2014 IEEE Congress on Evolutionary Computation (CEC)
July 6-11, 2014, Beijing, China

# Two Parameter Update Schemes for Recurrent Reinforcement Learning

Jin Zhang and Dietmar Maringer
Faculty of Business and Economics, University of Basel, Peter Merian-Weg 6,
CH-4002, Switzerland
Email: {jin.zhang, dietmar.maringer}@unibas.ch

***Abstract*—Recurrent reinforcement learning (RRL) is a machine learning algorithm which has been proposed by researchers for constructing financial trading platforms. When an analysis of RRL trading performance is conducted using low frequency financial data (e.g. daily data), the weakening autocorrelation in price changes may lead to a decrease in trading profits as compared to its applications in high frequency trading. There therefore is a need to improve RRL for the purposes of daily equity trading. This paper presents two parameter update schemes (the 'average elitist' and the 'multiple elitist') for RRL. The purpose of the first scheme is to improve out-of-sample performance of RRL-type trading systems. The second scheme aims to exploit serial dependence in stock returns to improve trading performance, when traders deal with highly correlated stocks. Profitability and stability of the trading system are examined by using four groups of S&P stocks for the period January 2009 to December 2012. It is found that the Sharpe ratios of the stocks increase after we use the two parameter update schemes in the RRL trading system.**

## I. Introduction

Recent developments in algorithmic trading show that there are ongoing attempts to find and develop new trading strategies. Recurrent reinforcement learning (RRL) which finds approximate solutions to stochastic dynamic programming problems, has been used to design online trading platforms [1]. It has been found that RRL-type trading systems are especially good at making profits by trading commodities and Forex for high frequency trading (see [2] and [3]). This is, because RRL-type trading systems are able to pick up the strong trends presented in price changes. Many extensions of the basic RRL trading system can be found in the literature, for example: a multi-layer RRL trading system which incorporates risk management and utility optimization into one online trading module [4]; and a regime-switching RRL trading system which aims to cope with the nonlinear behavior in price changes due to economic regime switches [5].

The RRL technique can be considered as a stochastic gradient ascent algorithm which continuously optimizes a utility measure by utilizing newly arrived market information. A major concern about RRL-type trading systems, however, is whether the parameters of the RRL-type trading system are well tuned for out-of-sample trading. In the literature, designers of RRL-type trading systems suggest that the trading system should first be turned during a training period. The trades made and performance of the system during the training period are used to update the parameters. Training of the trading system is usually repeated for many numbers of epochs to make sure that the parameters are well tuned before out-of-sample trading. The trades then made in an out-of-sample period are the actual trades for the period, and the update of the parameters is continuously guided by RRL in the period. It has been pointed out by researchers that the tuning process is important for out-of-sample trading performance since RRL-type trading systems may underperform in out-of-sample periods due to the over-fitting issue (see [3]).

In the context of daily equity trading, the use of large quantities of historical data may hamper an effective prediction of short-term price movements since there are only a limited number of observations which may carry valuable information for price movement prediction. In this paper, we propose two parameter update schemes for RRL. The first scheme (namely the 'average elitist'), which can be applied to general RRL-type trading systems, is designed to reduce the impact of over-fitting on out-of-sample trading profits. The second scheme (namely, the 'multiple elitist') aims to exploit serial dependence in stock returns, in order to improve the trading performance of RRL systems when traders deal with highly correlated stocks.

The rest of the paper is organized as follows. Section II introduces the two update schemes, Section III presents experimental results, and Section IV concludes.

## II. RRL trading systems

### A. Recurrent reinforcement learning

Recurrent reinforcement learning (RRL) has been used to tune financial trading systems for the purpose of utility maximization based on newly arrived market information [1]. In the literature, the market information usually refers to a series of lagged price returns, although RRL trading systems can easily accommodate technical indicators and financial fundamentals (see [6]).

The basic RRL trading system is designed to trade a single-asset with a two-position action (long/short), which is produced using linear combinations and a tanh function. In Figure 1, $\boldsymbol{x}$, $v$ and $F_t$ refer to the market information and $\boldsymbol{\theta_t}$ denotes a parameter set of the input signals.

We use the objective function which has been applied in most discussions on RRL-type trading systems. If we denote the utility function as $U_t$ which depends on the most recently realized trading reward $R_t$, the goal of the RRL trading system is to maximize the wealth measure $U_t$ by adjusting the parameter set $\boldsymbol{\theta}_t$ (hereafter referred to as signal parameters) in

978-1-4799-1488-3/14/$31.00 ©2014 IEEE

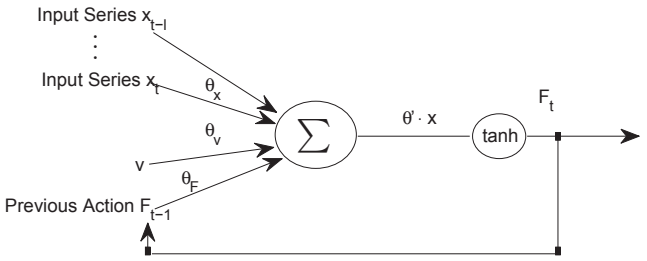


Fig. 1. Recurrent reinforcement learning

a continuous manner:

$$\max U_t(R_t; \boldsymbol{\theta}_t). \tag{1}$$

By denoting the closing price as $P_t$, the price change of an asset is defined as $r_t = \ln \frac{P_t}{P_{t-1}}$. At time $t$, the realized profit can be written as:

