Portfolio Choices with Orthogonal Bandit Learning

Weiwei Shen[†], Jun Wang[‡], Yu-Gang Jiang[♯], Hongyuan Zha[♮]
[†]GE Global Research Center, Niskayuna, NY, USA, weiwei.shen@ge.com
[‡]Alibaba Group, Seattle, WA, USA, j.wang@alibaba-inc.com
[♯]School of Computer Science, Fudan University, Shanghai, China, ygj@fudan.edu.cn
[♮]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA, zha@cc.gatech.edu

Abstract

The investigation and development of new methods from diverse perspectives to shed light on portfolio choice problems has never stagnated in financial research. Recently, multi-armed bandits have drawn intensive attention in various machine learning applications in online settings. The tradeoff between *exploration* and *exploitation* to maximize rewards in bandit algorithms naturally establishes a connection to portfolio choice problems. In this paper, we present a bandit algorithm for conducting online portfolio choices by effectually exploiting correlations among multiple arms. Through constructing orthogonal portfolios from multiple assets and integrating with the upper confidence bound bandit framework, we derive the optimal portfolio strategy that represents the combination of passive and active investments according to a risk-adjusted reward function. Compared with oft-quoted trading strategies in finance and machine learning fields across representative real-world market datasets, the proposed algorithm demonstrates superiority in both risk-adjusted return and cumulative wealth.

1 Introduction

Portfolio choice problems have had a profound influence on the finance industry ranging from pension fund to mutual fund management and from insurance to corporate risk management [Brandt, 2010]. Modern portfolio theory and analysis build upon the seminal work of Markowitz [Markowitz, 1952]. However, motivated by its noticeably poor performance in out-of-sample settings [Broadie, 1993], researchers have expended unremitting efforts of investigating novel approaches to attack portfolio choice problems, such as numerous variants of the Markowitz framework [DeMiguel *et al.*, 2009a], the optimal growth portfolio based on the Kelly criterion [Thorp, 1971], and the linear programming based portfolio optimization [Konno and Yamazaki, 1991].

Meanwhile, the globalization and the rapid growth of market integration yield massive amounts of data in the finance industry, which promotes the study of advanced data analysis tools. In particular, as leading-edge analytical techniques, machine learning algorithms emphasize on on-line building, updating and applying models based on its efficient automated processing of large datasets. Accordingly, machine learning researchers have paid significant efforts to design real time data stream based trading strategies [Helmbold *et al.*, 1998; Blum and Kalai, 1999; Borodin *et al.*, 2004; Agarwal *et al.*, 2006; Györfi *et al.*, 2006; Li and Hoi, 2012]. Among them, the Kelly criterion purporting to achieve optimal return growth shows a popular choice. For example, one representative work captures and utilizes the moving average reversion phenomena of the stock market to maximize the return growth on investment [Li and Hoi, 2012]. A comprehensive survey about on-line portfolio strategies can be referred to [Li and Hoi, 2014], and the extensive references therein.

As a potent tool for designing on-line sequential decision strategies, the multi-armed bandit problem has been studied since the early 1950's [Robbins, 1952]. The pivotal idea of bandit learning is to acquire new information while optimizing rewards based on existing knowledge, which is known as the tradeoff between *exploitation* and *exploration* in reinforcement learning. Such a tradeoff naturally establishes a connection to the sequential decision process in portfolio choice problems. The illuminating paper by [Hoffman *et al.*, 2011] adopts a multi-armed bandit strategy to design the portfolio of acquisition functions in Bayesian optimization. However, standard multi-armed bandits assume the rewards of each arm are drawn from i.i.d. (independent and identically distributed) random variables, whereas in practice financial asset returns are generally correlated. Moreover, standard bandit learning attempts to choose the best arm for action, while in portfolio choice problems investors tend to select multiple assets for investment. Although combinatorial multi-armed bandit algorithms have been proposed to select multiple arms, it makes binary decisions of selecting arms and equally distributes investments among them [Chen *et al.*, 2013]. In contrast, the crux of portfolio choice problems lies in determining the optimal distributing weights among assets.

To grapple with those challenges in applying conventional bandit algorithms to portfolio choice problems, in this paper, we provide an orthogonal bandit learning algorithm to effectively make portfolio choices. In particular, we take advantage of the principal component decomposition to orthogonalize correlated assets, choose the Sharpe ratio as the risk-adjusted reward function in the upper confidence bound bandit framework to make investment decisions, and combine the generated passive and active portfolio weights to construct a low-risk portfolio. Further, to validate the proposed strategy, we evaluate the performance from both risk-adjusted return and cumulative wealth. Our extensive empirical studies and comparisons over several market datasets lucidly illustrate the superiority of the proposed strategy. We believe this work is a step in the development of leveraging on-line machine learning algorithms for portfolio choice problems.

