ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Huo X, Fu F. 2017 Risk-aware multi-armed bandit problem with application to portfolio selection. *R. Soc. open sci.* **4**: 171377. http://dx.doi.org/10.1098/rsos.171377

Received: 14 September 2017
Accepted: 13 October 2017

Subject Category:
Physics

Subject Areas:
statistical physics/applied mathematics

Keywords:
multi-armed bandit, online learning, portfolio selection, graph theory, risk-awareness, conditional value-at-risk

Authors for correspondence:
Xiaoguang Huo
e-mail: xh84@cornell.edu
Feng Fu
e-mail: fufeng@gmail.com

Risk-aware multi-armed bandit problem with application to portfolio selection

Xiaoguang Huo[1] and Feng Fu[2,3]

[1]Department of Mathematics, Cornell University, Ithaca, NY 14850, USA
[2]Department of Mathematics, Dartmouth College, Hanover, NH 03755, USA
[3]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756, USA

FF, 0000-0001-8252-1990

Sequential portfolio selection has attracted increasing interest in the machine learning and quantitative finance communities in recent years. As a mathematical framework for reinforcement learning policies, the stochastic multi-armed bandit problem addresses the primary difficulty in sequential decision-making under uncertainty, namely the *exploration* versus *exploitation* dilemma, and therefore provides a natural connection to portfolio selection. In this paper, we incorporate risk awareness into the classic multi-armed bandit setting and introduce an algorithm to construct portfolio. Through filtering assets based on the topological structure of the financial market and combining the optimal multi-armed bandit policy with the minimization of a coherent risk measure, we achieve a balance between risk and return.

1. Introduction

Portfolio selection is a popular area of study in the financial industry ranging from academic researchers to fund managers. The problem involves determining the best combination of assets to be held in the portfolio in order to achieve the investor's objectives, such as maximizing the cumulative return relative to some risk measure. In the finance community, the traditional approach to this problem can be traced back to 1952 with Markowitz's seminal paper [1], which introduces mean-variance analysis, also known as the modern portfolio theory (MPT), and suggests choosing the allocation that maximizes the expected return for a certain risk level quantified by variance. On the other hand, sequential portfolio selection models have been developed in the mathematics and computer science communities; for example, Cover's universal portfolio strategy [2], Helmbold's multiplicative update portfolio strategy [3] and also see

THE ROYAL SOCIETY PUBLISHING

Li & Hoi [4] for a comprehensive survey. In recent years, with the unprecedented success of AI and machine learning methods evidenced by AlphaGo defeating the world champion and OpenAI's bot beating professional Dota players, more creative machine learning-based portfolio selection strategies also emerged [5,6].

Including portfolio selection, many practical problems such as clinical trials, online advertising and robotics can be modelled as sequential decision-making under uncertainty [7]. In such a process, at each trial the learner faces the trade-off between acting ambitiously to acquire new knowledge and acting conservatively to take advantage of current knowledge, which is commonly known as the *exploration* versus *exploitation* dilemma. Often understood as a single-state Markov decision process (MDP), the stochastic multi-armed bandit problem provides an extremely intuitive mathematical framework to study sequential decision-making.

An abstraction of this setting involves a set of K slot machines and a sequence of N trials. At each trial $t = 1, \ldots, N$, the learner chooses to play one of the machines $I_t \in \{1, \ldots, K\}$ and receives a reward $R_{I_t,t}$ drawn randomly from the corresponding fixed but unknown probability distribution ν_{I_t}, whose mean is μ_{I_t}. In the classic setting, the random rewards of the same machine across time are assumed to be independent and identically distributed, and the rewards of different machines are also independent. The objective of the learner is to develop a *policy*, an algorithm that specifies which machine to play at each trial, to maximize cumulative rewards. A popular measure for the performance of a policy is the *regret* after some n trials, which is defined to be

$$\xi(n) \stackrel{\text{def}}{=} \max_{\forall i \in [1,K]} \sum_{t=1}^{n} R_{i,t} - \sum_{t=1}^{n} R_{I_t,t}. \tag{1.1}$$

However, in a stochastic model it is more intuitive to compare rewards in expectation and use *pseudo-regret* [8]. Let $T_i(n)$ be the number of times machine i is played during the first n trials and let $\mu^* = \max\{\mu_1, \ldots, \mu_K\}$. Then,

$$\hat{\xi}(n) \stackrel{\text{def}}{=} n\mu^* - \mathbb{E}\sum_{t=1}^{n} R_{I_t,t} = \sum_{1 \le i \le K,\, \mu_i < \mu^*} (\mu^* - \mu_i)\mathbb{E}[T_i(n)]. \tag{1.2}$$

Thus, the learner's objective to maximize cumulative rewards is then equivalent to minimizing regret. The asymptotic lower bound on the best possible growth rate of total regret is proved by Lai & Robbins [9], which is $\mathcal{O}(\log n)$ with a coefficient determined by the suboptimality of each machine and the Kullback–Leibler divergence. Since then, various online learning policies have been proposed [10], among which the UCB1 policy developed in Auer *et al.* [11] is considered the optimal and will be introduced in detail in Methods and model section.

Although the classic multi-armed bandit has been well studied in academia, a number of variants of this problem are proposed to model different real-world scenarios. For example, Agrawal & Goyal [12] considers a contextual bandit with a linear reward function and analyses the performance of the Thompson sampling algorithm. Koulouriotis & Xanthopoulos [13] studies the non-stationary setting where the reward distributions of machines change at a fixed time. A more important variant is the risk-aware setting, where the learner considers risk in the objective instead of simply maximizing the cumulative reward. This variant is closely related to the portfolio selection problem, where risk management is an indispensable concern, and has been discussed in several papers. For example, Sani *et al.* [14] studies the problem where the learner's objective is to minimize the mean variance defined as $\sigma^2 - \rho\mu$ and proposes two algorithms, MV-LCB and ExpExp. In a similar setting, Vakili & Zhao [15] provides a finer analysis of the performance of algorithms proposed in Sani *et al.* [14]. In addition, Vakili & Zhao [16] extends this setting by considering the mean variance and value-at-risk of total rewards at the end of the time horizon. In a more generalized case, Zimin *et al.* [17] sets the objective to be a function of the mean and the variance $f(\mu, \sigma^2)$ and defines the φ-LCB algorithm that achieves desirable performance under certain conditions. Moreover, Galichet *et al.* [18] chooses the conditional value-at-risk to be the objective and proposes the MARAB algorithm.

These works serve as the inspiration for us to consider risk in the model, but they are not directly applicable to the portfolio selection problem, owing to the primary obstacle that these methods only choose the best single machine to play at each trial. To address this issue, a basket of candidate portfolios need to be first selected in the preliminary stage in a strategic and logical way. For example, Shen *et al.* [19] uses principal component analysis (PCA) to select candidate portfolios, namely the normalized eigenvectors of the covariance matrix of asset returns.

In our model, we first take a graph theory approach to filter and select a basket of assets, which we use to construct the portfolio. Then, at each trial we combine the single-asset portfolio determined by the optimal multi-armed bandit algorithm with the portfolio that globally minimizes a *coherent* risk measure, the conditional value-at-risk. The rest of this paper is organized as follows. In Methods and model section, we formulate the portfolio selection problem in the multi-armed bandit setting, and describe our methodology in detail. In Results section, we present our simulation results using the proposed method. In Discussion and conclusion section, we discuss results and also provide directions for future research.

