + P_n (x_n(t) - w_n)^2 \phi_{1,n}\phi_{1,n}^\top \right\} \mathrm{d}t \\
=& \frac{1}{2}[\gamma \phi_{2,n} - 2\phi_{2,n} \sigma\sigma^\top \phi_{2,n}] \mathrm{d}t.
\end{aligned}
$$

However, Lemma 1 yields

$$
\begin{aligned}
\mathbb{E}[(x_n(t)-w_n)^2] = {} & [(x_0-w_n)^2 + \frac{\langle \Sigma, \phi_{2,n}\rangle e^{\phi_3 T}}{-2(\mu-r)^\top \phi_{1,n} + \langle \Sigma, \phi_{1,n}\phi_{1,n}^\top\rangle + \phi_3}] e^{(-2(\mu-r)^\top \phi_{1,n} + \langle \Sigma, \phi_{1,n}\phi_{1,n}^\top\rangle)t} \\
& - \frac{\langle \Sigma, \phi_{2,n}\rangle e^{\phi_3 (T-t)}}{-2(\mu-r)^\top \phi_{1,n} + \langle \Sigma, \phi_{1,n}\phi_{1,n}^\top\rangle + \phi_3}.
\end{aligned}
$$

Integrating $\mathrm{d}\mathbb{E}[Z_{1,n}(t)]$ from 0 to T and plugging in the above expression of $\mathbb{E}(x_n(t)-w_n)^2$, we obtain

$$
h_1(\phi_{1,n}, \phi_{2,n}, w_n) = -R(\phi_{1,n}, \phi_{2,n}, w_n)(\mu - r - \Sigma\phi_{1,n}), \tag{51}
$$

where the function R is defined by

$$
R(\phi_1, \phi_2, w) = 2\left[\frac{(x_0-w)^2 e^{-\phi_3 T}(e^{Q(\phi_1)T}-1)}{Q(\phi_1)} + \frac{\langle \sigma\sigma^\top, \phi_2\rangle (e^{Q(\phi_1)T} - 1 - Q(\phi_1)T)}{Q(\phi_1)^2}\right], \tag{52}
$$

while

$$
Q(\phi_1) = -2(\mu-r)^\top \phi_1 + \langle \sigma\sigma^\top, \phi_1\phi_1^\top\rangle + \phi_3. \tag{53}
$$

Similarly (and more easily), we have

$$
h_2(\phi_{1,n}, \phi_{2,n}, w_n) = \left(\phi_{2,n}\Sigma\phi_{2,n} - \frac{\gamma}{2}\phi_{2,n}\right)T, \tag{54}
$$

which is quadratic in $\phi_{2,n}$. Moreover, Lemma 1 implies

$$
h_w(\phi_{1,n}, \phi_{2,n}, w_n) = \left(1 - e^{-(\mu-r)^\top \phi_{1,n} T}\right) w_n + (x_0 e^{-(\mu-r)^\top \phi_{1,n} T} - z), \tag{55}
$$

which is linear in w_n.

Properties of mean increments

With the explicit expressions of h_1, h_2 and h_w in (51), (54) and (55) respectively, we further investigate properties of these functions, which will be useful in the sequel. Recall that the function R defined in (52) depends on ϕ_3. We first show that this function has a positive lower bound when ϕ_3 is sufficiently large. Indeed, noting that $\sigma\sigma^\top$ is positive definite we have

$$
\begin{aligned}
Q(\phi_1) &= [\phi_1 - (\sigma\sigma^\top)^{-1}(\mu-r)]^\top (\sigma\sigma^\top)[\phi_1 - (\sigma\sigma^\top)^{-1}(\mu-r)] + \phi_3 - (\mu-r)^\top(\sigma\sigma^\top)^{-1}(\mu-r) \\
&> \phi_3 - (\mu-r)^\top(\sigma\sigma^\top)^{-1}(\mu-r) =: C_Q > 0
\end{aligned} \tag{56}
$$

when ϕ_3 is sufficiently large. Hence,

$$\begin{aligned} R(\phi_1,\phi_2,w) &= 2\left[\frac{(x_0-w)^2e^{-\phi_3 T}(e^{Q(\phi_1)T}-1)}{Q(\phi_1)}+\frac{\langle\sigma\sigma^\top,\phi_2\rangle(e^{Q(\phi_1)T}-1-Q(\phi_1)T)}{Q(\phi_1)^2}\right] \\ &\geqslant 2[(x_0-w)^2e^{-\phi_3 T}T+\frac{1}{2}\langle\sigma\sigma^\top,\phi_2\rangle T^2] =: C_R>0, \end{aligned} \tag{57}$$

where the inequality follows from the familiar general result $e^x-1-x-\frac{1}{2}x^2\geqslant 0\ \forall x\geqslant 0$.

On the other hand, Q is a quadratic function in ϕ_1; hence there exist constants $C_{Q_0},C_{Q_1}>0$ such that $Q(\phi_1)\leqslant C_{Q_0}+C_{Q_1}|\phi_1|^2$. As a consequence,

$$\begin{aligned} R(\phi_1,\phi_2,w) &\leqslant 2\left[\frac{C(1+|w|^2)e^{C_{Q_0}+C_{Q_1}|\phi_1|^2}}{C_Q}+\frac{C|\phi_2|e^{C_{Q_0}+C_{Q_1}|\phi_1|^2}}{C_Q^2}\right] \\ &\leqslant C_{R_0}(1+|w|^2+|\phi_2|)\exp\left(C_{Q_0}+C_{Q_1}|\phi_1|^2\right), \end{aligned}$$

where $C_{R_0}>0$ is some constant.

Next, we derive the upper bounds for h_1, h_2 and h_w. We have

$$\begin{aligned} |h_1(\phi_{1,n},\phi_{2,n},w_n)| =&R(\phi_{1,n},\phi_{2,n},w_n)|\mu-r-\Sigma\phi_{1,n}| \\ \leqslant&\left(C_{R_0}(1+|w_n|^2+|\phi_{2,n}|)\exp\left(C_{Q_0}+C_{Q_1}|\phi_{1,n}|^2\right)\right)|\mu-r-\Sigma\phi_{1,n}| \\ \leqslant&C\left(1+|\phi_{1,n}|+|\phi_{1,n}||w_n|^2e^{C|\phi_{1,n}|^2}+|\phi_{1,n}||\phi_{2,n}|e^{|\phi_{1,n}|^2}\right), \end{aligned} \tag{58}$$

and

$$|h_2(\phi_{1,n},\phi_{2,n},w_n)| =T\left|\phi_{2,n}\Sigma\phi_{2,n}-\frac{\gamma}{2}\phi_{2,n}\right|\leqslant C(1+|\phi_{2,n}|^2), \tag{59}$$

where the constant C only depends on Σ, γ and ϕ_3. Denoting $\boldsymbol{h}(\phi_1,\phi_2,w;\phi_3)=(h_1(\phi_{1,n},\phi_{2,n},w_n;\phi_3),h_2(\phi_1,\phi_2,w;\phi_3))^\top$, we conclude by (58) and (59) that

$$\begin{aligned} |\boldsymbol{h}(\phi_{1,n},\phi_{2,n},w_n)|^2 \leqslant&|h_1(\phi_{1,n},\phi_{2,n},w_n)|^2+|h_2(\phi_{1,n},\phi_{2,n},w_n)|^2 \\ \leqslant&\left(C(1+|\phi_{1,n}|+|\phi_{1,n}||w_n|^2e^{C|\phi_{1,n}|^2}+|\phi_{1,n}||\phi_{2,n}|e^{|\phi_{1,n}|^2})\right)^2+\left(C(1+|\phi_{2,n}|^2)\right)^2 \\ \leqslant&C\left(1+|\phi_{1,n}|^2+|\phi_{2,n}|^4+|\phi_{1,n}|^2|w_n|^4e^{C|\phi_{1,n}|^2}+|\phi_{1,n}|^2|\phi_{2,n}|^2e^{C|\phi_{1,n}|^2}\right). \end{aligned} \tag{60}$$

Furthermore, it follows from (55) that

$$|h_w(\phi_{1,n},\phi_{2,n},w_n)|^2\leqslant C(1+e^{C|\phi_{1,n}|})|w_n|. \tag{61}$$

Almost sure convergence of $\boldsymbol{\phi}_n$

We now prove the almost sure convergence of $\boldsymbol{\phi}_n$. Indeed, we present a more general result of such convergence, of which Theorem 1-(a) is a special case.

Theorem 4. *Let ϕ_3 be a sufficiently large constant, while $\boldsymbol{\phi}_n = (\phi_{1,n}, \phi_{2,n})^\top$ and w be updated according to (42). Assume that the noise vector $\xi_n = (\xi_{1,n}, \xi_{2,n})^\top$ satisfies $\mathbb{E}\left[\xi_{i,n+1}\middle|\mathcal{G}_n\right] = \beta_{i,n}$ for $i = 1,2$ and $\mathbb{E}\left[\xi_{w,n+1}\middle|\mathcal{G}_n\right] = \beta_{w,n}$, where $\mathcal{G}_n$ are the filtration generated by $\{\boldsymbol{\theta}_m, \boldsymbol{\phi}_m, w_m, m = 0,1,2,...,n\}$, with the following upper bounds:*

$$\begin{aligned}
\mathbb{E}\left[|\xi_{1,n+1} - \beta_{1,n}|^2 \middle|\mathcal{G}_n\right] &\leqslant C\bigg(1 + |w_n|^{16} + |\phi_{1,n}|^8 + |\phi_{2,n}|^8 + |b_n|^8\bigg)e^{C|\phi_{1,n}|^8},\\
\mathbb{E}\left[|\xi_{2,n+1} - \beta_{2,n}|^2 \middle|\mathcal{G}_n\right] &\leqslant C\bigg(1 + |w_n|^{16} + |\phi_{1,n}|^8 + |\phi_{2,n}|^8\bigg)e^{C|\phi_{1,n}|^8},
\end{aligned} \tag{62}$$

where $C > 0$ is a constant independent of n. Moreover, assume

$$\begin{aligned}
&(i) \quad \sum_n a_n = \infty, \quad \sum_n a_n|\beta_{i,n}| < \infty, \quad \text{for } i = 1,2;\\
&(ii) \quad c_{1,n} \uparrow \infty, \quad c_{2,n} \uparrow \infty, \quad c_{w,n} \uparrow \infty, \quad \sum_n a_n^2 b_n^8 c_{2,n}^8 c_{w,n}^{16} e^{c_{1,n}^8} < \infty;\\
&(iii) \quad b_n \uparrow \infty, \quad \sum_n \frac{a_n}{b_n} = \infty.
\end{aligned} \tag{63}$$

Then $\boldsymbol{\phi}_n = (\phi_{1,n}, \phi_{2,n})^\top$ almost surely converges to the unique equilibrium point $\boldsymbol{\phi}^* = (\phi_1^*, \phi_2^*)^\top$ where $\phi_1^* = \Sigma^{-1}(\mu - r)$ and $\phi_2^* = \frac{\gamma}{2}\Sigma^{-1}$.