The remainder of the paper is organized as follows. Section 2 briefly reviews the background of bandit learning. In Section 3 we propose our portfolio strategy based on orthogonal bandit learning. Section 4 covers our experiments and comparative studies, including the discussion of the datasets, the evaluation metrics and the performance demonstration. Finally, Section 5 concludes.

2 Background

In this section, we first review the problem setting and the solution of the standard multi-armed bandit. Then we discuss major difficulties of applying the standard bandit learning to portfolio choice problems.

Given n arms that represent n actions or assets and time steps t_k, $k = 1, \dots, m$, at time t_k each arm receives a bounded real-valued reward $r_i(t_k)$, $i = 1, \dots, n$. In practice, each arm may receive a fixed reward of the value 1 with certain probability, or otherwise a reward of 0, which is called the Bernoulli multi-armed bandit. Further, the objective of bandit learning is to choose a series of arms, one for each time, to maximize the total rewards or minimize the regret. For stochastic bandits, denote by ν_i the expectation of the reward from the i-th arm. The largest reward is $\nu^* = \max_{i=1,\dots,n} \nu_i$ and the maximum reward after m plays is $m\nu^*$. Thus, the pseudo regret after m plays is defined by

$$m\nu^* - \sum_{k=1}^{m} \mathbb{E}[r_{i_k}(t_k)], \tag{1}$$

where i_k is the index of selected arm at time t_k and $r_{i_k}(t_k)$ is the corresponding reward [Lai and Robbins, 1985]. Besides this standard bandit setting, various variants have been studied, such as non-stochastic bandits [Auer *et al.*, 2002], bandits in an adversarial environment [Auer *et al.*, 1995], and bandits in a contextual setting [Langford and Zhang, 2008]. Analyses of the reward and regret over numerous types of bandits can be found in the survey by [Bubeck and Cesa-Bianchi, 2012].

Further, to solve the multi-armed bandit problem, one straightforward policy is called the ϵ-greedy approach. At each time t_k, the player chooses the arm with the highest reward with a probability $1 - \epsilon_{t_k}$, or randomly chooses an arm with a probability of ϵ_{t_k}, where the two parts are corresponding to "exploitation" and "exploration", respectively. If we set $\epsilon_{t_k} = 12/(d^2 k)$ with $0 < d < \Delta^*$, the accumulated regret until the m-th step of the ϵ-greedy strategy is bounded by $\mathcal{O}(\Delta^* n \ln(m)/k)$, where $\triangle^*$ is the gap between the best expected reward and the expected reward [Auer *et al.*, 2002]. On the other hand, different from the randomized policy as the ϵ-greedy approach, the upper confidence bounds (UCB) strategy has emerged as another popular choice for multi-armed bandit problems [Lai and Robbins, 1985]. After playing each arm once, at each time t_k the best arm $i^*(t_k)$ is selected according to the following objective:

$$i^*(t_k) = \underset{i=1,\dots,n}{\arg\max}\, \bar{r}_i(t_k) + \sqrt{\frac{2\ln(k)}{k_i}}, \tag{2}$$

where $\bar{r}_i(t_k)$ is the mean reward of the i-th arm and k_i is the number of times that the i-th arm has been played so far. The second part of the above selection rule relates to the one-sided confidence interval for the average reward. It has been shown that the total regret at time t_m of the UCB policy is bounded by the following quantity [Auer *et al.*, 2002]:

$$\frac{8n}{\triangle^*}\ln(m) + 5n. \tag{3}$$

Further, Thompson Sampling and its variants have been another popular approach [Scott, 2010; Graepel *et al.*, 2010]. A recent empirical study shows that Thompson sampling outperforms other peer methods in several real-world benchmarks [Chapelle and Li, 2011]. The empirical success sparks the great use of bandit learning algorithms in a wide range of applications, such as clinic trials [Press, 2009], web analytics [Graepel *et al.*, 2010], algorithm selections [Gagliolo and Schmidhuber, 2011], and news recommendations [Li *et al.*, 2010]. In addition, the role of risk in bandit and on-line learning has started to be acknowledged and studied [Even-Dar *et al.*, 2006; Sani *et al.*, 2012; Maillard, 2013].

Given those successful examples, however, how to apply bandit learning to portfolio choice problems is less investigated. The intrinsic distinctions of financial assets and investment conventions call for a novel approach of altering and morphing the standard bandits. First, the standard multi-armed bandits assume i.i.d. rewards for each arm, which in principle does not hold for financial asset returns. Second, the standard multi-armed bandits tend to select the best arm, while the portfolio choices often aim to select a set of assets for investment. Third, the standard multi-armed bandits rest on the reward mean as the objective function, whereas in finance investors focus on the risk-adjusted return. Fourth, the standard multi-armed bandits assume no available historical data, whereas public traded financial assets generally have sufficient amounts of data. As such, we present a new bandit learning algorithm in the UCB framework to address those challenges in on-line portfolio choices.