2. Methods and model

2.1. Problem formulation

In this section, we modify the classic multi-armed bandit setting to model portfolio selection. Consider a financial market with a large set of assets, from which the learner selects a basket of K assets to invest in a sequence of N trials. At each trial $t = 1, \ldots, N$, the learner chooses a portfolio $\boldsymbol{\omega}_t = (\omega_{1,t}, \ldots, \omega_{K,t})^\top$ where $\omega_{i,t}$ is the weight of asset i. As we only consider long-only and self-financed trading, we must have $\boldsymbol{\omega}_t \in W$ where $W = \{\boldsymbol{u} \in \mathbb{R}_+^K : \boldsymbol{u}^\top \mathbf{1} = 1\}$ and $\mathbf{1}$ is a column vector of ones. The returns of assets are then revealed at trial $t+1$ and denoted by $\boldsymbol{R}_t = (R_{1,t}, \ldots, R_{K,t})^\top$. In particular, the return for each asset $R_{i,t}$ is viewed as a random draw from the corresponding probability distribution ν_i with mean μ_i and can be simply defined as the log price ratio $R_{i,t} = \log(P_{i,t+1}/P_{i,t})$, where we use the natural log, and $P_{i,t}$, $P_{i,t+1}$ denote the prices at trial t and $t+1$, respectively. For the trading period from t to $t+1$, the learner receives $\boldsymbol{\omega}_t{}^\top \boldsymbol{R}_t$ as the reward for his portfolio. The investment strategy of the learner is thus a sequence of N mappings from the accumulated knowledge to W.

We make the following assumptions. First, we assume we always have access to historical returns $H_{i,t}$ of every asset i in the market for $t = 1, \ldots, \delta$. The historical return is defined similarly to $R_{i,t}$ as the log price ratio but corresponds to the time horizon immediately before our investment period. They are only used to estimate the correlation structure and risk level. Second, we make no assumption on the dependency of returns either across time or across assets. We only assume that, for each trial t and for all $i \in \{1, \ldots, K\}$, $R_{i,t} \sim \nu_i$ and $H_{i,t} \sim \nu_i$ with a relatively small δ. Note that the UCB1 algorithm we use later is proved to be optimal under a weaker assumption, $\mathbb{E}[R_{i,t} \mid R_{i,1}, \ldots, R_{i,t-1}] = \mu_i$, allowing us to waive the assumptions in the classic setting [11]. Third, transaction costs and market liquidity will not be considered. See Model 1 for a summary of the problem.

Model 1: Sequential portfolio selection problem

Parameters: δ, N
Receive historical returns $H_{i,t}$ of each asset i for $t = 1, \ldots, \delta$;
Filter to select a basket of K assets;
for $t = 1, \ldots, N$ **do**
  Choose portfolio $\boldsymbol{\omega}_t = (\omega_{1,t}, \ldots, \omega_{K,t})^\top$;
  Observe $\boldsymbol{R}_t = (R_{1,t}, \ldots, R_{K,t})^\top$ and receive reward $\boldsymbol{\omega}_t{}^\top \boldsymbol{R}_t$;
end

2.2. Portfolio construction by filtering assets

Graph theory has been popularly applied in various disciplines to model networks, where the vertices represent individuals of interest and the edges represent their interactions. For example, in evolutionary game theory, graphs are used to analyse the dynamics of cooperation within different population structures [20–25]. In financial markets, the minimum spanning tree (MST) is accepted as a robust method to visualize the structure of assets [26], allowing one to capture different market sectors from empirical data [27–29].

For our purpose, as we have a large pool of assets, we first want to select a basket of K to invest in. Recall that the return of each asset is $R_{i,t} = \log(P_{i,t+1}/P_{i,t})$, where $P_{i,t}$ and $P_{i,t+1}$ are the prices at trial t and $t+1$. Following Mantegna [27] and Mantegna & Stanley [30], we use δ trials of historical returns to find the correlation matrix, whose entries are

$$\rho_{i,j} \overset{\text{def}}{=} \frac{\langle H_i H_j \rangle - \langle H_i \rangle \langle H_j \rangle}{\sqrt{(\langle H_i{}^2 \rangle - \langle H_i \rangle^2)(\langle H_j{}^2 \rangle - \langle H_j \rangle^2)}},$$

where $\langle\cdot\rangle$ is the historical mean, namely $\langle H_i\rangle = \sum_{t=1}^{\delta} H_{i,t}$ for each asset i in the market. For δ small, we can improve our estimation by taking advantage of the shrinkage method in Ledoit & Wolf [31]. We then define the metric distance between two vertices as $d_{i,j} \overset{\text{def}}{=} \sqrt{2(1-\rho_{i,j})}$. The Euclidean distance matrix $\boldsymbol{D}$ whose entries are $d_{i,j}$ is then used to compute the undirected graph $G = \{V, E\}$, where V is the set of vertices representing assets and E is the set of weighted edges representing distance. To extract the most important edges from G, we construct the MST T. In particular, T is the subgraph of G that connects all vertices without cycle and minimizes total edge weights.

One way to classify vertices is based on their relative positions in the graph, central versus peripheral. In financial markets, this classification method turns out to have significant implications in *systemic risk*, which is the risk that an economic shock causes the collapse of a chain of institutions [32]. Several empirical studies suggest that such risk can be associated with certain characteristics of the correlation structure of the market. For example, Kritzman *et al.* [33] defines the *absorption ratio* as the fraction of total variances explained by a fixed number of principal components, namely the eigenvectors of the covariance matrix, and shows this ratio increased dramatically during both domestic and global financial crises including the housing bubble, dot-com bubble, the 1997 Asian financial crisis and so on. Drozdz *et al.* [34] finds a similar result and suggests that the maximum eigenvalue of the correlation matrix rises during crisis and exhausts the total variances. Hence, graph theory can be naturally applied to this setting and provides significant insights into managing systemic risk. In particular, Huang *et al.* [35] gives an intuitive simulation of the contagion process of systemic risk on a bipartite graph. Onnela *et al.* [36] shows that the MST of assets shrinks during a crisis, which supports the above arguments on the compactness of the eigenvalues of correlation matrix. More importantly, Onnela *et al.* [36], Pozzi *et al.* [37] and Ren *et al.* [38] suggest that investing in the assets located on the peripheral parts of the MST can facilitate diversification and reduce the exposure to systemic risk during a crisis.