Proof. The main idea is to derive inductive upper bound of $|\boldsymbol{\phi}_n - \boldsymbol{\phi}^*|^2$, namely, to bound $|\boldsymbol{\phi}_{n+1} - \boldsymbol{\phi}^*|^2$ in terms of $|\boldsymbol{\phi}_n - \boldsymbol{\phi}^*|^2$.

First, for any closed, convex set $K \subset \mathbb{S}_+^d$ and $x \in K, y \in \mathbb{S}^d$, it follows from a property of projection that the function $f(t) = |t\Pi_K(y) + (1-t)x - y|^2$, $t \in \mathbb{R}$, achieves minimum at $t = 1$. However,

$$f(t) = t^2|\Pi_K(y) - y|^2 + (1-t)^2|x-y|^2 + 2t(1-t)\langle \Pi_K(y) - y, x - y\rangle.$$

The first-order condition at $t = 1$ yields

$$2|\Pi_K(y) - y|^2 - 2\langle \Pi_K(y) - y, x - y\rangle = 0.$$

Therefore,

$$|\Pi_K(y) - x|^2 = |\Pi_K(y) - y + y - x|^2 = |y-x|^2 + |\Pi_K(y) - y|^2 + 2\langle \Pi_K(y) - y, y - x\rangle = |y-x|^2 - |\Pi_K(y) - y|^2 \leqslant |y-x|^2.$$

Now, consider n sufficiently large such that $\boldsymbol{\phi}^* \in K_{1,n+1} \times K_{2,n+1}$ and denote

$$\boldsymbol{h}(\phi_1, \phi_2, w) = (h_1(\phi_1, \phi_2, w), h_2(\phi_1, \phi_2, w))^\top.$$

By the above general projection inequality, we have

$$|\boldsymbol{\phi}_{n+1} - \boldsymbol{\phi}^*|^2 \leqslant \left|\boldsymbol{\phi}_n - a_n[\boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n) + \xi_{n+1}] - \boldsymbol{\phi}^*\right|^2.$$

Denoting $U_n = \boldsymbol{\phi}_n - \boldsymbol{\phi}^*$ and $\boldsymbol{\beta}_n = (\beta_{1,n}, \beta_{2,n})^\top$, we have

$$\begin{aligned}
&\mathbb{E}\left[|U_{n+1}|^2\middle|\boldsymbol{\phi}_n, w_n\right] \\
\leqslant&\mathbb{E}\left[|U_n - a_n[\boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n) + \xi_{n+1}]|^2\middle|\boldsymbol{\phi}_n, w_n\right] \\
=&|U_n|^2 - 2a_n\langle U_n, \boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n) + \boldsymbol{\beta}_n\rangle + a_n^2\mathbb{E}\left[|\boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n) + \xi_{n+1}|^2\middle|\boldsymbol{\phi}_n, w_n\right] \\
=&|U_n|^2 - 2a_n\langle U_n, \boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n) + \boldsymbol{\beta}_n\rangle + a_n^2\mathbb{E}\left[|\boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n) + (\xi_{n+1} - \boldsymbol{\beta}_n) + \boldsymbol{\beta}_n|^2\middle|\boldsymbol{\phi}_n, w_n\right] \\
\leqslant&|U_n|^2 - 2a_n\langle U_n, \boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n))\rangle + 2a_n|\boldsymbol{\beta}_n||U_n| \\
&+ 3a_n^2\left(|\boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n)|^2 + |\boldsymbol{\beta}_n|^2 + \mathbb{E}\left[|\xi_{n+1} - \boldsymbol{\beta}_n|^2\middle|\boldsymbol{\phi}_n, w_n\right]\right) \\
\leqslant&|U_n|^2 - 2a_n\langle U_n, \boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n))\rangle + a_n|\boldsymbol{\beta}_n|(1 + |U_n|^2) \\
&+ 3a_n^2\left(|\boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n)|^2 + |\boldsymbol{\beta}_n|^2 + \mathbb{E}\left[|\xi_{n+1} - \boldsymbol{\beta}_n|^2\middle|\boldsymbol{\phi}_n, w_n\right]\right).
\end{aligned}$$

Recall that $|\phi_{1,n}| \leqslant c_{1,n}$, $|\phi_{2,n}| \leqslant c_{2,n}$, $|w_n| \leqslant c_{w,n}$ almost surely. By the estimate (60),

$$|\boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n)|^2 \leqslant C(1 + c_{1,n}^2 + c_{2,n}^4 + c_{1,n}^2 c_{w,n}^4 e^{Cc_{1,n}^2} + c_{1,n}^2 c_{2,n}^2 e^{Cc_{1,n}^2}). \tag{64}$$

However, the assumption (62) yields

$$\begin{aligned}
\mathbb{E}\left[|\xi_{n+1} - \boldsymbol{\beta}_n|^2\middle|\boldsymbol{\phi}_n, w_n\right] \leqslant&\mathbb{E}\left[|\xi_{1,n+1} - \beta_{1,n}|^2\middle|\boldsymbol{\phi}_n, w_n\right] + \mathbb{E}\left[|\xi_{2,n+1} - \beta_{2,n}|^2\middle|\boldsymbol{\phi}_n, w_n\right] \\
\leqslant&C\Big(1 + (1 + |w_n|^4 + |\phi_{2,n}|^2)\exp\left(C|\phi_{1,n}|^4\right)(b_n^2 + |\phi_{2,n}|^2 + |\phi_{1,n}|^4) \\
&+ (1 + |w_n|^8 + |\phi_{2,n}|^4)\exp(C|\phi_{1,n}|^8)(1 + b_n^4 + |\phi_{1,n}|^4) + (d\log b_n)^4\Big) \\
&+ C\left(1 + |\phi_{2,n}|^4 + (1 + |w_n|^8 + |\phi_{2,n}|^4)\exp(C|\phi_{1,n}|^8)(1 + |\phi_{1,n}|^4)\right) \\
\leqslant&C\Big(1 + |\phi_{2,n}|^4 + (d\log b_n)^4 + \exp(C|\phi_{1,n}|^4)(1 + |w_n|^8| + |\phi_{2,n}|^4 + b_n^4 + |\phi_{1,n}|^8) \\
&+ \exp(C|\phi_{1,n}|^8)(1 + |w_n|^{16}| + |\phi_{2,n}|^8 + b_n^8 + |\phi_{1,n}|^8)\Big) \\
\leqslant&C\left(1 + c_{2,n}^4 + (d\log b_n)^4 + \exp(Cc_{1,n}^8)(1 + c_{w,n}^{16} + c_{2,n}^8 + b_n^8 + c_{1,n}^8)\right)
\end{aligned}$$

almost surely, for some positive constant C that only depends on the model primitives μ, Σ, d.

Therefore,

$$
\begin{aligned}
&\mathbb{E}\left[|U_{n+1}|^2 \middle| \boldsymbol{\phi}_n, w_n\right] \\
\leqslant &|U_n|^2 - 2a_n\langle U_n, \boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n)\rangle + a_n|\boldsymbol{\beta}_n||U_n|^2 + a_n|\boldsymbol{\beta}_n| \\
&+ 3a_n^2\left(|\boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n)|^2 + |\boldsymbol{\beta}_n|^2 + \mathbb{E}\left[|\xi_{n+1} - \boldsymbol{\beta}_n|^2 \middle| \boldsymbol{\phi}_n, w_n\right]\right) \\
\leqslant &|U_n|^2 - 2a_n\langle U_n, \boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n)\rangle + a_n|\boldsymbol{\beta}_n||U_n|^2 + a_n|\boldsymbol{\beta}_n| \\
&+ 3a_n^2\bigg(C(1 + c_{1,n}^2 + c_{2,n}^4 + c_{1,n}^2 c_{w,n}^4 e^{Cc_{1,n}^2} + c_{1,n}^2 c_{2,n}^2 e^{Cc_{1,n}^2}) + |\boldsymbol{\beta}_n|^2 \\
&+ C(1 + c_{2,n}^4 + (d\log b_n)^4 + e^{Cc_{1,n}^4}(1 + c_{w,n}^8 + c_{2,n}^4 + b_n^4 + c_{1,n}^8) + e^{Cc_{1,n}^8}(1 + c_{w,n}^{16} + c_{2,n}^8 + b_n^8 + c_{1,n}^8))\bigg) \\
= &(1 + a_n|\boldsymbol{\beta}_n|)|U_n|^2 - 2a_n\langle U_n, \boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n)\rangle + a_n|\boldsymbol{\beta}_n| + \\
&+ 3a_n^2\bigg(C(1 + c_{1,n}^2 + c_{2,n}^4 + c_{1,n}^2 c_{w,n}^4 e^{Cc_{1,n}^2} + c_{1,n}^2 c_{2,n}^2 e^{Cc_{1,n}^2}) + |\boldsymbol{\beta}_n|^2 \\
&+ C(1 + (d\log b_n)^4 + e^{Cc_{1,n}^4}(1 + c_{w,n}^8 + c_{2,n}^4 + b_n^4 + c_{1,n}^8) + e^{Cc_{1,n}^8}(1 + c_{w,n}^{16} + c_{2,n}^8 + b_n^8 + c_{1,n}^8))\bigg) \\
=: &(1 + \gamma_n)|U_n|^2 - \zeta_n + \eta_n,
\end{aligned}
\tag{65}
$$

where $\gamma_n = a_n|\beta_n|$, $\zeta_n = 2a_n\langle U_n, h(\phi_{1,n}, \phi_{2,n}, w_n)\rangle$, and

$$
\begin{aligned}
\eta_n = &a_n|\boldsymbol{\beta}_n| + 3a_n^2|\boldsymbol{\beta}_n|^2 + 3a_n^2\bigg(C(1 + c_{1,n}^2 + c_{2,n}^4 + c_{1,n}^2 c_{w,n}^4 e^{Cc_{1,n}^2} + c_{1,n}^2 c_{2,n}^2 e^{Cc_{1,n}^2}) + |\boldsymbol{\beta}_n|^2 \\
&+ C(1 + (d\log b_n)^4 + e^{Cc_{1,n}^4}(1 + c_{w,n}^8 + c_{2,n}^4 + b_n^4 + c_{1,n}^8) + e^{Cc_{1,n}^8}(1 + c_{w,n}^{16} + c_{2,n}^8 + b_n^8 + c_{1,n}^8))\bigg).
\end{aligned}
\tag{66}
$$