3 Methodology

In this section, we first introduce the notations and financial terms in this paper. Then we will derive the proposed orthogonal bandit algorithm to make portfolio choices.

3.1 Notation

We consider a frictionless, self-financing, discrete-time and finite horizon investment environment. The trading periods consist of $t_k = k\Delta t$, $k = 0, \dots, m$. In particular, Δt could represent one week or one month in our study. We use k for short as the time step index to indicate the trading period at time t_k hereafter. We assume that investors have access to n risky assets with the gross return from time t_{k-1} to t_k as

$\mathbf{R}_k = (R_{k,1}, \dots, R_{k,i}, \dots, R_{k,n})^\top$. The gross return $R_{k,i}$ for the i-th asset is computed by $R_{k,i} = S_{k,i}/S_{k-1,i}$, where $S_{k,i}$ and $S_{k-1,i}$ represent the prices of the i-th asset at time steps t_k and t_{k-1}, respectively. In time, investors decide how to invest in those assets. The investment decision over a set of assets at time t_k is determined by a column vector representing portfolio weights $\boldsymbol{\omega}_k = (\omega_{k,1}, \dots, \omega_{k,i}, \dots, \omega_{k,n})^\top$, where the portfolio weight $\omega_{k,i}$ specifies the invested percentage of wealth in the i-th asset. The sum of all the portfolio weights always equals one, i.e.,

$$\boldsymbol{\omega}_k^\top \mathbf{1} = \sum_{i=1}^{n} \omega_{k,i} = 1, \tag{4}$$

where $\mathbf{1}$ denotes a column vector with ones as its entities. In addition, $\omega_{k,i} > 0$ indicates that investors take a *long* position of the i-th asset, and $\omega_{k,i} < 0$ represents that investors take a *short sale* position of the i-th asset. In particular, a short sale position means that investors first borrow an asset for sale and then invest its liquidation in other assets. The negative sign of the short sale weight reveals that investors will confront with a loss if the price of this asset starts to mount.

3.2 Bandit with Orthogonal Portfolio

We denote by $\boldsymbol{\Sigma}_k$ the positive definite covariance matrix of the n asset returns $\mathbf{R}_k$ at time t_k. The principal component decomposition of the covariance matrix $\boldsymbol{\Sigma}_k$ derives:

$$\boldsymbol{\Sigma}_k = \mathbf{H}_k \boldsymbol{\Lambda}_k \mathbf{H}_k^\top, \tag{5}$$

where the diagonal matrix $\boldsymbol{\Lambda}_k$ contains the eigenvalues of $\boldsymbol{\Sigma}_k$ in decreasing order, i.e., $\lambda_{k,1} > \lambda_{k,2} > \dots \lambda_{k,n} > 0$ and the columns of the orthogonal matrix $\mathbf{H}_k = (H_{k,1}, \dots, H_{k,i}, \dots, H_{k,n})$ are the corresponding eigenvectors of $\boldsymbol{\Sigma}_k$. In particular, the principal eigenvectors define a set of n uncorrelated portfolios with the return $\mathbf{H}_k^\top \mathbf{R}_k$; the eigenvalues that are all nonnegative represent the variances of those uncorrelated portfolios; the orthonormal property of the matrix $\mathbf{H}_k$ implies:

$$H_{k,i}^\top H_{k,j} = \delta_{i,j} \tag{6}$$

where $\delta_{i,j}$ is the Kronecker delta function.

Further, to convert the principal eigenvectors into the portfolio weights satisfying the condition (4), we normalize each eigenvector by its sum and define $\tilde{\mathbf{H}}_k = (\tilde{H}_{k,1}, \dots, \tilde{H}_{k,i}, \dots, \tilde{H}_{k,n})$ as the normalized matrix with each column computed by[1]:

$$\tilde{H}_{k,i} = \frac{H_{k,i}}{H_{k,i}^\top \mathbf{1}}. \tag{7}$$

Therefore, the new set of n uncorrelated portfolios has the return $\tilde{\mathbf{H}}_k^\top \mathbf{R}_k$ and the covariance matrix of the returns:

$$\tilde{\boldsymbol{\Sigma}}_k = \tilde{\mathbf{H}}_k \boldsymbol{\Sigma}_k \tilde{\mathbf{H}}_k^\top = \tilde{\boldsymbol{\Lambda}}_k, \tag{8}$$

where the diagonal matrix $\tilde{\boldsymbol{\Lambda}}_k$ with the i-th diagonal entity as $\tilde{\lambda}_{k,i} = \lambda_{k,i}/(H_{k,i}^\top \mathbf{1})^2$ characterizes the variances. For ease of presentation, we call the new set of portfolios the *orthogonal portfolios*. The orthogonal portfolios represent the risk factors in the market. Market fluctuations can be characterized as moves along the eigenvector directions [Meucci, 2009]. At time t_k the return and the variance of the i-th orthogonal portfolio are estimated as $\tilde{H}_{k,i} R_{k,i}$ and $\tilde{\lambda}_{k,i}$, respectively.