For our study, we select 30 S&P 500 stocks, which consist of 15 financial institutions (JPM, WFC, BAC, C, GS, USB, MS, KEY, PNC, COF, AXP, PRU, SCHW, BBT, STI) and 15 randomly selected companies from other sectors (KR, PFE, XOM, WMT, DAL, CSCO, HCP, EQIX, DUK, NFLX, GE, APA, F, REGN, CMS). We use the daily close price of 44 trading days during the subprime mortgage crisis to construct the MST and investigate the advantage of investing in peripheral vertices using the equally weighted portfolio strategy. Although the number of stocks is small, our results similarly show that investing in peripheral vertices can reduce loss during financial crisis (figure 1). Figure 1*a* shows the complete graph of 30 stocks. Figure 1*b* is the MST we obtain following the above method. Observe that this tree has a total of 14 leaves (WFC, C, GS, KEY, PNC, SCHW, KR, DAL, HCP, EQIX, DUK, NFLX, GE, F), and selecting from these leaves to construct a portfolio almost always reduces the median daily loss compared with the portfolio with all vertices. For example, figure 1*c* provides the performance of the portfolio with 10 randomly selected vertices from the 14 leaves, which increases the median daily log price ratio from -0.0101 to -0.0079 and the median daily percentage return from -0.0095 to -0.0070. Furthermore, figure 1*d* shows that the eigenvalue spectrum of the covariance matrix becomes less compact. Finally, we acknowledge the dynamic nature of the market structure, but for simplicity this aspect will not be considered in our study.

Therefore, we select the K most peripheral vertices from the MST T as our basket of assets to invest in. We note that for any graph G with distinct edge weight, which is often the case for financial data with high precision, the MST T is proved to be unique. Our selection of vertices tends to lie on the leaves for a star-like graph, on the two ends of the longest edge for a cycle, and on the corners for a lattice. Among the numerous centrality measures discussed in graph theory [39], we use the most straightforward measure and select the K vertices with the least *degree*. The value of K is subjective and can be determined based on the learner's view of the economic state. Assuming K assets are selected, we proceed to portfolio construction as described in what follows.

2.3. Combined sequential portfolio selection algorithm

We design a sequential portfolio selection algorithm by combining the optimal multi-armed bandit policy, namely the UCB1 proposed in Auer *et al.* [11], with the minimization of a coherent risk measure, namely the conditional value-at-risk. Recall that the return $R_{i,t}$ of each asset i is defined as the log price ratio, namely $R_{i,t} = \log(P_{i,t+1}/P_{i,t})$. The UCB1 policy is defined as follows. First, select each asset once and observe return during the first K trials. Then, for each trial select the asset that maximizes an estimated upper confidence bound of return with a certain confidence level. Precisely, at each trial t

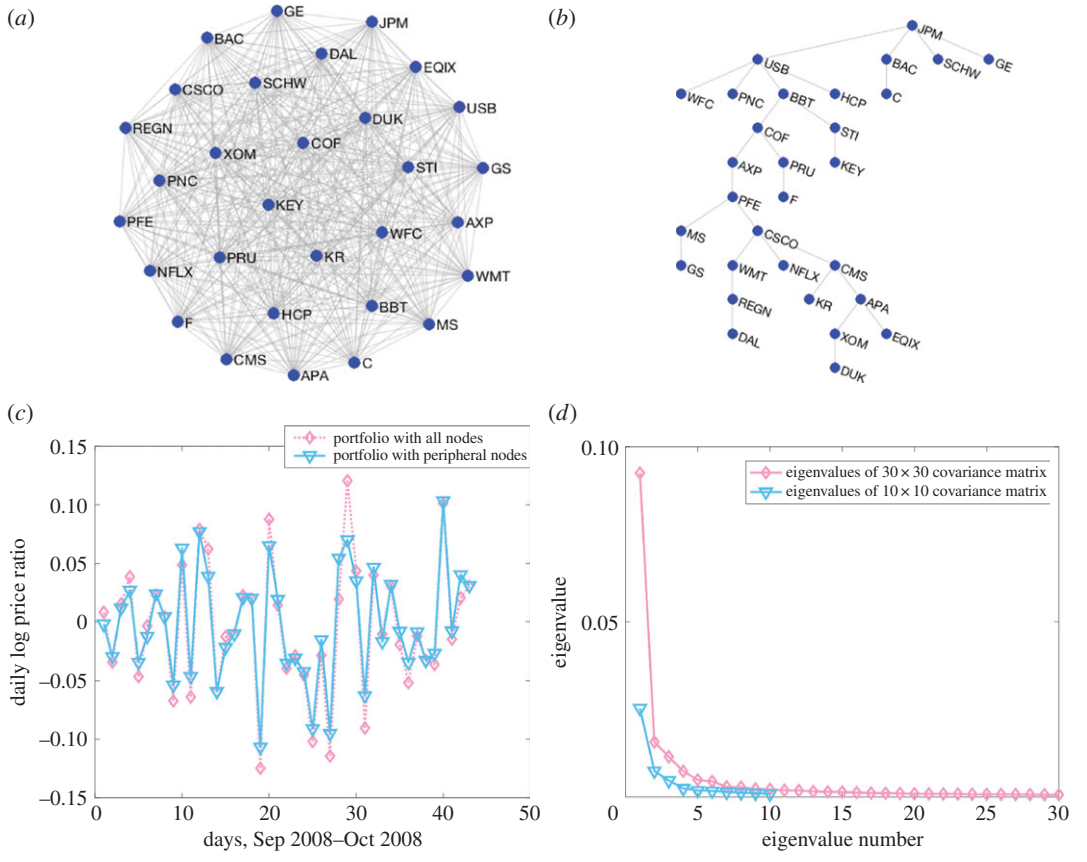

Figure 1. Portfolio selection based on the MST. (*a*) The complete graph and (*b*) the corresponding MST constructed from the 30 selected S&P 500 stocks during the period September 2008 to October 2008. (*c*) The performance of the portfolio of 10 randomly selected vertices from the 14 leaves shown in *b*. (*d*) The eigenvalue spectrum of the covariance matrix of the 30 selected S&P 500 stocks in *a* with that of 10 stocks randomly chosen from the peripheral nodes from the MST in *c*.

we select

$$I_t^* \stackrel{\text{def}}{=} \begin{cases} t & \text{if } t \le K, \\ \arg\max_{\forall i \in \{1,\ldots,K\}} \bar{R}_i(t) + \sqrt{\dfrac{2\log t}{T_i(t-1)}} & \text{otherwise,} \end{cases} \tag{2.1}$$

where $\bar{R}_i(t)$ is the empirical mean of return for asset i and recall that $T_i(t-1)$ is the number of times asset i has been selected during the past $t-1$ trials. Theorem 2.1 below provided in Auer *et al*. [11] proves the optimality of UCB1.

Theorem 2.1 [11]. *For all $K > 1$ assets whose mean returns are in the support $[0,1]$, the regret of UCB1 algorithm after any number n of trials satisfies*

$$\hat{\xi}(n) \le \left[8 \sum_{i:\mu_i < \mu^*} \left(\frac{\log n}{\mu^* - \mu_i} \right) \right] + \left(1 + \frac{\pi^2}{3} \right) \left[\sum_{i=1}^{K} (\mu^* - \mu_i) \right],$$

where μ_i is the mean return of asset i and $\mu^* = \max\{\mu_1, \ldots, \mu_K\}$.

The proof makes no assumption on the dependency and distribution of asset returns besides $\mathbb{E}[R_{i,t} \mid R_{i,1}, \ldots, R_{i,t-1}] = \mu_i$. Therefore, by scaling the values we can achieve optimality. In addition, we can use historical returns and observed returns of unselected assets to further improve performance, but we do not discuss details here. Let $\boldsymbol{e}_i \in \mathbb{R}^K$ be the vector of a single 1 on entry i and 0 on the others. Our single-asset multi-armed bandit portfolio at t chosen according to equation (2.1) is

$$\boldsymbol{\omega}_t^M \stackrel{\text{def}}{=} \boldsymbol{e}_{I_t^*}. \tag{2.2}$$

Now, let us incorporate risk awareness into our algorithm by finding the portfolio that achieves the global minimum of the conditional value-at-risk. We define risk measure and associated properties following Artzner *et al.* [40] and Bäuerle & Rieder [41].