By Assumptions (i)–(ii), we know $\sum_n \gamma_n < \infty$ and $\sum_n \eta_n < \infty$. It then follows from Robbins and Siegmund (1971, Theorem 1) that $|U_n|^2$ converges to a finite limit and $\sum_n \zeta_n < \infty$ almost surely.

It remains to show $|U_n| \to 0$ almost surely. Consider the term

$$
\begin{aligned}
&\langle \boldsymbol{\phi} - \boldsymbol{\phi}^*, \boldsymbol{h}(\phi_1, \phi_2, w)\rangle \\
= &\langle \phi_1 - \phi_1^*, h_1(\phi_1, \phi_2, w)\rangle + \langle \phi_2 - \phi_2^*, h_2(\phi_2)\rangle \\
= &\langle \phi_1 - \phi_1^*, R(\phi_1, \phi_2, w)\Sigma(\phi_1 - \phi_1^*)\rangle + \langle \phi_2 - \phi_2^*, \phi_{2,n}^\top \Sigma(\phi_2 - \phi_2^*)\rangle \\
= &R(\phi_1, \phi_2, w)\langle \Sigma, (\phi_1 - \phi_1^*)(\phi_1 - \phi_1^*)^\top\rangle + \langle \Sigma\phi_2, (\phi_2 - \phi_2^*)(\phi_2 - \phi_2^*)^\top\rangle.
\end{aligned}
$$

Note that $\langle \Sigma, (\phi_1 - \phi_1^*)(\phi_1 - \phi_1^*)^\top\rangle \geqslant 0$ because $\Sigma \in \mathbb{S}_{++}^d$ and $(\phi_1 - \phi_1^*)(\phi_1 - \phi_1^*)^\top \in \mathbb{S}_+^d$.

To proceed, let us first consider a spacial case when $\Sigma = I$ to get the main idea of the rest of the proof. Indeed, when $\Sigma = I$,

$$
\langle \Sigma, (\phi_1 - \phi_1^*)(\phi_1 - \phi_1^*)^\top\rangle = \langle I, (\phi_1 - \phi_1^*)(\phi_1 - \phi_1^*)^\top\rangle \geqslant |\phi_1 - \phi_1^*|^2 \geqslant \delta^2,
$$

whenever $|\phi_1 - \phi_1^*| \geqslant \delta > 0$. In this case,

$$R(\phi_1, \phi_2, w)\langle I, (\phi_1 - \phi_1^*)(\phi_1 - \phi_1^*)^\top\rangle \geqslant C_R\delta^2$$

because $R(\phi_1, \phi_2, w; \phi_3) \geqslant C_R > 0$ due to (57). Moreover, $\langle \phi_2, (\phi_2 - \phi_2^*)(\phi_2 - \phi_2^*)^\top\rangle \geqslant 0$ because $\phi_2 \in \mathbb{S}_+^d$ and $(\phi_2 - \phi_2^*)(\phi_2 - \phi_2^*)^\top \in \mathbb{S}_+^d$. In particular, when $|\phi_2 - \phi_2^*| \geqslant \delta > 0$ and $\phi_2 - \frac{1}{b_n}I \in \mathbb{S}_+^d$, we have

$$\begin{aligned}\langle \phi_2, (\phi_2 - \phi_2^*)(\phi_2 - \phi_2^*)^\top\rangle =& \langle \phi_2 - \frac{1}{b_n}I, (\phi_2 - \phi_2^*)(\phi_2 - \phi_2^*)^\top\rangle + \frac{1}{b_n}\langle I, (\phi_2 - \phi_2^*)(\phi_2 - \phi_2^*)^\top\rangle \\ \geqslant& \frac{1}{b_n}|\phi_2 - \phi_2^*|^2 \geqslant \frac{\delta^2}{b_n}.\end{aligned}$$

Now, suppose $|U_n| \nrightarrow 0$ almost surely. Then there exists a set $Z \in \mathcal{F}$ with $\mathbb{P}(Z) = 1$ so that for every $\omega \in Z$, there is $\delta(\omega) > 0$ such that for all n sufficiently large, at least one of the following two cases holds true: (a) $|\phi_1(\omega) - \phi_1^*| \geqslant \delta(\omega) > 0$; (b) $|\phi_2(\omega) - \phi_2^*| \geqslant \delta(\omega) > 0$.

Recall that $\boldsymbol{\phi}_n(\omega) \in K_{1,n} \times K_{2,n}$. If (a) is true, then the above analysis yields

$$\langle U_n(\omega), \boldsymbol{h}(\boldsymbol{\phi}_n(\omega), w_n(\omega))\rangle \geqslant \delta(\omega)^2.$$

Thus, by Assumption (iii), we have

$$\sum_n \zeta_n(\omega) = 2\sum_n a_n\langle U_n(\omega), \boldsymbol{h}(\boldsymbol{\phi}_n(\omega), w_n(\omega))\rangle \geqslant 2C_R\delta(\omega)^2\sum_n a_n = \infty.$$

This is a contradiction.

If (b) is true, then

$$\langle U_n(\omega), \boldsymbol{h}(\boldsymbol{\phi}_n(\omega), w_n(\omega))\rangle \geqslant \frac{\delta(\omega)^2}{b_n},$$

and hence, by Assumption-(iii),

$$\sum_n \zeta_n(\omega) = 2\sum_n a_n\langle U_n(\omega), \boldsymbol{h}(\boldsymbol{\phi}_n(\omega), w_n(\omega))\rangle \geqslant 2\delta(\omega)^2\sum_n \frac{a_n}{b_n} = \infty.$$

This is again a contradiction.

Now let us consider the general case when $\Sigma \neq I$. Introduce a different inner product and norm on $\mathbb{R}^d \times \mathbb{R}^{d\times d}$ induced by $\Sigma \in \mathbb{S}_{++}^d$:

$$\langle (A_1, A_2)^\top, (B_1, B_2)^\top\rangle_\Sigma := \langle A_1, B_1\rangle + \langle \Sigma A_2\Sigma, B_2\rangle,$$

$$|(A_1, A_2)^\top|_\Sigma := |A_1| + \sqrt{\langle A, A\rangle_\Sigma} = |A_1| + |\Sigma^{1/2}A_2\Sigma^{1/2}|.$$

It is straightforward to verify that $\langle\cdot,\cdot\rangle_\Sigma$ is indeed an inner product and $|\cdot|_\Sigma$ is the associated norm. Moreover, since all norms on a finite dimensional space are equivalent, there exist constants $\overline{C} > \underline{C} > 0$ depending only on Σ and the dimension d such that

$$\underline{C}|(A_1, A_2)^\top| \leqslant |(A_1, A_2)^\top|_\Sigma \leqslant \overline{C}|(A_1, A_2)^\top|,$$

for any $A_1 \in \mathbb{R}^d$, $A_2 \in \mathbb{R}^{d\times d}$.

When n is sufficiently large such that $\boldsymbol{\phi}^* \in K_{n+1}$,

$$|\boldsymbol{\phi}_{n+1} - \boldsymbol{\phi}^*|^2 \leqslant |\boldsymbol{\phi}_n - a_n[\boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n) + \xi_{n+1}] - \boldsymbol{\phi}^*|^2,$$

or $|\boldsymbol{\phi}_{n+1} - \boldsymbol{\phi}^*|_\Sigma^2 \leqslant \frac{\overline{C}^2}{\underline{C}^2}|\boldsymbol{\phi}_n - a_n[\boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n) + \xi_{n+1}] - \boldsymbol{\phi}^*|_\Sigma^2$. Hence the estimate (65) for U_{n+1} still holds true under the new norm $|\cdot|_\Sigma$. It follows that

$$\sum a_n \langle U_n, \boldsymbol{h}(\phi_{1,n}, \phi_{2,n}, w_n)\rangle_\Sigma < \infty$$

and $|U_n|_\Sigma^2$ converges to a finite limit almost surely.