Algorithm 1 Orthogonal Bandit Portfolio

```
Inputs: $m$, $n$, $l$, $\Delta t$, $\mathbf{R}_k$, $\tau$
for $k = 1 \to m$ do
    Estimate the average return $\mathbb{E}[\mathbf{R}_k]$ and covariance matrix of asset returns $\boldsymbol{\Sigma}_k$ by $\{\mathbf{R}_{-\tau+k}, \dots, \mathbf{R}_{k-1}\}$;
    Implement the principal component decomposition as equation (5): $\boldsymbol{\Sigma}_k = \mathbf{H}_k \boldsymbol{\Lambda}_k \mathbf{H}_k^\top$;
    Compute the renormalized similarity matrices and eigenvalues (8): $\tilde{\boldsymbol{\Sigma}}_k = \tilde{\mathbf{H}}_k \boldsymbol{\Sigma}_k \tilde{\mathbf{H}}_k^\top = \tilde{\boldsymbol{\Lambda}}_k$;
    Compute the Sharpe ratio of each arm (10);
    Compute the adjusted reward function of each arm (11);
    Select the optimal arms according to (11) from the first $l$ and the next $n - l$ orthogonal portfolios, respectively;
    Compute the optimal mixture weight $\theta_k^*$ by (12);
    Compute the optimal portfolio weight $\boldsymbol{\omega}_k$ by (13);
Output:
The portfolio weight vectors $\boldsymbol{\omega}_k$ and the portfolio returns $\mu_k$ for $k = 1, \dots, m$.
```

In addition, numerous empirical studies in finance show the covariance matrix of asset returns consists of a few significant factors and other relatively unimportant factors [Bai and Ng, 2002; Meucci, 2009]. In other words, the covariance matrix can be decomposed as the sum of n rank-one matrices:

$$\tilde{\boldsymbol{\Sigma}}_k = \underbrace{\sum_{i=1}^{l} \tilde{\lambda}_{k,i} \tilde{H}_{k,i} \tilde{H}_{k,i}^\top}_{\text{significant, passive}} + \underbrace{\sum_{i=l+1}^{n} \tilde{\lambda}_{k,i} \tilde{H}_{k,i} \tilde{H}_{k,i}^\top}_{\text{insignificant, active}}, \tag{9}$$

where the first l factors are viewed as the systematic movement in market, industry and sector that investors should follow, and the next $n - l$ factors are considered as the idiosyncratic risks that investors may explore to generate extra return. Current research shows no consensus on the cutoff number l across different markets or asset classes. Three to five significant factors are commonly observed in empirical research and for specific market main factors are relatively stable along with time period [Fama and French, 1993]. In our empirical study, the cutoff l chosen for different markets according to the critical point where we observe a dramatic drop in the descending eigenvalues agrees with the above observation. Thus, we choose one portfolio from the first l orthogonal portfolios as a passive investment and another from the next $n - l$ orthogonal portfolios as an active investment. The former attempts to follow the market trend and enjoy the passive return; the latter represents the endeavor of adding potential extra returns from small factors [Grinold and Kahn, 2000].

Furthermore, to determine how to invest in those two subsets, we apply the UCB algorithm to the multi-armed bandit setting with l arms and $n - l$ arms, respectively [Auer *et al.*, 2002]. While standard multi-armed bandit algorithms hinge

[1] The denominator generally is non-zero, or it leads to a potential arbitrage chance by investing in the dollar-neutral portfolio $H_{k,i}$.

Table 1: Summary of the tested datasets

Dataset	Frequency	Time Period	m	n
FF48	Monthly	07/01/1963 - 12/31/2004	498	48
FF100	Monthly	07/01/1963 - 12/31/2004	498	100
ETF139	Weekly	01/01/2008 - 10/30/2012	252	139
EQ181	Weekly	01/01/2008 - 10/30/2012	252	181

on the expected return of each arm as the proxy of reward, we capture the tradeoff between risk and return by adopting the Sharpe ratio as our proxy of reward $\bar{r}_i(t_k)$ [Sharpe, 1966]. Unlike the general utility function that needs some subjective parameters quantifying the risk-averse degree [Sani *et al.*, 2012], the Sharpe ratio takes into account the return per risk unit directly. Specifically, at time t_k the Sharpe ratio of the i-th arm, i.e., the i-th orthogonal portfolio, is computed by:

$$\bar{r}_i(t_k) \equiv \mathrm{SR}_{k,i} = \frac{\mathbb{E}[\tilde{H}_{k,i} R_{k,i}]}{\sqrt{\tilde{\lambda}_{k,i}}} = \frac{H_{k,i}\mathbb{E}[R_{k,i}]}{\sqrt{\lambda_{k,i}}}. \tag{10}$$

Next, by incorporating the one-sided confidence bound into the reward function, we determine the optimal arm for each subset by the objective function:

$$i_k^* = \underset{i=1,\ldots,n}{\arg\max}\, \bar{r}_i(t_k) + \sqrt{\frac{2\ln(k+\tau)}{\tau + k_i}}, \tag{11}$$

where k_i stands for the number of times the i-th orthogonal portfolio has been chosen so far and τ represents the size of training data. As the historical information of assets is available, we incorporate it into the one-sided confidence bound through the length τ [Shivaswamy and Joachims, 2012].