Definition 2.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and denote by $\mathcal{L}(\Omega, \mathcal{F}, \mathbb{P})$ the set of integrable random variables, where any instance of $\mathcal{L}(\Omega, \mathcal{F}, \mathbb{P})$ represents portfolio return. A function $\Psi : \mathcal{L}(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ is called a risk measure.

Definition 2.3. Let Ψ be a risk measure; we say Ψ is a coherent risk measure if, for all $X_1, X_2 \in \mathcal{L}(\Omega, \mathcal{F}, \mathbb{P})$, $c \in \mathbb{R}$ and $d \in \mathbb{R}_+ \cup \{0\}$, it satisfies

— *translation invariance*: $\Psi(X_1 + c) = \Psi(X_1) - c$
— *subadditivity*: $\Psi(X_1 + X_2) \leq \Psi(X_1) + \Psi(X_2)$
— *positive homogeneity*: $\Psi(dX_1) = d\Psi(X_1)$
— *monotonicity*: $X_1 \leq X_2 \Rightarrow \Psi(X_1) \geq \Psi(X_2)$

Definition 2.4. Let $X \in \mathcal{L}(\Omega, \mathcal{F}, \mathbb{P})$; the risk measure value-at-risk of X at confidence level $\beta \in (0, 1)$ is defined as

$$\mathrm{VaR}_\beta(X) \stackrel{\mathrm{def}}{=} \inf\{x \in \mathbb{R} : \mathbb{P}(x + X < 0) \leq 1 - \beta\}.$$

In addition, the risk measure conditional value-at-risk at confidence level $\gamma \in (0, 1)$ is defined as

$$\mathrm{CVaR}_\gamma(X) \stackrel{\mathrm{def}}{=} \frac{1}{1-\gamma} \int_\gamma^1 \mathrm{VaR}_\beta(X)\, \mathrm{d}\beta.$$

In the literature, the above risk measures are sometimes expressed in terms of the portfolio loss variable, namely positive values represent loss and negative values represent gain. We note that these definitions are equivalent. Intuitively, the value-at-risk denotes the maximum threshold of loss under a certain confidence level, and conditional value-at-risk is the conditional expectation of loss given that it exceeds such a threshold. Although more popularly used in practice, value-at-risk fails certain mathematical properties such as subadditivity, which contradicts with Markowitz's MPT and implies that diversification may not reduce investment risk. As a result, it is not a coherent risk measure. On the other hand, Pflug [42] proves that conditional value-at-risk is coherent and satisfies some extra properties such as convexity, monotonicity with respect to first-order stochastic dominance (FSD) and second-order monotonic dominance.

Theorem 2.5 [42]. *The conditional value-at-risk is a coherent risk measure.*

Therefore, we would like to minimize risk using the conditional value-at-risk at confidence level γ as the risk measure. We recall that $W = \{\boldsymbol{u} \in \mathbb{R}_+^K : \boldsymbol{u}^\top \mathbf{1} = 1\}$ is the set of possible portfolios. At each trial t, the learner would like to solve the following optimization problem:

$$\underset{\boldsymbol{u} \in W}{\text{minimize}} \quad CVaR_\gamma(\boldsymbol{u}^\top \boldsymbol{R}_t)$$

Note that as $\gamma \to 0$, the problem becomes minimizing expected loss, and as $\gamma \to 1$, it becomes minimizing the worst outcome. In this study, we use $\gamma = 0.95$. Rockafellar & Uryasev [43] provides a convenient method to solve this problem. Recall that we assume that both historical returns and present returns follow the same distribution; let $p(\boldsymbol{R}_t)$ be the density. Define the performance function as

$$F_\gamma(\boldsymbol{u}, \alpha) \stackrel{\mathrm{def}}{=} \alpha + \frac{1}{1-\gamma} \int_{\boldsymbol{R}_t \in \mathbb{R}^K} [-\boldsymbol{u}^\top \boldsymbol{R}_t - \alpha]^+ p(\boldsymbol{R}_t)\, \mathrm{d}\boldsymbol{R}_t,$$

where $[m]^+ \stackrel{\mathrm{def}}{=} \max\{m, 0\}$. Then, we have the following theorem.

Theorem 2.6 [43]. *The minimization of* $\mathrm{CVaR}_\gamma(\boldsymbol{u}^\top \boldsymbol{R}_t)$ *over* $\boldsymbol{u} \in W$ *is equivalent to the minimization of* $F_\gamma(\boldsymbol{u}, \alpha)$ *over all pairs of* $(\boldsymbol{u}, \alpha) \in W \times \mathbb{R}$. *Moreover, as* $F_\gamma(\boldsymbol{u}, \alpha)$ *is convex with respect to* $(\boldsymbol{u}, \alpha)$, *the loss function* $-\boldsymbol{u}^\top \boldsymbol{R}_t$ *is convex with respect to* $\boldsymbol{u}$ *and* W *is a convex set due to linearity, the minimization of* $F_\gamma(\boldsymbol{u}, \alpha)$ *is an instance of convex programming.*

Moreover, as the density $p(\boldsymbol{R}_t)$ is unknown, we would like to approximate the performance function using not only historical returns but also knowledge gained as we proceed in this learning process. From the received $H_{i,1}, \ldots, H_{i,\delta}$ for all i, we extract historical returns of our K assets $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_\delta \in \mathbb{R}^K$. Let

$\boldsymbol{R}_1, \ldots, \boldsymbol{R}_{t-1}$ be the $t-1$ trials of returns observed so far. Then our approximation of $F_\gamma(\boldsymbol{u}, \alpha)$ at trial t is the following convex and piecewise linear function

$$\tilde{F}_\gamma(\boldsymbol{u}, \alpha, t) \stackrel{\text{def}}{=} \alpha + \frac{1}{(\delta + t - 1)(1-\gamma)} \left[\sum_{s=1}^{\delta} [-\boldsymbol{u}^\top \boldsymbol{H}_s - \alpha]^+ + \sum_{s=1}^{t-1} [-\boldsymbol{u}^\top \boldsymbol{R}_s - \alpha]^+ \right]. \tag{2.3}$$

Note that the approximation function is implicitly also a function of the current trial t, hence we have added an extra parameter and denote it as $\tilde{F}_\gamma(\boldsymbol{u}, \alpha, t)$. As the learner proceeds in time, she accumulates data information and obtains a more and more precise approximation. As a result, the minimization of conditional value-at-risk is solved by convex programming and generates the following optimal solution. At each trial t, the risk-aware portfolio constructed according to equation (2.3) is

$$\boldsymbol{\omega}_t^C \stackrel{\text{def}}{=} \underset{(\boldsymbol{u}, \alpha) \in W \times \mathbb{R}}{\arg\min} \ \tilde{F}_\gamma(\boldsymbol{u}, \alpha, t). \tag{2.4}$$

Now, we have found both the single-asset multi-armed bandit portfolio by (2.2) and the risk-aware portfolio by (2.4). Note that they are dynamic and update based on the learner's accumulated knowledge. For each trial t, the learner combines them with a factor $\lambda \in [0, 1]$ to form the balanced portfolio

$$\boldsymbol{\omega}_t^* \stackrel{\text{def}}{=} \lambda \boldsymbol{\omega}_t^M + (1 - \lambda) \boldsymbol{\omega}_t^C. \tag{2.5}$$

In particular, λ is the proportion of wealth invested in the single-asset multi-armed bandit portfolio and $1 - \lambda$ is the proportion invested in the risk-aware portfolio. The value of λ denotes the risk preference of the learner. As $\lambda \to 1$, our algorithm reverts to the UCB1 policy, whereas for $\lambda \to 0$, it becomes the minimization of conditional value-at-risk. Therefore, the commonly discussed trade-off between reward and risk is illustrated here in the choice of λ. Finally, the following algorithm summarizes our sequential portfolio selection algorithm.