Consider the term

$$\begin{aligned}
&\langle \boldsymbol{\phi} - \boldsymbol{\phi}^*, \boldsymbol{h}(\phi_1, \phi_2, w)\rangle_\Sigma \\
=&R(\phi_1, \phi_2, w)\langle \Sigma, (\phi_1 - \phi_1^*)(\phi_1 - \phi_1^*)^\top\rangle + \langle \Sigma(\phi_2 - \phi_2^*)\Sigma, \phi_{2,n}^\top \Sigma(\phi_2 - \phi_2^*)\rangle \\
=&R(\phi_1, \phi_2, w)\langle \Sigma, (\phi_1 - \phi_1^*)(\phi_1 - \phi_1^*)^\top\rangle + \langle \Sigma^{1/2}(\phi_2 - \phi_2^*)\Sigma^{1/2}, \Sigma^{1/2}\phi_{2,n}^\top \Sigma^{1/2}\Sigma^{1/2}(\phi_2 - \phi_2^*)\Sigma^{1/2}\rangle \\
=&R(\phi_1, \phi_2, w)\langle \Sigma, (\phi_1 - \phi_1^*)(\phi_1 - \phi_1^*)^\top\rangle + \langle \tilde{\phi_2} - \tilde{\phi_2^*}, \tilde{\phi_{2,n}^\top}(\tilde{\phi_2} - \tilde{\phi_2^*})\rangle \\
=&R(\phi_1, \phi_2, w)\langle \Sigma, (\phi_1 - \phi_1^*)(\phi_1 - \phi_1^*)^\top\rangle + \langle \tilde{\phi_2}, (\tilde{\phi_2} - \tilde{\phi_2^*})(\tilde{\phi_2} - \tilde{\phi^{(2)*\top}})\rangle,
\end{aligned}$$

where $\tilde{\phi_2} = \Sigma^{1/2}\phi_2\Sigma^{1/2}$ and $\tilde{\phi_2^*} = \Sigma^{1/2}\phi_2^*\Sigma^{1/2}$. As before, we need to prove $|U_n|_\Sigma \to 0$ almost surely. If not, then there exists a set $Z \in \mathcal{F}$ with $\mathbb{P}(Z) = 1$ so that for every $\omega \in Z$, there is $\delta(\omega) > 0$ such that for all n sufficiently large, at least one of the following two cases are true: (a) $|\phi_1(\omega) - \phi_1^*| \geqslant \delta(\omega) > 0$; (b) $|\phi_2(\omega) - \phi_2^*|_\Sigma \geqslant \delta(\omega) > 0$.

If (a) is true, then there is a contradiction based on the same argument before. If (b) is true, then when

$|\phi_2(\omega) - \phi_2^*|_\Sigma \geqslant \delta(\omega) > 0$ and $\phi_2(\omega) - \frac{1}{b_n} I \in \mathbb{S}_+^d$, we have $\Sigma^{1/2}(\phi_2(\omega) - \frac{1}{b_n} I)\Sigma^{1/2} \in \mathbb{S}_+^d$, and

$$\begin{aligned}
\langle \boldsymbol{\phi} - \boldsymbol{\phi}^*, \boldsymbol{h}(\phi_1, \phi_2, w) \rangle_\Sigma \geqslant & \langle \Sigma^{1/2}\phi_2\Sigma^{1/2}, (\tilde{\phi_2} - \tilde{\phi_2^*})(\tilde{\phi_2} - \tilde{\phi_2^*})^\top \rangle \\
= & \langle \Sigma^{1/2}(\phi_2 - \frac{1}{b_n} I)\Sigma^{1/2}, (\tilde{\phi_2} - \tilde{\phi_2^*})(\tilde{\phi_2} - \tilde{\phi_2^*})^\top \rangle + \frac{1}{b_n}\langle \Sigma, (\tilde{\phi_2} - \tilde{\phi_2^*})(\tilde{\phi_2} - \tilde{\phi_2^{*\top}}) \rangle \\
\geqslant & \frac{1}{b_n}\langle \Sigma, (\tilde{\phi_2} - \tilde{\phi_2^*})(\tilde{\phi_2} - \tilde{\phi^{(2)*\top}}) \rangle \\
\geqslant & \frac{\lambda_{min}}{b_n}\langle I, (\tilde{\phi_2} - \tilde{\phi_2^*})(\tilde{\phi_2^*} - \tilde{\phi_2^*})^\top \rangle \\
= & \frac{\lambda_{min}}{b_n}|\tilde{\phi_2} - \tilde{\phi_2^*}|^2 = \frac{\lambda_{min}}{b_n}|\Sigma^{1/2}(\phi_2 - \phi_2^*)\Sigma^{1/2}|^2 \\
\geqslant & \frac{\lambda_{min}\delta^2}{b_n},
\end{aligned}$$

where $\lambda_{min} > 0$ is the smallest eigenvalue of Σ. Hence

$$\langle U_n(\omega), \boldsymbol{h}(\boldsymbol{\phi}_n(\omega), w_n(\omega)) \rangle_\Sigma \geqslant \frac{\lambda_{min}\delta^2}{b_n}.$$

Thus, Assumption-(iii) implies

$$\sum_n a_n \langle U_n(\omega), \boldsymbol{h}(\boldsymbol{\phi}_n(\omega), w_n(\omega)) \rangle \geqslant \lambda_{min}\delta(\omega)^2 \sum_n \frac{a_n}{b_n} = \infty,$$

which is a contradiction. The proof is now complete. □

Remark 1. *When $\beta_{1,n} = 0$, $\beta_{2,n} = 0$, $\beta_{w,n} = 0$ for all n, which holds true in our mean–variance problem, a typical choice of the sequences satisfying Assumptions (i)–(iii) is $a_n = \frac{\alpha}{n+\beta}$ with constants $\alpha > 0$ and $\beta > 0$, $b_n = 1 \vee (\log\log n)^{\frac{1}{8}}, c_{1,n} = 1 \vee (\log\log n)^{\frac{1}{8}}, c_{2,n} = 1 \vee (\log\log n)^{\frac{1}{8}}$ and $c_{w,n} = 1 \vee (\log\log n)^{\frac{1}{16}}$. This is because $\sum \frac{1}{n(\log\log n)^\kappa} = \infty$ and $\sum \frac{(\log n)^{\kappa_1}(\log\log n)^{\kappa_2}}{n^2} < \infty$, for any $\kappa, \kappa_1, \kappa_2 > 0$.*

Mean–squared error of $\phi_{1,n} - \phi_1^*$

Now we move forward to derive the error bound of $\phi_{1,n} - \phi_1^*$ in the mean-squared sense, which is Theorem 1-(b). Note that this result is also necessary for subsequently proving the almost sure convergence of w_n, because h_w not only depends on w, but also on ϕ_1. Moreover, the error bound of $\phi_{1,n} - \phi_1^*$ affects the property of h_w.

We first need a general recursive relation satisfied by a typical learning rate sequence.

Lemma 3. *For any $A > 0$, there exist positive numbers $\alpha > \frac{1}{A}$ and $\beta \geqslant \frac{1}{A\alpha - 1}$ such that the learning rate sequence $a_n = \frac{\alpha}{n+\beta}$, $n \geqslant 0$, satisfies $a_n \leqslant a_{n+1}(1 + Aa_{n+1})$ for any $n \geqslant 0$.*

Proof. It is clear that $a_n \leqslant a_{n+1}(1 + Aa_{n+1})$ is equivalent to $n + 1 + \beta \leqslant A\alpha n + A\alpha\beta$. However, the latter holds true when $\alpha > \frac{1}{A}$, $\beta \geqslant \frac{1}{A\alpha - 1}$. □

With Lemma 3, we present the following result for the mean-squared error of $\phi_{1,n}$.

Theorem 5. *Under the assumptions of Theorem 4, if the sequence $\{a_n\}$ further satisfies*

$$a_n \leqslant a_{n+1}(1 + Aa_{n+1}),$$

for some sufficiently small constant $A > 0$ and $|\beta_n| = O(a_n^{\frac{1}{2}})$, then there exists an increasing sequence $\{\hat{\eta}_n\}$ and a constant $C' > 0$ such that

$$\mathbb{E}[|\phi_{1,n+1} - \phi_1^*|^2] \leqslant C' a_n \hat{\eta}_{1,n}.$$

In particular, if we set the parameters $a_n, b_n, c_{1,n}, \beta_{1,n}$ as in Remark 1, then

$$\mathbb{E}[|\phi_{1,n+1} - \phi_1^*|^2] \leqslant C \frac{(\log n)^p (\log \log n)}{n}$$

for any n, where C and p are positive constants that only depend on model primitives.

Proof. Denote $n_0 = \inf\{n \geqslant 0 : \boldsymbol{\phi}^* \in K_{1,n+1} \times K_{2,n+1}\}$ and $U_{1,n} = \phi_{1,n} - \phi_1^*$. It follows from (51) and (57) that

$$\langle U_{1,n}, h_1(\phi_{1,n}, \phi_{2,n}, w_n; \phi_3)\rangle \geqslant C_R' |\phi_{1,n} - \phi_1^*|^2 = C_R' |U_{1,n}|^2$$

with some constant $C_R' > 0$. When $n \geqslant n_0$, this together with a similar argument to the proof of Theorem 4 yields

$$\begin{aligned}
&\mathbb{E}\left[|U_{1,n+1}|^2 \middle| \boldsymbol{\phi}_n, w_n\right] \\
\leqslant & |U_{1,n}|^2 - 2a_n \langle U_{1,n}, h_1(\phi_{1,n}, \phi_{2,n}, w_n; \phi_3)\rangle + 2a_n |\beta_{1,n}||U_{1,n}| + 3a_n^2 \Big(|h_1(\phi_{1,n}, \phi_{2,n}, w_n; \phi_3)|^2 + |\beta_{1,n}|^2 \\
&+ \mathbb{E}\left[|\xi_{1,n+1} - \beta_{1,n}|^2 \middle| \boldsymbol{\phi}_n, w_n\right] \Big) \\
\leqslant & |U_{1,n}|^2 - 2a_n \langle U_{1,n}, h_1(\phi_{1,n}, \phi_{2,n}, w_n; \phi_3)\rangle + a_n \left(\frac{1}{C_R'} |\beta_{1,n}|^2 + C_R' |U_{1,n}|^2 \right) \\
&+ 3a_n^2 \Big(|h_1(\phi_{1,n}, \phi_{2,n}, w_n; \phi_3)|^2 + |\beta_{1,n}|^2 + \mathbb{E}\left[|\xi_{1,n+1} - \beta_{1,n}|^2 \middle| \boldsymbol{\phi}_n, w_n\right] \Big) \\
\leqslant & (1 - a_n C_R') |U_{1,n}|^2 + 3a_n^2 \hat{\eta}_n.
\end{aligned} \tag{67}$$