Furthermore, after obtaining the optimal arms from the two subsets, we take the weighted average of them such that the total variance of the selected portfolio, i.e., $\lambda_{k,p}$, is minimized. Specifically, assume the i-th and j-th arms are selected from the two subsets, respectively. $\tilde{H}_{k,i_k^*}$ and $\tilde{H}_{k,j_k^*}$ are the corresponding uncorrelated portfolios. Thus, the portfolio mixture weight θ_k^* is computed by minimizing the total variance $\lambda_{k,p} = \theta_k^2 \tilde{\lambda}_{k,j_k^*} + (1-\theta_k)^2 \tilde{\lambda}_{k,i_k^*}$:

$$\theta_k^* = \underset{\theta_k}{\arg\min}\, \lambda_{k,p} = \frac{\tilde{\lambda}_{k,i_k^*}}{\tilde{\lambda}_{k,i_k^*} + \tilde{\lambda}_{k,j_k^*}}. \tag{12}$$

Therefore, we attain the portfolio mixed by the passive and active investment as:

$$\boldsymbol{\omega}_k = (1-\theta_k^*)\tilde{H}_{k,i_k^*} + \theta_k^* \tilde{H}_{k,j_k^*}. \tag{13}$$

Accordingly, the realized portfolio net return μ_k from time t_{k-1} to t_k will be $\mu_k = \boldsymbol{\omega}_k^\top \mathbf{R}_k - 1$. In the above formulation, we estimate the covariance matrix $\boldsymbol{\Sigma}_k$ by a factor model [Fan *et al.*, 2008] and estimate the average return $\mathbb{E}[\mathbf{R}_k]$ by the James-Stein shrinkage estimator [Meucci, 2009]. They both rest on the historical data in sliding windows with the size of τ training data. Algorithm 1 succinctly summarizes the steps of constructing the proposed orthogonal bandit portfolio.

4 Experiments

In this section, we describe the experimental settings and report the out-of-sample performance. We conduct the experiments on several empirical datasets and compare with representative portfolio choice methods in both finance and machine learning communities.

4.1 Data

In our experiments, we follow [Shen *et al.*, 2014] to choose two types of datasets for performance validation and comparison. The first benchmarks are the Fama and French (FF) datasets [Fama and French, 1992]. With the raw data from the US stock market, the FF benchmarks construct the portfolios for different financial segments. Specifically, the FF48 dataset contains monthly returns of 48 portfolios representing different industrial sectors, and the FF100 dataset includes monthly returns of 100 portfolios on the basis of size and book-to-market ratio. By virtue of the extensive coverage to asset classes and lengthy periods, the FF datasets are recognized as standard evaluation protocols and oft-adopted testbeds in the finance community. The second type benchmarks contain two datasets of actively traded assets, i.e., ETF139 and EQ181, which are crawled from *Yahoo! Finance* on a weekly base from 2008 to 2012. The ETF139 dataset consists of typical accessible asset classes for investors, i.e., exchange-traded funds, which have the advantages over conventional mutual funds of low costs, tax efficiency, and stock-like features. The EQ181 dataset represents the selection of the individual equities sampled from the large-cap segment of the Russell 200 index, covering 63% of total market capitalization. To avoid selection bias, we remove those stocks with missing historical data during our testing periods, thereby having a total of 181 stocks in the EQ181 dataset.

As summarized in Table 1, these two types of datasets embody different perspectives for performance assessment. The FF datasets essentially underscore the long-term performance of the proposed strategy. The period range contains highly volatile times in the stock market, such as "Black Monday" in 1987, the Internet bubble burst and September 11 terrorist attacks in 2001. Such long historical datasets would introduce limited selection bias and performance manipulation. On the other hand, the ETF and equity datasets underline the robustness of the proposed portfolio with respect to the higher trading frequency and the special market environment after the recent financial crisis that began in 2007 with the default of subprime mortgage loans.

4.2 Experimental Settings

Following the "rolling-horizon" settings in [Shen *et al.*, 2014], we use sliding windows with the size of $\tau = 120$ months/weeks of training data to construct portfolios for the subsequent month/week. For our comparative study, we consider the following competing methods: a) equally-weighted portfolio (EW); b) value-weighted portfolio (VW); c) conventional minimum-variance portfolio (MVP); d) on-line moving average reversion (MAR) based portfolio; e) naive bandit portfolio (NBP); f) the proposed orthogonal bandit portfolio (OBP). The first three portfolio strategies are typical

Table 2: Portfolio Sharpe ratios (%) with the significance level measured by p-values with respect to EW.