Algorithm 1: Our proposed sequential portfolio selection algorithm

Input: K, γ, λ
Select K peripheral assets from the market according to §2.2;
for $t = 1, \ldots, N$ **do**
  Compute the single-asset multi-armed bandit portfolio $\boldsymbol{\omega}_t^M$ by (2.2);
  Compute the risk-aware portfolio $\boldsymbol{\omega}_t^C$ at confidence level γ by (2.4);
  Select the combined portfolio $\boldsymbol{\omega}_t^*$ with a factor λ by (2.5);
  Observe returns $\boldsymbol{R}_t$ and update accumulated knowledge for (2.2) and (2.4);
  Receive portfolio reward $\boldsymbol{\omega}_t^{*\top} \boldsymbol{R}_t$;
end

3. Results

In this section, we design experiments and report the performance of the proposed algorithm (algorithm 1) in comparison with several benchmarks.

3.1. Monte Carlo simulation method

For simplicity, we consider stocks as our assets and adopt the Black–Scholes model [44] to simulate stock prices as geometric Brownian motion (GBM) paths. As a Nobel Prize-winning model, it provides a partial differential equation to price a European option by computing the initial wealth for perfectly hedging a short position in that option. The underlying asset, usually a stock, is modelled to follow a GBM. Although this assumption may not hold perfectly in reality, it provides an extremely convenient and popularly used method to simulate any number of stock paths. For our purpose, as we never make any assumption on the dependency of asset returns, we consider the general case where stock paths can be correlated as it is almost always the case in the financial market. We use definitions similar to ch. 4 of Shreve [45] and describe our method below.

Definition 3.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The stock price $P_i(t)$ is said to follow a GBM if it satisfies the following stochastic differential equation:

$$\mathrm{d}P_i(t) = \alpha_i P_i(t)\,\mathrm{d}t + \sigma_i P(t)\,\mathrm{d}W_i(t),$$

where $W_i(t)$ is a Brownian motion, α_i is the drift and σ_i is the volatility.

Definition 3.2. Two stock paths $P_i(t)$ and $P_j(t)$ modelled by GBMs are correlated if their associated Brownian motions satisfy

$$\mathrm{d}W_i(t)\,\mathrm{d}W_j(t) = \rho_{i,j} \cdot \mathrm{d}t$$

for some non-zero constant $\rho_{i,j} \in [-1, 1]$ where $\rho_{i,i} = \rho_{j,j} = 1$.

Proposition 3.3. *For two correlated stock prices $P_i(t)$ and $P_j(t)$ that satisfy $\mathrm{d}W_i(t)\,\mathrm{d}W_j(t) = \rho_{i,j} \cdot \mathrm{d}t$, the following properties hold*:

— $\mathbb{E}[W_i(t)W_j(t)] = \rho_{i,j} \cdot t$
— $\mathrm{Cov}[W_i(t), W_j(t)] = \rho_{i,j} \cdot t$
— $\mathrm{Cov}[\sigma_i W_i(t), \sigma_j W_j(t)] = \sigma_i \sigma_j \rho_{i,j} \cdot t$,

where σ_i and σ_j are volatility parameters of $P_i(t)$ and $P_j(t)$, respectively.

Proof. We prove the first claim and the rest follow immediately after some computations. By the Itô–Doeblin formula, which can be found in Shreve [45], we have

$$\mathrm{d}(W_i(t)W_j(t)) = W_i(t)\,\mathrm{d}W_j(t) + W_j(t)\,\mathrm{d}W_i(t) + \rho_{i,j} \cdot \mathrm{d}t.$$

Integrating on both sides, we have

$$W_i(t)W_j(t) = \int_0^t W_i(t)\,\mathrm{d}W_j(t) + \int_0^t W_j(t)\,\mathrm{d}W_i(t) + \rho_{i,j} \cdot t.$$

By the Martingale property of Itô integrals, we simply take the expectation on both sides to obtain $\mathbb{E}[W_i(t)W_j(t)] = \rho_{i,j} \cdot t$. ∎

Recall that we have K stocks whose prices $P_1(t), \ldots, P_K(t)$ are modelled by correlated GBMs. By definition, they must satisfy the following two equations:

$$\frac{\mathrm{d}P_i(t)}{P_i(t)} = \alpha_i\,\mathrm{d}t + \sigma_i\,\mathrm{d}W_i(t) \tag{3.1}$$

and

$$\mathrm{d}W_i(t)\,\mathrm{d}W_j(t) = \rho_{i,j} \cdot \mathrm{d}t. \tag{3.2}$$

In particular, the solution to equation (3.1) can be expressed as follows [46]. For any time $u < l$, we have

$$P_i(l) = P_i(u) \cdot \exp\left\{\left(\alpha_i - \tfrac{1}{2}{\sigma_i}^2\right)(l-u) + \sigma_i(W_i(l) - W_i(u))\right\}. \tag{3.3}$$

We first would like to express the scaled correlated Brownian motions $\sigma_i W_i(t)$ using independent ones. By proposition 3.3, we have the following instantaneous covariance matrix:

$$\Theta = \begin{bmatrix} {\sigma_1}^2 & \sigma_1\sigma_2\rho_{1,2} & \ldots & \sigma_1\sigma_K\rho_{1,K} \\ \sigma_2\sigma_1\rho_{2,1} & {\sigma_2}^2 & \ldots & \sigma_2\sigma_K\rho_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_K\sigma_1\rho_{K,1} & \sigma_K\sigma_2\rho_{K,2} & \ldots & {\sigma_K}^2 \end{bmatrix}$$

As Θ has to be symmetric and positive definite, it has a square root and we apply Cholesky decomposition to find the matrix A such that $AA^{\mathrm{T}} = \Theta$. By Shreve [45], there exists K independent

Brownian motions $X_1(t), \ldots, X_K(t)$ such that

$$\sigma_i W_i(t) = \sum_{m=1}^{K} A_{i,m} X_m(t).$$

Then equation (3.1) becomes

$$\frac{\mathrm{d}P_i(t)}{P_i(t)} = \alpha_i \,\mathrm{d}t + \sum_{m=1}^{K} A_{i,m} \,\mathrm{d}X_m(t), \tag{3.4}$$

and equation (3.3) becomes, for any time $u < l$,

$$P_i(l) = P_i(u) \exp\left\{\left(\alpha_i - \frac{1}{2}{\sigma_i}^2\right)(l-u) + \sum_{m=1}^{K} A_{i,m}(X_m(l) - X_m(u))\right\}. \tag{3.5}$$