Now, by the proof of Theorem 4,

$$\begin{aligned}
&|h_1(\phi_{1,n}, \phi_{2,n}, w_n; \phi_3)|^2 + \mathbb{E}\left[|\xi_{1,n+1} - \beta_{1,n}|^2 \middle| \boldsymbol{\theta}_n, \boldsymbol{\phi}_n, w_n\right] \\
\leqslant & C \Big(1 + c_{1,n}^2 + c_{1,n}^2 c_{w,n}^4 e^{C c_{1,n}^2} + c_{1,n}^2 c_{2,n}^2 e^{C c_{1,n}^2} \\
&+ \exp\{C c_{1,n}^4\}(1 + c_{w,n}^8 + c_{2,n}^4 + b_n^4 + c_{2,n}^4 + c_{1,n}^8) \\
&+ \exp\{C c_{1,n}^8\}(1 + c_{w,n}^{16} + c_{2,n}^8 + b_n^8 + c_{1,n}^8) + (d \log b_n)^8 \Big).
\end{aligned} \tag{68}$$

Moreover, the assumption $|\beta_n| = O(a_n^{\frac{1}{2}})$ imply that $\frac{|\beta_n|^2}{a_n} \leqslant c$, where $c > 0$ is a constant. When $n \geqslant n_0$, it follows from (67) that

$$\mathbb{E}\left[|U_{1,n+1}|^2 \middle| \boldsymbol{\phi}_n, w_n\right] \leqslant (1 - a_n C'_R)|U_{1,n}|^2 + 3a_n^2 \hat{\eta}_n,$$

where

$$\begin{aligned}
\hat{\eta}_n =& C\bigg(1 + c_{1,n}^2 + c_{1,n}^2 c_{w,n}^4 e^{Cc_{1,n}^2} + c_{1,n}^2 c_{2,n}^2 e^{Cc_{1,n}^2} \\
&+ \exp\{Cc_{1,n}^4\}(1 + c_{w,n}^8 + c_{2,n}^4 + b_n^4 + c_{2,n}^4 + c_{1,n}^8) \\
&+ \exp\{Cc_{1,n}^8\}(1 + c_{w,n}^{16} + c_{2,n}^8 + b_n^8 + c_{1,n}^8) + (d \log b_n)^8\bigg),
\end{aligned} \tag{69}$$

which is monotonically increasing because so are $c_{1,n}$, $c_{2,n}$, $c_{w,n}$, b_n by the assumptions. Taking expectation on both sides of the above and denoting $\rho_n = \mathbb{E}[|U_{1,n}|^2]$, we get

$$\rho_{n+1} \leqslant (1 - a_n C'_R)\rho_n + 3a_n^2 \hat{\eta}_n, \tag{70}$$

where $n \geqslant n_0$.

Next, we show $\rho_{n+1} \leqslant C' a_n \hat{\eta}_n$ for all $n \geqslant 0$ by induction, where $C' = max\{\frac{\rho_1}{a_0 \hat{\eta}_0}, \frac{\rho_2}{a_1 \hat{\eta}_1}, \cdots, \frac{\rho_{n_0+1}}{a_{n_0} \hat{\eta}_{n_0}}, \frac{3}{C'_R}\} + 1$. Indeed, it is true when $n \leqslant n_0$. Assume that $\rho_{k+1} \leqslant c' a_k \hat{\eta}_{1,k}$ is true for $n_0 \leqslant k \leqslant n-1$. Then (70) yields

$$\begin{aligned}
\rho_{n+1} &\leqslant (1 - a_n C'_R)\rho_n + 3a_n^2 \hat{\eta}_n \\
&\leqslant (1 - a_n C'_R) C' a_{n-1} \hat{\eta}_{n-1} + 3a_n^2 \hat{\eta}_n \\
&\leqslant (1 - a_n C'_R) C' a_n (1 + A a_n) \hat{\eta}_{n-1} + 3a_n^2 \hat{\eta}_n \\
&\leqslant (1 - a_n C'_R) C' a_n (1 + A a_n) \hat{\eta}_n + 3a_n^2 \hat{\eta}_n \\
&= C' a_n \hat{\eta}_n + C' \hat{\eta}_n a_n^2 \left(-A C'_R a_n + (A - C'_R) + \frac{3}{C'}\right).
\end{aligned}$$

Consider the function

$$f(x) = C' \hat{\eta}_n x^2 \left(-A C'_R x + (A - C'_R) + \frac{3}{C'}\right),$$

which has two roots at $x_{1,2} = 0$ and one root at $x_3 = \frac{A - C'_R + \frac{3}{C'}}{A C'_R}$. Because $C'_R - \frac{3}{C'} > 0$, we can choose $0 < A < C'_R - \frac{3}{C'}$ so that $x_3 < 0$. So $f(x) < 0$ when $x > 0$, leading to

$$C' \hat{\eta}_n a_n^2 \left(-A C'_R a_n + (A - C'_R) + \frac{3}{C'}\right) < 0, \quad \forall n,$$

since $a_n > 0$. We have now proved $\mathbb{E}[|U_{1,n+1}|^2] \leqslant C' a_n \hat{\eta}_n$.

In particular, under the settings of Remark 1, it is straightforward to verify that $|\beta_n| = O(a_n^{\frac{1}{2}})$. Then

$$
\begin{aligned}
\hat{\eta}_n =& C\bigg(1 + c_{1,n}^2 + c_{1,n}^2 c_{w,n}^4 e^{Cc_{1,n}^2} + c_{1,n}^2 c_{2,n}^2 e^{Cc_{1,n}^2} \\
&+ \exp\{Cc_{1,n}^4\}(1 + c_{w,n}^8 + c_{2,n}^4 + b_n^4 + c_{2,n}^4 + c_{1,n}^8) \\
&+ \exp\{Cc_{1,n}^8\}(1 + c_{w,n}^{16} + c_{2,n}^8 + b_n^8 + c_{1,n}^8) + (d \log b_n)^8\bigg) \\
\leqslant& C\bigg(1 + \log\log n + \log\log n(\log n)^p + (\log n)^p(1 + \log\log n)\bigg) \\
\leqslant& C(\log n)^p(\log\log n),
\end{aligned}
\tag{71}
$$

where C and p are positive constants independent of n. The proof is now complete.

$\square$

Almost sure convergence of w_n

We finally prove the almost sure convergence of w_n.

Theorem 6. *Let w_n be updated following* (42)*, and the assumptions* (63) *and the following additional assumptions be satisfied:*

$$
\begin{aligned}
&(i) \quad \sum_n a_{w,n} = \infty, \quad \sum_n a_{w,n}|\beta_{w,n}| < \infty; \\
&(ii) \quad c_{1,n} \uparrow \infty, \quad c_{2,n} \uparrow \infty, \quad c_{w,n} \uparrow \infty, \quad \sum_n a_{w,n}^2 c_{2,n} c_{w,n}^2 e^{c_{1,n}^2} < \infty; \\
&(iii) \quad \sum_n a_{w,n} a_n c_{1,n}^8 c_{2,n}^8 b_n^8 c_{w,n}^{16} e^{c_{1,n}^8} < \infty.
\end{aligned}
\tag{72}
$$

Then $w_n \to w^* = \frac{ze^k - x_0}{e^k - 1}$ almost surely as $n \to \infty$, where $k = (\mu - r)^\top \Sigma^{-1}(\mu - r)T$.

Proof. Recall from Lemma 1 that $\mathbb{E}\left[|\xi_{w,n+1} - \beta_{w,n}|^2 \,\middle|\, \phi_n\right] \leqslant C(1 + |w_n|^2 + |\phi_{2,n}|)e^{C|\phi_{1,n}|^2} \leqslant C(1 + c_{w,n}^2 + c_{2,n})e^{Cc_{1,n}^2}$ and $\beta_{w,n} = 0$ in our case.

Also, we can estimate the upper bound of function $|h_w|$ as

$$
|h_w(\phi_{1,n}, w_n)| \leqslant C(1 + |w_n|)e^{C|\phi_{1,n}|}.
$$

Then,

$$
|h_w(\phi_{1,n}, w_n)|^2 \leqslant C(1 + |w_n|^2)e^{C|\phi_{1,n}|} \leqslant C(1 + c_{w,n}^2)e^{Cc_{1,n}}.
$$

Denote $U_{w,n} = w_n - w^*$. Then similarly as in the proof of Theorem 4, we have

$$
\begin{aligned}
&\mathbb{E}\left[|U_{w,n+1}|^2\Big|\boldsymbol{\phi}_n, w_n\right] \\
\leqslant&|U_{w,n}|^2 - 2a_{w,n}U_{w,n}h_w(\phi_{1,n}, w_n) + a_{w,n}|\beta_{w,n}|(1 + |U_{w,n}|^2) \\
&+ 3a_{w,n}^2\left(|h_w(\phi_{1,n}, w_n)|^2 + |\beta_{w,n}|^2 + \mathbb{E}\left[|\xi_{w,n+1} - \beta_{w,n}|^2\Big|\boldsymbol{\phi}_n, w_n\right]\right) \\
\leqslant&|U_{w,n}|^2 - 2a_{w,n}U_{w,n}h_w(\phi_{1,n}, w_n) + a_{w,n}|\beta_{w,n}|(1 + |U_{w,n}|^2) \\
&+ 3a_{w,n}^2\left(C(1 + c_{w,n}^2)e^{Cc_{1,n}} + |\beta_{w,n}|^2 + C(1 + c_{w,n}^2 + c_{2,n})e^{Cc_{1,n}^2}\right) \\
\leqslant&\left[1 + a_{w,n}|\beta_{w,n}|\right]|U_{w,n}|^2 - 2a_{w,n}U_{w,n}h_w(\phi_{1,n}, w_n) + a_{w,n}|\beta_{w,n}| \\
&+ 3a_{w,n}^2\left(C(1 + c_{w,n}^2)e^{Cc_{1,n}} + |\beta_{w,n}|^2 + C(1 + c_{w,n}^2 + c_{2,n})e^{Cc_{1,n}^2}\right) \\
=&\left[1 + a_{w,n}|\beta_{w,n}| + 4a_{w,n}(1 - e^{-(\mu-r)^\top\phi_{1,n}T})^-\right]|U_{w,n}|^2 \\
&- 2a_{w,n}\left(U_{w,n}h_w(\phi_{1,n}, w_n) + 2(1 - e^{-(\mu-r)^\top\phi_{1,n}T})^-|U_{w,n}|^2 + M(\phi_{1,n})\right) \\
&+ a_{w,n}|\beta_{w,n}| + 3a_{w,n}^2\left(C(1 + c_{w,n}^2)e^{Cc_{1,n}} + |\beta_{w,n}|^2 + C(1 + c_{w,n}^2 + c_{2,n})e^{Cc_{1,n}^2}\right)+2a_{w,n}M(\phi_{1,n}) \\
=&:(1 + \gamma_n)|U_{w,n}|^2 - \zeta_n + \eta_n,
\end{aligned}
\tag{73}
$$