Dataset	OBP	NBP	EW	VW	MVP	MAR
FF48	**26.15**	25.68	24.30	23.37	22.38	24.48
	(0.64)	(0.80)	(1.00)	(0.22)	(0.72)	(0.93)
FF100	**34.89**	26.09	26.97	29.76	18.01	23.60
	(0.01)	(0.82)	(1.00)	(0.00)	(0.19)	(0.22)
ETF139	**25.47**	15.45	6.01	5.85	6.82	7.61
	(0.05)	(0.04)	(1.00)	(0.22)	(0.94)	(0.44)
EQ181	**18.10**	11.62	9.03	8.95	13.62	12.15
	(0.19)	(0.74)	(1.00)	(0.86)	(0.66)	(0.27)

Table 3: Portfolio terminal cumulative wealth ($).

Dataset	OBP	NBP	EW	VW	MVP	MAR
FF48	**61.75**	35.23	54.77	48.06	25.10	42.34
FF100	**626.04**	76.91	123.92	198.32	74.73	57.74
ETF139	**1.42**	1.35	1.20	1.19	1.05	1.21
EQ181	**1.69**	1.28	1.30	1.29	1.42	1.34

baselines circulated in the finance community. For example, the EW portfolio is a naive approach yet has been empirically shown to mostly outperform 14 models across seven empirical datasets [DeMiguel *et al.*, 2009b]. The VW portfolio mimics a market portfolio by weighing individual market components according to their market capitalization. The MVP portfolio has shown significant performance improvement over the Markowitz mean-variance portfolio [Jagannathan and Ma, 2003]. On the other hand, the MAR portfolio represents a more data-driven approach developed by machine learning researchers and outperforms 12 different portfolio strategies across five datasets [Li and Hoi, 2012]. The NBP portfolio implements the UCB bandit algorithm under the standard assumption of i.i.d. rewards without considering the correlations between assets. Specifically, it selects one asset at a time for investment.

4.3 Performance Metrics

We compare the out-of-sample performance of the portfolios by the standard criteria in finance [Brandt, 2010]: (i) *Sharpe ratios*; and (ii) *cumulative wealth*. The *Sharpe ratio* (SR) measures the reward-to-risk ratio of a portfolio strategy, which is computed as the portfolio return normalized by its standard deviation:

$$\text{SR} = \frac{\hat{\mu}}{\hat{\sigma}} \tag{14}$$

where the mean of portfolio net returns $\hat{\mu}$ and the corresponding standard deviation $\hat{\sigma}$ are computed as

$$\hat{\mu} = \frac{1}{m}\sum_{k=1}^{m}\mu_k, \quad \hat{\sigma} = \sqrt{\frac{1}{m-1}\sum_{k=1}^{m}(\mu_k - \hat{\mu})^2}. \tag{15}$$

Since SR is a summary statistic of returns, we supplement this commonly used measure of investment performance with the time series plot of *cumulative wealth* (CW). While CW measures the total profit yield from the portfolio strategy across investment periods without considering any risks and costs, investors are commonly concerned about the growth of their investment. Starting the investment period with one dollar, CW is computed by

$$\text{CW} = \prod_{k=1}^{m} \boldsymbol{\omega}_k^{\top}\mathbf{R}_k. \tag{16}$$

To further quantify the statistical significance of the difference in SR between two comparing portfolios, we also report the p-values under the corresponding SR results. To compute the p-values for the case of non-i.i.d. returns, we adopt the studentized circular block bootstrapping methodology in [Ledoit and Wolf, 2008]. In particular, we set the EW portfolio as the benchmark with 1000 bootstrap resamples, 95% significance level, and a block with the size of 5.

4.4 Results

Table 2 reports the SR values with the p-values over the entire investment period, where the best performance is highlighted in bold. Apparently, the proposed orthogonal bandit portfolio has achieved the highest risk-adjusted returns, i.e., Sharpe ratios, in all the four datasets. Since NBP has only achieved comparable performance with OBP in the FF48 dataset, it reveals that asset correlations play a crucial role in portfolio choice problems. In particular, in the ETF139 and the EQ181 datasets, the proposed OBP method has generated the highest SR with significant margins. In addition, we can gain confirmation from the p-values that the OBP method is statistically distinguishable from the simple yet powerful EW strategy in FF100 and ETF139. NBP works relatively better in ETF139 and EQ181. This observation likely indicates that the on-line bandit strategy is more efficacious for short-term investment.