As each Brownian motion $X_m(t)$ for $m \in [1, K]$ above is independent and the increment $X_m(l) - X_m(u)$ is Gaussian with mean 0 and variance $l - u$, let $\boldsymbol{Z}(t) = (Z_1(t), \ldots, Z_K(t))^\top$ be standard multivariate Gaussian, then equation (3.5) becomes

$$P_i(l) = P_i(u) \exp\left\{\left(\alpha_i - \frac{1}{2}{\sigma_i}^2\right)(l-u) + \sqrt{l-u} \sum_{m=1}^{K} A_{i,m} Z_m(l)\right\}. \tag{3.6}$$

Therefore, at each time we can conveniently generate a sample from $\boldsymbol{Z}(t)$ to compute the price increment. Specifically, equation (3.6) leads to the following recursive algorithm that can also be found in Glasserman [46]. For $0 = t_0 < t_1 < \cdots < t_\infty$, we have

$$P_i(t_{s+1}) = P_i(t_s) \cdot \exp\left\{\left(\alpha_i - \frac{1}{2}{\sigma_i}^2\right)(t_{s+1} - t_s) + \sqrt{t_{s+1} - t_s} \sum_{m=1}^{K} A_{i,m} Z_m(t_{s+1})\right\}.$$

Also note that when the paths are independent, $\mathrm{d}W_i(t)\,\mathrm{d}W_j(t) = \boldsymbol{\delta}_{i,j}\,\mathrm{d}t$, where $\boldsymbol{\delta}_{i,j}$ is the Kronecker delta function, and the covariance matrix Θ is diagonal. In this special case, it is equivalent to compute K paths separately in the one-dimensional space. For our purpose, we first find some appropriate covariance matrix and generate K price paths following the above algorithm. We then uniformly divide the total time horizon into $\delta + N$ trials and use the prices at the beginning and end of each trial to calculate return, which was defined earlier as the log price ratio. We run our sequential portfolio selection algorithm on these data and compare the performance with four benchmark portfolios, namely UCB1 (2.2), risk-aware portfolio (2.4), ϵ-greedy and the equally weighted portfolio.

3.2. Simulation results

After we repeatedly generate price paths and compare the performance, we can see that the results agree well with our prediction (figure 2). The UCB1 portfolio almost always achieves the most cumulative wealth but has high variations in its path. On the other hand, the risk-aware portfolio achieves a relatively low cumulative wealth but also has low variations. As a result, our combined portfolio achieves a middle ground between the two extremes of maximizing reward and minimizing risk. For example, figure 2*a–c* illustrate a typical simulation, where figure 2*a* shows $K = 5$ GBM paths, figure 2*b* shows the optimality of UCB1 compared to ϵ-greedy and figure 2*c* shows the cumulative wealth at the end of $N = 200$ trials. With an initial wealth of 1 and $\lambda = 0.9$, the cumulative wealth is 2.1615 for UCB1, 2.1024 for combined portfolio, 1.9168 for ϵ-greedy, 1.6355 for the risk-aware portfolio and 1.4640 for the equally weighted portfolio.

In addition, we observe that when the market is volatile and when different stock paths are similar in expectation, it takes more trials for the UCB1 policy to reach optimality (figure 2*d–f*). In this case, the risk-aware portfolio achieves the most cumulative wealth with a similarly low variation in its path. Different from the simulation presented in figure 2*a–c*, where the volatility parameters of GBMs are bounded in the interval $[0.02, 0.025]$, we now choose values from the interval $[0.03, 0.035]$ for figure 2*d–f*. Specifically, figure 2*d–f* demonstrate such a simulation, where figure 2*d* shows the GBM paths, figure 2*e* shows the suboptimality of UCB1 and figure 2*f* shows the cumulative wealth at the end of 200 trials. With an initial wealth of 1 and $\lambda = 0.9$, the cumulative wealth is 1.5412 for the risk-aware portfolio, 1.4409 for the combined portfolio, 1.4294 for UCB1, 1.4132 for the equally-weighted portfolio and finally, 1.3298 for ϵ-greedy.