where $f^- = \max(-f, 0)$, $\gamma_n = a_{w,n}|\beta_{w,n}| + 4a_{w,n}(1 - e^{-(\mu-r)^\top\phi_{1,n}T})^-$, $\zeta_n = 2a_{w,n}\Big[U_{w,n}h_w(\phi_{1,n}, w_n) + 2(1 - e^{-(\mu-r)^\top\phi_{1,n}T})^-U_{w,n}^2 + M(\phi_{1,n})\Big]$, $\eta_n = a_{w,n}|\beta_{w,n}| + 3a_{w,n}^2\Big(C(1 + c_{w,n}^2)e^{Cc_{1,n}} + |\beta_{w,n}|^2 + C(1 + c_{w,n}^2 + c_{2,n})e^{Cc_{1,n}^2}\Big)+2a_{w,n}M(\phi_{1,n})$, while

$$
M(\phi_{1,n}) = \frac{(z - x_0)^2}{4(e^k - 1)^2}\frac{(e^{-(\mu-r)^\top(\phi_{1,n}-\phi_1^*)T} - 1)^2}{|1 - e^{-(\mu-r)^\top\phi_{1,n}T}|} \geqslant 0.
$$

First, we consider the term γ_n. By Theorem 4, almost surely, $\phi_{1,n} \to \phi_1^* = \Sigma^{-1}(\mu - r)$ as $n \to \infty$. Hence $(\mu - r)^\top\phi_{1,n} \to (\mu - r)^\top\Sigma^{-1}(\mu - r) > 0$ since $\Sigma \in \mathbb{S}_{++}^d$ and $\mu \neq r$. Then for any $\epsilon > 0$ and any $\omega \in \Omega$ except for a zero-probability set, there exists $0 < N_1(\epsilon, \omega) < \infty$ such that when $n \geqslant N_1(\epsilon, \omega)$, $1 - e^{-(\mu-r)^\top\phi_{1,n}(\omega)T} > \epsilon > 0$. Then,

$$
\begin{aligned}
\sum_{n=1}^{\infty}\gamma_n &= 4\sum_{n=1}^{\infty}a_{w,n}(1 - e^{-(\mu-r)^\top\phi_{1,n}(\omega)T})^- \\
&= 4\sum_{n=1}^{N_1(\epsilon,\omega)}a_{w,n}(1 - e^{-(\mu-r)^\top\phi_{1,n}(\omega)T})^- < \infty.
\end{aligned}
$$

Next, we consider the term $U_{w,n}h_w(\phi_{1,n}, w_n)\rangle + 2\left(1 - e^{-(\mu-r)^\top\phi_{1,n}T}\right)^- U_{w,n}^2$. When $1 - e^{-(\mu-r)^\top\phi_{1,n}T} \geqslant$

0,

$$
\begin{aligned}
&U_{w,n}h_w(\phi_{1,n},w_n)+2\left(1-e^{-(\mu-r)^\top\phi_{1,n}T}\right)^- U_{w,n}^2\\
=&(w_n-w^*)\left[\left(1-e^{-(\mu-r)^\top\phi_{1,n}T}\right)w_n+\left(x_0e^{-(\mu-r)^\top\phi_{1,n}T}-z\right)\right]\\
=&\left(1-e^{-(\mu-r)^\top\phi_{1,n}T}\right)w_n^2+\frac{1}{e^k-1}\left\{\left[e^k(x_0+z)-2x_0\right]e^{-(\mu-r)^\top\phi_{1,n}T}-2ze^k+z+x_0\right\}w_n\\
&-\frac{1}{e^k-1}(ze^k-x_0)\left[x_0e^{-(\mu-r)^\top\phi_{1,n}T}-z\right],
\end{aligned}
$$

which is a convex quadratic function of w_n with the minimum value

$$
-\frac{(z-x_0)^2}{4(e^k-1)^2}\frac{(e^{-(\mu-r)^\top(\phi_{1,n}-\phi_1^*)T}-1)^2}{1-e^{-(\mu-r)^\top\phi_{1,n}T}}\leqslant 0.
$$

When $1-e^{-(\mu-r)^\top\phi_{1,n}T}<0$,

$$
\begin{aligned}
&U_{w,n}h_w(\phi_{1,n},w_n)+2\left(1-e^{-(\mu-r)^\top\phi_{1,n}T}\right)^- U_{w,n}^2\\
=&(w_n-w^*)\left[\left(1-e^{-(\mu-r)^\top\phi_{1,n}T}\right)w_n+\left(x_0e^{-(\mu-r)^\top\phi_{1,n}T}-z\right)\right]+2\left(e^{-(\mu-r)^\top\phi_{1,n}T}-1\right)U_{w,n}^2\\
=&\left(e^{-(\mu-r)^\top\phi_{1,n}T}-1\right)w_n^2+\frac{1}{e^k-1}\left\{e^{-(\mu-r)^\top\phi_{1,n}T}\left[(-3z+x_0)e^k+2x_0\right]+2ze^k-3x_0+z\right\}w_n\\
&+\frac{ze^k-x_0}{(e^k-1)^2}\left\{e^{-(\mu-r)^\top\phi_{1,n}T}\left[(2z-x_0)e^k-x_0\right]+2x_0-z-ze^k\right\},
\end{aligned}
$$

which is also a convex quadratic function of w_n with the minimum value of

$$
-\frac{(z-x_0)^2}{4(e^k-1)^2}\frac{(e^{-(\mu-r)^\top(\phi_{1,n}-\phi_1^*)T}-1)^2}{e^{-(\mu-r)^\top\phi_{1,n}T}-1}\leqslant 0.
$$

To sum up, in both cases $U_{w,n}h_w(\phi_{1,n},w_n)+2(1-e^{-(\mu-r)^\top\phi_{1,n}})^-U_{w,n}^2$ is a convex quadratic function of w_n with the minimum value of $-M(\phi_{1,n})$. This implies that

$$
\zeta_n=2a_{w,n}\left(U_{w,n}h_w(\phi_{1,n},w_n)+2(1-e^{-(\mu-r)^\top\phi_{1,n}})^-U_{w,n}^2+M(\phi_{1,n})\right)\geqslant 0
$$

is always true for any n.

Third, we aim to prove $\sum\eta_n<\infty$ almost surely. By Theorem 4, $\phi_{1,n}\to\phi_1^*$; hence, there exists $0<N_2(\omega)<\infty$ such that $-1<(\mu-r)^\top(\phi_{1,n}-\phi_1^*)T<1$ for all $n\geqslant N_2(\omega)$. Additionally, for any $\delta>0$ there exists $N_3(\epsilon,\delta,\omega)>0$ such that $|\frac{z-x_0}{e^k-1}(1-e^{-(\mu-r)^\top(\phi_{1,n}-\phi_1^*)T})|<\frac{\epsilon\delta}{2}$ when $n\geqslant N_3(\epsilon,\delta,\omega)$.

Choose $N(\epsilon,\delta,\omega)=\max\{N_1(\epsilon,\omega),N_2(\omega),N_3(\epsilon,\delta,\omega)\}$. Notice that $(e^{-x}-1)^2\leqslant 4x^2$ when $-1\leqslant x\leqslant 1$. So when $n\geqslant N(\epsilon,\delta,\omega)$,

$$
(e^{-(\mu-r)^\top(\phi_{1,n}-\phi_1^*)T}-1)^2\leqslant 4|(\mu-r)^\top(\phi_{1,n}-\phi_1^*)|^2T^2\leqslant 4T^2|\mu-r|^2|\phi_{1,n}-\phi_1^*|^2.
$$

Furthermore, when $n \geqslant N(\epsilon, \delta, \omega)$, we have

$$M(\phi_{1,n}) \leqslant \frac{(z-x_0)^2}{(e^k-1)^2} \frac{T^2|\mu-r|^2|\phi_{1,n}-\phi_1^*|^2}{\epsilon} \leqslant C_\epsilon |\phi_{1,n}-\phi_1^*|^2.$$

By Theorem 5, the definition of $\hat{\eta}_n$ in (69) and the assumption (72) on $\{a_{w,n}\}$, we know

$$\sum_{n=1}^{\infty} a_{w,n} \mathbb{E}[|\phi_{1,n}-\phi_1^*|^2] \leqslant C' \sum_{n=1}^{\infty} a_{w,n} a_n \hat{\eta}_n < \infty.$$

Consider the sequence $S_m = \sum_{n=1}^m a_{w,n}|\phi_{1,n}-\phi_1^*|^2$, which is a monotone increasing sequence and $S_m \to S = \sum_{n=1}^\infty a_{w,n}|\phi_{1,n}-\phi_1^*|^2$. By the monotone convergence theorem, we have $\mathbb{E}[S_m] \to \mathbb{E}[S] = \sum_{n=1}^\infty a_{w,n}\mathbb{E}[|\phi_{1,n}-\phi_1^*|^2] < \infty$. It follows that $S = \sum_{n=1}^\infty a_{w,n}|\phi_{1,n}-\phi_1^*|^2 < \infty$ almost surely. This implies $\sum_{n=1}^\infty a_{w,n}M(\phi_{1,n}) \leqslant \sum_{n=1}^{N(\epsilon,\delta,\omega)-1} a_{w,n}M(\phi_{1,n}) + C_\epsilon \sum_{n=N(\epsilon,\delta,\omega)}^\infty a_{w,n}|\phi_{1,n}-\phi_1^*|^2 < \infty$ almost surely. Furthermore, if assumptions in (63) in Theorem 4 and assumptions in (72) in Theorem 6 are satisfied, then we have $\sum \eta_n < \infty$.