Table 3 summarizes the terminal cumulative wealth of different portfolios in all the datasets. Echoing with its superior performance in risk-adjusted return, OBP has generated the greatest increase in wealth among all the competing approaches. For the long investment periods such as the FF48

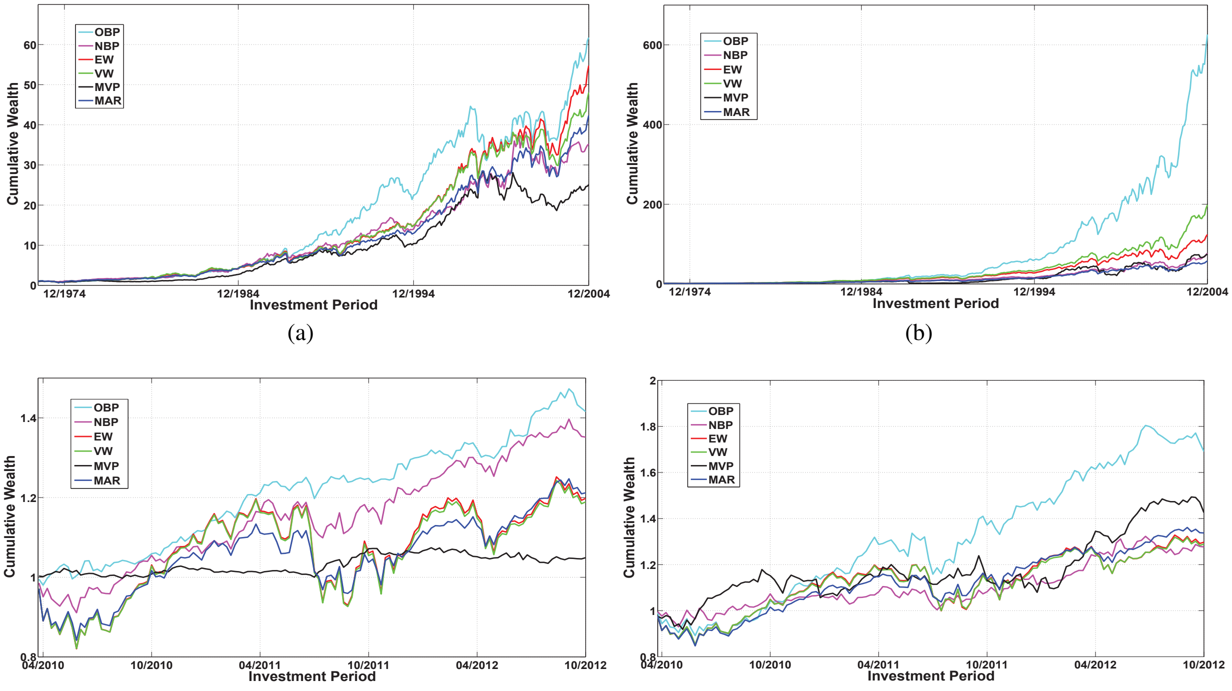

Figure 1: The curves of *cumulative wealth* across the investment periods for different portfolios yield from a) FF48, b) FF100, c) ETF139, and d) EQ181.

and FF100 datasets, the OBP method has produced significantly more wealth than the market itself, increasing the final wealth by 28.5% and 215.7%, respectively.

Figure 1 shows the time series curves of the cumulative wealth over the investment periods. OBP has demonstrated the highest wealth level in most of time. In particular, the illustration of its wealth cumulative trends that are generally disparate from others implicitly implies the distinction of its design principles. Our approach may perform even better if investors inject domain knowledge into the OBP framework that we illustrate here.

5 Conclusion

In this paper, we tackle the portfolio choice problems by an orthogonal bandit algorithm. Our novel algorithm has addressed the conundrum of making portfolio choices via multi-armed bandits with orthogonal portfolios. In particular, we orthogonalize generally correlated financial assets to create orthogonal portfolios through the principal component decomposition for the standard bandit learning framework; we incorporate the Sharpe ratio from the finance field as a risk-adjusted reward function into the UCB algorithm to direct investment; we further synergistically combine the generated passive and active investments to construct a low-risk portfolio suitable to normal investors. Our future work not only includes appropriately generalizing and enriching the current framework to encompass more practical concerns in portfolio choice problems, such as position constraints, transaction costs and taxes [Shen and Wang, 2015; Dammon and Spatt, 2012], but also contains studying theoretical underpinning of the bandit problems they may bring.

References

[Agarwal *et al.*, 2006] A. Agarwal, E. Hazan, S. Kale, and R. E. Schapire. Algorithms for portfolio management based on the Newton method. In *Proceedings of the 23th international conference on machine learning*, pages 9–16, 2006.

[Auer *et al.*, 1995] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331, 1995.

[Auer *et al.*, 2002] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[Bai and Ng, 2002] J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[Blum and Kalai, 1999] A. Blum and A. Kalai. Universal portfolios with and without transaction costs. *Machine Learning*, 35(3):193–205, 1999.