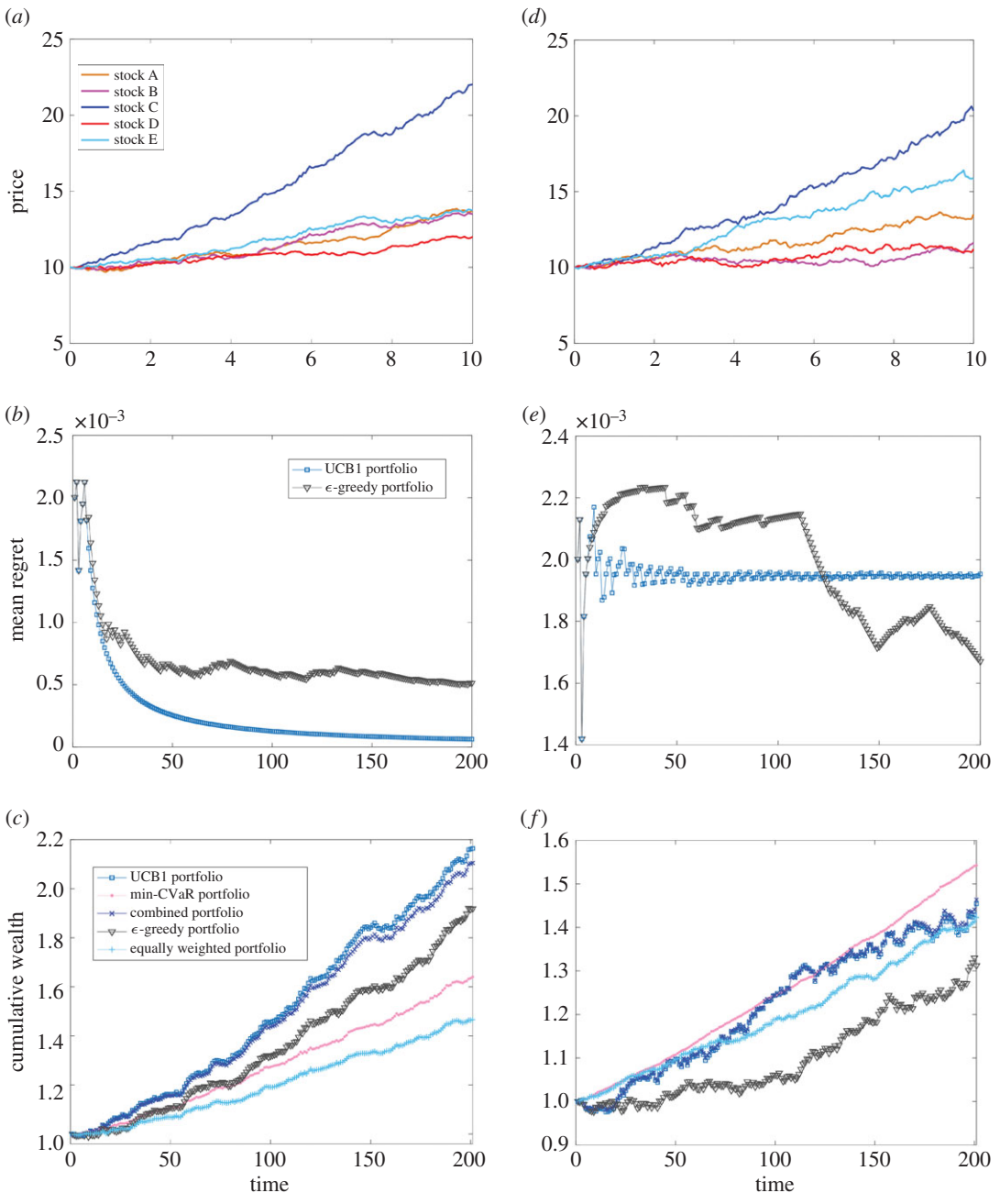

Figure 2. Combined sequential portfolio selection algorithm can achieve a balance between risk and return. (*a*,*d*) The simulated stock paths based on the GBM. (*b*,*e*) The performance of two portfolio selection algorithms, UCB1 versus ϵ-greedy. Panels (*c*,*f*) compare the cumulative wealth obtained with our sequential portfolio selection algorithm that combines the single-asset multi-armed bandit portfolio by (2.2) and the risk-aware portfolio by (2.4) with the other four benchmarks of portfolio selection algorithms. To quantify and compare the role of volatility in the performance of portfolio selection algorithms, we present the simulation results of low volatility in (*a*)–(*c*) and high volatility in (*d*)–(*f*). Parameters: the same vector (0.04, 0.035, 0.08, 0.02, 0.03) for drift terms α_i is used for simulating the stock paths in (*a*) and (*d*). For each trial, the volatility terms σ_i are uniformly and randomly generated from the interval [0.02, 0.025] in (*a*) and from the interval [0.03, 0.035] in (*d*). $\lambda = 0.9$.

From the above discussion, it is evident that the value of λ is vital to the performance of our sequential portfolio selection algorithm and should be determined based on the market condition. In particular, Way *et al.* [47] discusses the trade-off between specialization to achieve high rewards and diversification to hedge against risk, and similarly shows that such choice depends on the underlying parameters and initial conditions.

4. Discussion and conclusion

In this paper, we have studied the multi-armed bandit problem as a mathematical model for sequential decision-making under uncertainty. In particular, we focus on its application in financial markets and construct a sequential portfolio selection algorithm. We first apply graph theory and select the peripheral assets from the market to invest. Then at each trial, we combine the optimal multi-armed bandit policy with the minimization of a coherent risk measure. By adjusting the parameter, we are able to achieve the balance between maximizing reward and minimizing risk. We adopt the Black–Scholes model to repeatedly simulate stock paths and observe the performance of our algorithm. We conclude that the results agree well with our prediction when the market is stable. In addition, when the market is volatile, risk awareness becomes more crucial to achieving high performance. Therefore, parameter selection should be based on the market condition.

For future research, one may consider the optimal selection of the parameter λ for combining the two portfolios. One may also consider portfolio selection strategies based on the MDP, which is a generalization of the multi-armed bandit to multiple states. In addition, one may pay more attention to a chaotic market environment where stock paths can be affected by various factors instead of simply following a stochastic process. For example, Junior & Mart [48] uses random matrix theory and transfer entropy to show that news articles can possibly affect the market. Finally, one may consider transaction costs and market liquidity. For example, Reiter *et al.* [49] illustrates the trade-off between reward and cost in a biological auction setting and might provide some important insights for the researcher.

Data accessibility. Our data and simulation codes are deposited at Dryad: https://doi.org/10.5061/dryad.h628h [50].

Authors' contributions. X.H. & F.F. conceived the project, X.H. performed analyses and simulations, X.H. & F.F. analysed results. X.H. wrote the first draft of the main text. Both the authors reviewed the manuscript.

Competing interests. The authors declare no competing financial interests.

Funding. Financial support came from the Dartmouth Faculty Startup Fund and Walter & Constance Burke Research Initiation Award. X.H. is thankful for financial support from the National Science Foundation and Dartmouth College. F.F. is grateful for support from the Dartmouth Faculty Startup Fund, Walter & Constance Burke Research Initiation Award, NIH under grant no. C16A12652 (A10712), and DARPA under grant no. D17PC00002-002.

References

1. Markowitz H. 1952 Portfolio selection. *J. Finance* **7**, 77–91. (doi:10.1111/j.1540-6261.1952.tb01525.x)
2. Cover TM. 1991 Universal portfolios. *Math. Finance* **1**, 1–29. (doi:10.1111/j.1467-9965.1991.tb00002.x)
3. Helmbold DP, Schapire RE, Singer Y, Warmuth MK. 1998 On-line portfolio selection using multiplicative updates. *Math. Finance* **8**, 325–347. (doi:10.1111/1467-9965.00058)
4. Li B, Hoi SC. 2014 Online portfolio selection: a survey. *ACM Comput. Surv. (CSUR)* **46**, 35. (doi:10.1145/2512962)
5. Heaton JB, Polson NG, Witte JH. 2017 Deep learning for finance: deep portfolios. *Appl. Stochastic Models Bus. Ind.* **33**, 3–12. (doi:10.1002/asmb.2209)
6. Song Q, Liu A, Yang SY. 2017 Stock portfolio selection using learning-to-rank algorithms with news sentiment. *Neurocomputing* **264**, 20–28. (doi:10.1016/j.neucom.2017.02.097)
7. Ghavamzadeh M, Mannor S, Pineau J, Tamar A. 2015 Bayesian reinforcement learning: a survey. *Found. Trends® Mach. Learn.* **8**, 359–483. (doi:10.1561/2200000049)
8. Bubeck S, Cesa-Bianchi N. 2012 Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends® Mach. Learn.* **5**, 1–22. (doi:10.1561/2200000024)
9. Lai TL, Robbins H. 1985 Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6**, 4–22. (doi:10.1016/0196-8858(85)90002-8)
10. Kuleshov V, Precup D. 2014 Algorithms for multi-armed bandit problems. (https://arxiv.org/abs/1402.6028).
11. Auer P, Cesa-Bianchi N, Fischer P. 2002 Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47**, 235–256. (doi:10.1023/A:1013689704352)
12. Agrawal S, Goyal N. 2013 Thompson sampling for contextual bandits with linear payoffs. In *Int. Conf. on Machine Learning, Atlanta, GA, 13 February*, pp. 127–135. PMLR. See http://proceedings.mlr.press/v28/agrawal13.html.
13. Koulouriotis DE, Xanthopoulos A. 2008 Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Appl. Math. Comput.* **196**, 913–922. (doi:10.1016/j.amc.2007.07.043)
14. Sani A, Lazaric A, Munos R. 2012 Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems* (eds F Pereira, CJC Burges, L Bottou, KQ Weinberger), pp. 3275–3283. Curran Associates, Inc. See http://papers.nips.cc/paper/4753-risk-aversion-in-multi-armed-bandits.
15. Vakili S, Zhao Q. 2016 Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE J. Sel. Top. Signal Process.* **10**, 1093–1111. (doi:10.1109/JSTSP.2016.2592622)
16. Vakili S, Zhao Q. 2015 Mean-variance and value at risk in multi-armed bandit problems. In *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Urbana-Champaign, IL, 29 September–2 October*, pp. 1330–1335. IEEE. See http://ieeexplore.ieee.org/abstract/document/7447162/.
17. Zimin A, Ibsen-Jensen R, Chatterjee K. 2014 Generalized risk-aversion in stochastic multi-armed bandits. (https://arxiv.org/abs/1405.0833).
18. Galichet N, Sebag M, Teytaud O. 2013 Exploration versus exploitation versus safety: risk-aware multi-armed bandits. In *Asian Conference on Machine Learning, Canberra, Australia, 21 October*, pp. 245–260. See https://arxiv.org/abs/1401.1123.
19. Shen W, Wang J, Jiang YG, Zha H. 2015 Portfolio choices with orthogonal bandit learning. In *IJCAI, Buenos Aires, Argentina, 25 July*, p. 974. See http://dl.acm.org/citation.cfm?id=2832384.
20. Ohtsuki H, Hauert C, Lieberman E, Nowak MA. 2006 A simple rule for the evolution of cooperation on graphs. *Nature* **441**, 502–505. (doi:10.1038/nature04605)
21. Fu F, Nowak MA. 2013 Global migration can lead to stronger spatial selection than local migration. *J. Stat. Phys.* **151**, 637–653. (doi:10.1007/s10955-012-0631-6)
22. Tarnita CE, Ohtsuki H, Antal T, Fu F, Nowak MA. 2009 Strategy selection in structured populations. *J. Theor. Biol.* **259**, 570–581. (doi:10.1016/j.jtbi.2009.03.035)
23. Szolnoki A, Perc M. 2015 Antisocial pool rewarding does not deter public cooperation. *Proc. R. Soc. B* **282**, 20151975. (doi:10.1098/rspb.2015.1975)