The above analysis yields $\sum \gamma_n < \infty$, $\sum \eta_n < \infty$ and ζ_n is non-negative. It follows from Robbins and Siegmund (1971, Theorem 1) that $|U_{w,n}|^2$ converges to a finite limit and $\sum \zeta_n < \infty$ almost surely.

Finally, we show $|U_{w,n}| \to 0$. Otherwise, there exists a set $Z \in \mathcal{F}$ with $\mathbb{P}(Z) > 0$, for every $\omega \in Z$, there exists $\delta(\omega) > 0$ such that for all n sufficiently large, $|w_n(\omega) - w^*| \geqslant \delta(\omega) > 0$. Consider the following function:

$$f(\phi_{1,n}, w_n) = U_{w,n}h_w(\phi_{1,n}, w_n) = (w_n - w^*)\left[\left(1-e^{-(\mu-r)^\top \phi_{1,n}T}\right)w_n + \left(x_0 e^{-(\mu-r)^\top\phi_{1,n}T} - z\right)\right].$$

When $n > N(\epsilon, \delta, \omega)$, we have

$$f(\phi_{1,n}(\omega), w^* + \delta(\omega)) = \delta(\omega)\left[\left(1-e^{-(\mu-r)^\top\phi_{1,n}(\omega)T}\right)\delta(\omega) + \frac{z-x_0}{e^k-1}\left(1-e^{-(\mu-r)^\top(\phi_{1,n}(\omega)-\phi_1^*)T}\right)\right],$$

and

$$f(\phi_{1,n}(\omega), w^* - \delta(\omega)) = -\delta(\omega)\left[-\left(1-e^{-(\mu-r)^\top\phi_{1,n}(\omega)}\right)\delta(\omega) + \frac{z-x_0}{e^k-1}\left(1-e^{-(\mu-r)^\top(\phi_{1,n}(\omega)-\phi_1^*)T}\right)\right].$$

Recall that for $n > N(\epsilon, \delta, \omega)$, $|\frac{z-x_0}{e^k-1}(1-e^{-(\mu-r)^\top(\phi_{1,n}(\omega)-\phi_1^*)T})| < \frac{\epsilon\delta(\omega)}{2}$ holds true, and f is a convex quadratic function of w_n with one root to be w^*. Then we have $f(\phi_{1,n}(\omega), w^*+\delta(\omega)) \geqslant \frac{\epsilon\delta(\omega)^2}{2} > 0$ and $f(\phi_{1,n}(\omega), w^*-\delta(\omega)) \geqslant \frac{\epsilon\delta(\omega)^2}{2} > 0$. Moreover, by the property of quadratic functions, we obtain $f(\phi_{1,n}(\omega), w) > \frac{\epsilon\delta(\omega)^2}{2} > 0$ for all $w \in (-\infty, w^*-\delta(\omega)] \cup [w^*+\delta(\omega), \infty)$. Thus, if $|w_n(\omega) - w^*| > \delta(\omega)$ for any $n > N(\epsilon, \delta, \omega)$,

$$\begin{aligned}\zeta_n(\omega) =& 2a_{w,n}U_{w,n}(\omega)h_w(\phi_{1,n}(\omega), w_n(\omega)) + 2a_{w,n}M(\phi_{1,n}(\omega)) \\ \geqslant& 2a_{w,n}U_{w,n}(\omega)h_w(\phi_{1,n}(\omega)w_n(\omega)) \geqslant a_{w,n}\epsilon\delta(\omega)^2.\end{aligned}$$

Then

$$\sum_{n=1}^{\infty}\zeta_n(\omega)=\sum_{n=1}^{N(\epsilon,\delta,\omega)-1}\zeta_n(\omega)+\sum_{n=N(\epsilon,\delta,\omega)}^{\infty}\zeta_n(\omega)\geqslant\sum_{n=1}^{N(\epsilon,\delta,\omega)-1}\zeta_n(\omega)+\sum_{n=N(\epsilon,\delta,\omega)}^{\infty}a_{w,n}\epsilon\delta(\omega)^2=\infty,$$

which contradicts the fact that $\sum_{n=1}^{\infty}\zeta_n<\infty$ almost surely. Therefore, $w_n\to w^*$ almost surely.

□

Now, Theorem 1 follows from combining Theorems 4, 5, 6, and Remark 1.

G.2 Proof of Theorem 2

We first recall a simple result regarding the inner product between two positive semi-definite matrices.

Lemma 4. *For two matrices $M,N\in\mathbb{S}_+^d$, we have $\langle M,N\rangle\geqslant 0$.*

Proof. Since $M\in\mathbb{S}_+^d$, it can be represented as $M=Q^\top DQ$, where $D=diag(\lambda_1,\lambda_2,...\lambda_d)$ is a diagonal matrix with the diagonal entries being the (nonnegative) eigenvalues of M, and $Q=(q_1,q_2,...,q_d)$ is a matrix consisting of the corresponding eigenvectors of M. Then,

$$\begin{aligned}\langle M,N\rangle&=\langle\sum_{i=1}^{d}\lambda_i q_i q_i^\top,N\rangle=\sum_{i=1}^{d}\lambda_i\langle q_i q_i^\top,N\rangle\\&=\sum_{i=1}^{d}\lambda_i(q_i^\top N q_i)\geqslant 0,\end{aligned}\tag{74}$$

noting that $\lambda_i\geqslant 0$ and $N\in\mathbb{S}_+^d$. □

We now prove Theorem 2. Note that the wealth processes $x^{u^{\boldsymbol{\pi}}}$ and $x^{\boldsymbol{\pi}}$ have identical distributions. It follows from (24) that the wealth processes $\{x^{\boldsymbol{\pi}}(t):0\leqslant t\leqslant T\}$ and $\{x^{\hat{\boldsymbol{\pi}}}(t):0\leqslant t\leqslant T\}$ follow the dynamics:

$$\mathrm{d}x^{\boldsymbol{\pi}}(t)=-(\mu-r)^\top\phi_1(x^{\boldsymbol{\pi}}(t)-w)\mathrm{d}t+\sqrt{\langle\sigma\sigma^\top,\phi_1\phi_1^\top(x^{\boldsymbol{\pi}}(t)-w)^2+C(t)\rangle}\mathrm{d}W(t),$$

and

$$\mathrm{d}x^{\hat{\boldsymbol{\pi}}}(t)=-(\mu-r)^\top\phi_1(x^{\hat{\boldsymbol{\pi}}}(t)-w)\mathrm{d}t+\sqrt{\langle\sigma\sigma^\top,\phi_1\phi_1^\top(x^{\hat{\boldsymbol{\pi}}}(t)-w)^2+\hat{C}(t)\rangle}\mathrm{d}W(t).$$

Taking integration and then expectation on both equations and denoting $g(t)=\mathbb{E}[x^{\boldsymbol{\pi}}(t)]$ and $\hat{g}(t)=\mathbb{E}[x^{\hat{\boldsymbol{\pi}}}(t)]$, we find that g and $\hat{g}$ satisfy the same ODE:

$$g'(t)=-Ag(t)+Aw,\;\; g(0)=x_0;\;\; \hat{g}'(t)=-A\hat{g}(t)+Aw,\;\; \hat{g}(0)=x_0,\tag{75}$$

where $A=(\mu-r)^\top\phi_1$. The uniqueness of solution to this ODE implies $g\equiv\hat{g}$ and, hence, $\mathbb{E}[x^{\boldsymbol{\pi}}(T)]=\mathbb{E}[x^{\hat{\boldsymbol{\pi}}}(T)]$.

Next, applying Itô's formula to $(x^{\boldsymbol{\pi}}(t))^2$, and then integrating and taking expectation, we obtain that $k(t) = \mathbb{E}\left[(x^{\boldsymbol{\pi}}(t))^2\right]$ satisfies

$$k'(t) = (-2A + B)k(t) + 2w(A - B)g(t) + w^2 B + \langle \sigma\sigma^\top, C(t) \rangle, \tag{76}$$

where $B = \langle \sigma\sigma^\top, \phi_1\phi_1^\top \rangle$. Similarly, $\hat{k}(t) = \mathbb{E}\left[(x^{\hat{\boldsymbol{\pi}}}(t))^2\right]$ satisfies

$$\hat{k}'(t) = (-2A + B)\hat{k}(t) + 2w(A - B)\hat{g}(t) + w^2 B + \langle \sigma\sigma^\top, \hat{C}(t) \rangle. \tag{77}$$

However, Lemma 4 yields $\langle \sigma\sigma^\top, C(t) \rangle \geqslant \langle \sigma\sigma^\top, \hat{C}(t) \rangle$. Thus it follows from applying the comparison theorem of ODEs to (76) and (77) that $k(t) \geqslant \hat{k}(t)$ $\forall t \in [0, T]$. The desired result that $\mathrm{Var}(x^{\boldsymbol{\pi}}(T)) \geqslant \mathrm{Var}(x^{\hat{\boldsymbol{\pi}}}(T))$ follows immediately.