[Borodin *et al.*, 2004] A. Borodin, R. El-Yaniv, and V. Gogan. Can we learn to beat the best stock? *Journal of Artificial Intelligence*, 21:579–594, 2004.

[Brandt, 2010] M. W. Brandt. Portfolio choice problems. In Y. Ait-Sahalia and L. P. Hansen, editors, *Handbooks of Financial Econometrics*, pages 269–336. Elsevier, 2010.

[Broadie, 1993] M. Broadie. Computing efficient frontiers using estimated parameters. *Annals of Operations Research*, 45(1):21–58, 1993.

[Bubeck and Cesa-Bianchi, 2012] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 2012.

[Chapelle and Li, 2011] O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[Chen *et al.*, 2013] W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

[Dammon and Spatt, 2012] R. Dammon and C. Spatt. Taxes and investment choice. *The Annual Review of Financial Economics*, 4(1):411–429, 2012.

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

[DeMiguel *et al.*, 2009b] V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? *The Review of Financial Study*, 22:1915–1953, 2009.

[Even-Dar *et al.*, 2006] E. Even-Dar, M. Kearns, and J. Wortman. Risk-sensitive online learning. In *Algorithmic Learning Theory*, pages 199–213. Springer, 2006.

[Fama and French, 1992] E. F. Fama and K. R. French. The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465, 1992.

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

[Fan *et al.*, 2008] J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147:186–197, 2008.

[Gagliolo and Schmidhuber, 2011] M. Gagliolo and J. Schmidhuber. Algorithm portfolio selection as a bandit problem with unbounded losses. *Annals of Mathematics and Artificial Intelligence*, 61(2):49–86, 2011.

[Graepel *et al.*, 2010] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. In *Proceedings of the 27th International Conference on Machine Learning*, pages 13–20, 2010.

[Grinold and Kahn, 2000] R. C. Grinold and R. N. Kahn. *Active portfolio management*. McGraw Hill New York, NY, 2000.

[Györfi *et al.*, 2006] L. Györfi, G. Lugosi, and F. Udina. Nonparametric kernal-based sequential investment strategies. *Mathematical Finance*, 16:337–357, 2006.

[Helmbold *et al.*, 1998] D. P. Helmbold, R. E. Schapire, Y. Singer, and M. K. Singer. On-line portfolio selection using multiplicative updates. *Mathematical Finance*, 8:325–347, 1998.

[Hoffman *et al.*, 2011] M. D. Hoffman, E. Brochu, and N. de Freitas. Portfolio allocation for Bayesian optimization. In *The Conference on Uncertainty in Artificial Intelligence*, 2011.

[Jagannathan and Ma, 2003] R. Jagannathan and T. Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58:1651–1684, 2003.

[Konno and Yamazaki, 1991] H. Konno and H. Yamazaki. Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. *Management science*, 37(5):519–531, 1991.

[Lai and Robbins, 1985] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[Langford and Zhang, 2008] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

[Ledoit and Wolf, 2008] O. Ledoit and M. Wolf. Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15:850–859, 2008.

[Li and Hoi, 2012] B. Li and S. C. Hoi. On-line portfolio selection with moving average reversion. In *Proceedings of the 29th international conference on machine learning*, 2012.

[Li and Hoi, 2014] B. Li and S. C. Hoi. Online portfolio selection: A survey. *ACM Computing Survey*, 46(3):35, 2014.

[Li *et al.*, 2010] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[Maillard, 2013] O. A. Maillard. Robust risk-averse stochastic multi-armed bandits. In *Algorithmic Learning Theory*, pages 218–233. Springer, 2013.

[Markowitz, 1952] H. Markowitz. Portfolio selection. *Journal of Finance*, 7:77–91, 1952.

[Meucci, 2009] A. Meucci. *Risk and asset allocation*. Springer Science & Business Media, 2009.

[Press, 2009] W. H. Press. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proceedings of the National Academy of Sciences*, 106(52):22387–22392, 2009.

[Robbins, 1952] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

[Sani *et al.*, 2012] A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

[Scott, 2010] S. L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

[Sharpe, 1966] W. F. Sharpe. Mutual fund performance. *Journal of Business*, 39:119–138, 1966.

[Shen and Wang, 2015] W. Shen and J. Wang. Transaction costs-aware portfolio optimization via fast Löwner-John ellipsoid approximation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1854–1860, 2015.

[Shen *et al.*, 2014] W. Shen, J. Wang, and S. Ma. Doubly regularized portfolio with risk minimization. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1286–1292, 2014.

[Shivaswamy and Joachims, 2012] P. K. Shivaswamy and T. Joachims. Multi-armed bandit problems with history. In *International Conference on Artificial Intelligence and Statistics*, pages 1046–1054, 2012.

[Thorp, 1971] E. O. Thorp. Portfolio choice and the Kelly criterion. *Business and Economics Statistics Section of Proceedings of the American Statistical Association*, pages 215–224, 1971.