24. Chen X, Zhang Y, Huang TZ, Perc M. 2014 Solving the collective-risk social dilemma with risky assets in well-mixed and structured populations. *Phys. Rev. E* **90**, 052823. (doi:10.1103/PhysRevE.90.052823)
25. Szolnoki A, Perc M. 2015 Conformity enhances network reciprocity in evolutionary social dilemmas. *J. R. Soc. Interface* **12**, 20141299. (doi:10.1098/rsif.2014.1299)
26. Aste T, Shaw W, Di Matteo T. 2010 Correlation structure and dynamics in volatile markets. *New J. Phys.* **12**, 085009. (doi:10.1088/1367-2630/12/8/085009)
27. Mantegna RN. 1999 Hierarchical structure in financial markets. *Eur. Phys. J. B* **11**, 193–197. (doi:10.1007/s100510050929)
28. Bonanno G, Caldarelli G, Lillo F, Mantegna RN. 2003 Topology of correlation-based minimal spanning trees in real and model markets. *Phys. Rev. E* **68**, 046130. (doi:10.1103/PhysRevE.68.046130)
29. Bonanno G, Caldarelli G, Lillo F, Micciche S, Vandewalle N, Mantegna RN. 2004 Networks of equities in financial markets. *Eur. Phys. J. B* **38**, 363–371. (doi:10.1140/epjb/e2004-00129-6)
30. Mantegna RN, Stanley HE. 1999 *Introduction to econophysics: correlations and complexity in finance*. Cambridge, UK: Cambridge University Press.
31. Ledoit O, Wolf M. 2004 A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88**, 365–411. (doi:10.1016/S0047-259X(03)00096-4)
32. Schwarcz SL. 2008 Systemic risk. *Geo. LJ* **97**, 193–249. See https://www.iiiglobal.org/sites/default/files/systemicrisk.pdf.
33. Kritzman M, Li Y, Page S, Rigobon R. 2011 Principal components as a measure of systemic risk. *J. Portfolio Manag.* **37**, 112–126. (doi:10.3905/jpm.2011.37.4.112)
34. Drozdz S, Gümrmer F, Górski AZ, Ruf F, Speth J. 2000 Dynamics of competition between collectivity and noise in the stock market. *Phys. A: Stat. Mech. Appl.* **287**, 440–449. (doi:10.1016/S0378-4371(00)00383-6)
35. Huang X, Vodenska I, Havlin S, Stanley HE. 2013 Cascading failures in bi-partite graphs: model for systemic risk propagation. *Sci. Rep.* **3**, 1219. (doi:10.1038/srep01219)
36. Onnela JP, Chakraborti A, Kaski K, Kertesz J, Kanto A. 2003 Dynamics of market correlations: taxonomy and portfolio analysis. *Phys. Rev. E* **68**, 056110. (doi:10.1103/PhysRevE.68.056110)
37. Pozzi F, Di Matteo T, Aste T. 2013 Spread of risk across financial markets: better to invest in the peripheries. *Sci. Rep.* **3**, 1665. (doi:10.1038/srep01665)
38. Ren F, Lu YN, Li SP, Jiang XF, Zhong LX, Qiu T. 2017 Dynamic portfolio strategy using clustering approach. *PLoS ONE* **12**, e0169299. (doi:10.1371/journal.pone.0169299)
39. Freeman LC. 1978 Centrality in social networks conceptual clarification. *Social Netw.* **1**, 215–239. (doi:10.1016/0378-8733(78)90021-7)
40. Artzner P, Delbaen F, Eber JM, Heath D. 1999 Coherent measures of risk. *Math. Fin.* **9**, 203–228. (doi:10.1111/1467-9965.00068)
41. Bäuerle N, Rieder U. 2011 *Markov decision processes with applications to finance*. Berlin, Germany: Springer Science & Business Media.
42. Pflug GC. 2000 Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic constrained optimization* (ed. S Uryasev), pp. 272–281. Boston, MA: Springer. https://link.springer.com/chapter/10.1007/978-1-4757-3150-7_15.
43. Rockafellar RT, Uryasev S. 2000 Optimization of conditional value-at-risk. *J. Risk* **2**, 21–42. (doi:10.21314/JOR.2000.038)
44. Black F, Scholes M. 1973 The pricing of options and corporate liabilities. *J. Political Econ.* **81**, 637–654. (doi:10.1086/260062)
45. Shreve SE. 2004 *Stochastic calculus for finance II: continuous-time models*. Berlin, Germany: Springer Science & Business Media.
46. Glasserman P. 2013 *Monte Carlo methods in financial engineering*. Berlin, Germany: Springer Science & Business Media.
47. Way R, Lafond F, Farmer JD, Lillo F, Panchenko V. 2017 Wright meets Markowitz: How standard portfolio theory changes when assets are technologies following experience curves. (http://arxiv.org/abs/1705.03423)
48. Junior LS, Mart AM. 2017 Correlations and flow of information between The New York Times and Stock Markets. (http://arxiv.org/abs/1707.05778)
49. Reiter JG, Kanodia A, Gupta R, Nowak MA, Chatterjee K. 2015 Biological auctions with multiple rewards. *Proc. R. Soc. B* **282**, 20151041. (doi:10.1098/rspb.2015.1041)
50. Huo X, Fu F. 2017 Data from: Risk-aware multi-armed bandit problem with application to portfolio selection. Dryad Digital Repository. (doi:10.5061/dryad.h628h)