G.3 Proof of Theorem 3

We first show that the Sharpe ratio is a function of just ϕ_1. For ease of exposition, the wealth process $x^{\boldsymbol{u}}(t)$ will henceforth be denoted simply as $x(t)$. Indeed, under the deterministic policy (17), $\mathbb{E}[x(\cdot)]$ satisfies the same ODE (75). Solving it we get

$$\mathbb{E}[x(t)] = w + (x_0 - w)e^{-At}.$$

Moreover, solving the ODE (76) with $C = 0$, we obtain

$$\mathbb{E}[x(t)^2] = e^{(-2A+B)t}(w-1)^2 - 2e^{-At}(w^2 - w) + w^2.$$

Hence

$$\mathrm{Var}(x(t)) = (x_0 - w)^2 e^{-2At}(e^{Bt} - 1),$$

leading to

$$\mathrm{SR}(\phi_1) = \frac{(\mathbb{E}[x(T)] - x_0)/x_0}{\sqrt{\mathrm{Var}(x(T)/x_0)}} = \frac{e^{AT} - 1}{\sqrt{e^{BT} - 1}}. \tag{78}$$

Next we prove that $\mathrm{SR}(\phi_1)$ is uniformly bounded in $\phi_1 \in \mathbb{R}^d$. To this end, first note that $\mathrm{SR}(\phi_1)$ is a continuous function of ϕ_1 except at $\phi_1 = 0$. Denote by λ_{min} the smallest eigenvalue of the positive semi-definite matrix Σ. Then, on one hand,

$$
\begin{aligned}
\limsup_{|\phi_1|\to 0} |\operatorname{SR}(\phi_1)| &\leqslant \limsup_{|\phi_1|\to 0} \frac{T|(\mu-r)^\top\phi_1 + \frac{1}{2}((\mu-r)^\top\phi_1)^2 + O(|\phi_1|^3)|}{\sqrt{T\phi_1^\top\Sigma\phi_1}} \\
&\leqslant \limsup_{|\phi_1|\to 0} \frac{|\mu-r||\phi_1| + \frac{1}{2}|\mu-r|^2|\phi_1|^2 + O(|\phi_1|^3)}{\sqrt{\lambda_{min}|\phi_1|^2}}\sqrt{T} \\
&= \frac{|\mu-r|\sqrt{T}}{\sqrt{\lambda_{min}}}.
\end{aligned}
$$

On the other hand, note that $B = \phi_1^\top\Sigma\phi_1 \geqslant \lambda_{min}|\phi_1|^2 \to \infty$ as $|\phi_1| \to \infty$. In particular, when $|\phi_1| > \frac{1}{\sqrt{\lambda_{min}T}}$, $e^{BT} - 1 \geqslant \frac{1}{4}e^{BT}$. Therefore,

$$
\begin{aligned}
\limsup_{|\phi_1|\to\infty} |\operatorname{SR}(\phi_1)| \leqslant \limsup_{|\phi_1|\to\infty} \frac{e^{AT}}{\sqrt{e^{BT}-1}} &\leqslant \limsup_{|\phi_1|\to\infty} \frac{e^{|\mu-r||\phi_1|T}}{\sqrt{\frac{1}{4}e^{BT}}} \\
&\leqslant \limsup_{|\phi_1|\to\infty} 2e^{|\mu-r||\phi_1|T - \frac{1}{2}\phi_1^\top\Sigma\phi_1 T} \\
&\leqslant \limsup_{|\phi_1|\to\infty} 2e^{|\mu-r||\phi_1|T - \frac{1}{2}\lambda_{min}|\phi_1|^2 T} = 0.
\end{aligned}
$$

It follows then $|\operatorname{SR}(\phi_1)| \leqslant C_1\ \forall \phi_1 \in \mathbb{R}^d$ for some constant $C_1 > 0$.

Now, SR reaches its maximum at $\phi_1 = \phi_1^*$; hence $\operatorname{SR}'(\phi_1^*) = 0$. Next we show $\operatorname{SR}''(\phi_1^*) \leqslant 0$. Recall that $k = (\mu-r)^\top\Sigma^{-1}(\mu-r)T$,

$$
\operatorname{SR}''(\phi_1^*) = -\frac{1}{2}(e^k-1)^{-\frac{3}{2}}e^k[(e^k-1)\Sigma T - (\mu-r)(\mu-r)^\top T^2],
$$

where

$$
\begin{aligned}
(e^k-1)\Sigma T - (\mu-r)(\mu-r)^\top T^2 &\geqslant k\Sigma T - (\mu-r)(\mu-r)^\top T^2 \\
&= T^2((\mu-r)^\top\Sigma^{-1}(\mu-r)\Sigma - (\mu-r)(\mu-r)^\top).
\end{aligned}
$$

Consider the matrix $(\mu-r)^\top\Sigma^{-1}(\mu-r)\Sigma - (\mu-r)(\mu-r)^\top$, by the Cauchy–Schwarz inequality, for any vector $x \in \mathbb{R}^d$,

$$
\begin{aligned}
& x^\top((\mu-r)^\top\Sigma^{-1}(\mu-r)\Sigma - (\mu-r)(\mu-r)^\top)x \\
=& (\mu-r)^\top\Sigma^{-1}(\mu-r)x^\top\Sigma x - x^\top(\mu-r)(\mu-r)^\top x \\
=& (\mu-r)^\top\Sigma^{-1}(\mu-r)(x^\top\Sigma x) - (x^\top(\mu-r))^2 \\
\geqslant& 0.
\end{aligned}
$$

Therefore, we have $\operatorname{SR}''(\phi_1^*) \leqslant 0$.

Fix a constant $\delta < |\phi_1^*|$. Then for any ϕ_1 such that $|\phi_1 - \phi_1^*| < \delta$, we have $\operatorname{SR}''(\phi_1) \geq -\bar{C}I$ for some constant $\bar{C} > 0$, because SR'' is continuous in this region.

By Taylor's expansion, for any ϕ_1 with $|\phi_1 - \phi_1^*| < \delta$, we have

$$\begin{aligned}
\mathrm{SR}(\phi_1) - \mathrm{SR}(\phi_1^*) &= \mathrm{SR}'(\phi_1^*)(\phi_1 - \phi_1^*) + \int_0^1 (1-t)(\phi_1 - \phi_1^*)^\top \mathrm{SR}''(\phi_1^* + t(\phi_1 - \phi_1^*))(\phi_1 - \phi_1^*)\mathrm{d}t \\
&= \int_0^1 (1-t)(\phi_1 - \phi_1^*)^\top \mathrm{SR}''(\phi_1^* + t(\phi_1 - \phi_1^*))(\phi_1 - \phi_1^*)\mathrm{d}t \\
&\geqslant -\int_0^1 (1-t)\bar{C}|\phi_1 - \phi_1^*|^2 \mathrm{d}t = -\frac{1}{2}\bar{C}|\phi_1 - \phi_1^*|^2,
\end{aligned}$$

or $\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_1) \leqslant \frac{1}{2}\bar{C}|\phi_1 - \phi_1^*|^2$.

Recall that Theorem 1-(b) yields that

$$\begin{aligned}
\mathbb{E}[|\phi_{1,n} - \phi_1^*|^2] &\leqslant C\frac{(\log(n-1))^p \log\log(n-1)}{n-1} \\
&\leqslant C\frac{(\log n)^p \log\log n}{n-1} \\
&= C\frac{(\log n)^p \log\log n}{n} * \frac{n}{n-1} \\
&\leqslant \check{C}\frac{(\log n)^p \log\log n}{n},
\end{aligned}$$

where $\check{C}$ is a constant independent of n.

Set $\delta_n' = (4\frac{C_1\check{C}}{\bar{C}}\frac{(\log n)^p \log\log n}{n})^{\frac{1}{4}}$, $n \in \mathbb{N}$, and $n_0 = \inf\{n : \delta_n' < \delta\}$. Further, define $\delta_n = \delta$ for $n < n_0$, and $\delta_n = \delta_n'$ for $n \geqslant n_0$. Then, for $n \in \mathbb{N}$, we have

$$\begin{aligned}
&\mathbb{E}[\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_{1,n})] \\
&= \int_{|\phi_{1,n} - \phi_1^*| \leqslant \delta_n} [\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_{1,n})]\mathrm{d}\mathbb{P} + \int_{|\phi_{1,n} - \phi_1^*| > \delta_n} [\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_{1,n})]\mathrm{d}\mathbb{P} \\
&\leqslant \int_{|\phi_{1,n} - \phi_1^*| \leqslant \delta_n} \frac{1}{2}\bar{C}|\phi_{1,n} - \phi_1^*|^2 \mathrm{d}\mathbb{P} + \int_{|\phi_{1,n} - \phi_1^*| > \delta_n} 2C_1 \mathrm{d}\mathbb{P} \\
&\leqslant \frac{1}{2}\bar{C}\delta_n^2 + 2C_1\mathbb{P}(|\phi_{1,n} - \phi_1^*| > \delta_n).
\end{aligned}$$

When $n < n_0$, we have $\mathbb{E}[\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_{1,n})] \leqslant \frac{1}{2}\bar{C}\delta^2 + 2C_1$. When $n > n_0$, we have

$$\begin{aligned}
&\mathbb{E}[\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_{1,n})] \\
&\leqslant \frac{1}{2}\bar{C}\delta_n^2 + 2C_1\mathbb{P}(|\phi_{1,n} - \phi_1^*| > \delta_n) \\
&\leqslant \frac{1}{2}\bar{C}\delta_n^2 + 2C_1\frac{1}{\delta_n^2}\mathbb{E}[|\phi_{1,n} - \phi_1^*|^2] \\
&\leqslant \frac{1}{2}\bar{C}\delta_n^2 + 2C_1\frac{\check{C}}{\delta_n^2}\frac{(\log n)^p \log\log n}{n} \\
&= 2\sqrt{\bar{C}C_1\check{C}\frac{(\log n)^p \log\log n}{n}}.
\end{aligned}$$

Consequently,

$$
\begin{aligned}
&\mathbb{E}[\sum_{n=1}^{N}(\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_{1,n}))] \\
&= \sum_{n=1}^{N} \mathbb{E}[\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_{1,n})] \\
&= \sum_{n=1}^{n_0} \mathbb{E}[\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_{1,n})] + \sum_{n=n_0}^{N} \mathbb{E}[\mathrm{SR}(\phi_1^*) - \mathrm{SR}(\phi_{1,n})] \\
&\leqslant (\frac{1}{2}\bar{C}\delta^2 + 2C_1)n_0 + 2\sum_{n=n_0}^{N}\sqrt{\bar{C}C_1\check{C}\frac{(\log n)^p \log\log n}{n}} \\
&\leqslant C + C\sqrt{N(\log N)^p \log\log N}.
\end{aligned}
$$

The proof is complete.