$$R_t = \nu \cdot (\mathrm{sgn}(F_{t-1}) \cdot r_t - \delta \cdot |\,\mathrm{sgn}(F_t) - \mathrm{sgn}(F_{t-1})|), \tag{2}$$

where the $\nu$ is the number of shares and the $\delta$ is the transactions cost rate. The $\mathrm{sgn}(F_{t-1})$ refers to the current holding position, and the $\mathrm{sgn}(F_t)$ denotes the holding position for next period. The $F_t$ is defined as:

$$F_t = \tanh\left(\boldsymbol{\theta}_t \times \mathbf{I}_t\right), \tag{3}$$

where $\mathbf{I}_t$ denotes a set of input signals usually including a prior trading signal $F_{t-1}$, a constant $v$ with a value of 1, a set of lagged returns $r_t, r_{t-1}, r_{t-2}, \ldots, r_{t-l+1}$, $t = 1, \ldots, T$. $l$ is an integer number representing a length of the lags. Since RRL updates the signal parameters $\boldsymbol{\theta}_t$ using the stochastic gradient ascent, the gradients of $U_t$ with respect to the signal parameter set $\boldsymbol{\theta}_t$ can be written as:

$$\frac{dU_t(\boldsymbol{\theta}_t)}{d\boldsymbol{\theta}_t} = \frac{dU_t}{dR_t}\left\{\frac{dR_t}{dF_t}\frac{dF_t}{d\boldsymbol{\theta}_t} + \frac{dR_t}{dF_{t-1}}\frac{dF_{t-1}}{d\boldsymbol{\theta}_{t-1}}\right\}, \tag{4}$$

with

$$\frac{dF_t}{d\boldsymbol{\theta}_t} \approx \frac{\partial F_t}{\partial \boldsymbol{\theta}_t} + \frac{\partial F_t}{\partial F_{t-1}}\frac{dF_{t-1}}{d\boldsymbol{\theta}_{t-1}}, \tag{5}$$

$$\frac{dR_t}{dF_{t-1}} = \nu \cdot (r_t + \delta \cdot \mathrm{sgn}(F_t - F_{t-1})), \tag{6}$$

and

$$\frac{dR_t}{dF_t} = -\nu \cdot \delta \cdot \mathrm{sgn}(F_t - F_{t-1}). \tag{7}$$

At any time $t$, the update of signal parameters $\boldsymbol{\theta}_t$ follows:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \rho\frac{dU_t(\boldsymbol{\theta}_t)}{d\boldsymbol{\theta}_t}, \tag{8}$$

where $\rho$ is the learning rate.

In the literature, a utility measure which is widely used in RRL-type trading systems is the so-called Differential Sharpe Ratio (DSR). The DSR is the first-order term after taking the Taylor series expansion of a performance measure, namely exponential moving average Sharpe ratio (EMSR), at $\eta \to 0$. The EMSR is defined as:

$$EMSR_t = \frac{A_t}{K_\eta \cdot (B_t - A_t^2)^{1/2}}, \tag{9}$$

where $A_t = A_{t-1} + \eta \cdot (R_t - A_{t-1})$, $B_t = B_{t-1} + \eta \cdot (R_t^2 - B_{t-1})$, and $K_\eta = (\frac{1-\eta/2}{1-\eta})^{1/2}$. In other words, the utility measure $U_t$, i.e. the first-order term can be written as:

$$DSR_t = \frac{B_{t-1} \cdot (R_t - A_{t-1}) - \frac{1}{2} \cdot A_{t-1} \cdot (R_t^2 - B_{t-1})}{(B_{t-1} - A_{t-1}^2)^{\frac{3}{2}}}, \tag{10}$$

and the derivative of $U_t$ with respect to $R_t$ can be written as:

$$\frac{dU_t}{dR_t} = \frac{B_{t-1} - A_{t-1} \cdot R_t}{(B_{t-1} - A_{t-1}^2)^{3/2}}. \tag{11}$$

### *B. Two proposed update schemes*

*1) The 'average elitist' scheme:* In this section we introduce our proposed 'average elitist' scheme. Unlike most discussions on RRL trading systems in the literature, we use a group of simulation traders (sim-trader) rather than a single trader to derive the signal parameters for out-of-sample trading. We initialize the signal parameters of these sim-traders using random numbers. We do so, because, in the real world, traders may have different levels of information asymmetry and different expectations about future price changes.

These sim-traders are then trained with the data in a training period $T_{train}$. The training period covers an evaluation period $T_{eva}$. At the end of the training period, we set up an elitist set $\mathcal{E}$, in which members are selected according to the Sharpe ratio rankings of the sim-traders in the evaluation period $T_{eva}$. The signal parameters in the average elitist update scheme are the average values of the elitist members' signal parameters.

Assuming that the elitist set $\mathcal{E}$ consists of a number of $N$ 'elitist' members, at the beginning of the out-of-sample period $\mathfrak{t}_0 = T_{train} + 1$, the parameter set of the 'average elitist' RRL trading system is defined as

$$\widehat{\boldsymbol{\theta}}_{\mathfrak{t}_0} = \frac{1}{N}\sum_i \boldsymbol{\theta}_{\mathfrak{t}_0,i}, \quad i \in \mathcal{E}. \tag{12}$$

The update of signal parameters in the out-of-sample trading period follows:

$$\Delta\widehat{\boldsymbol{\theta}}_{\mathfrak{t}} = \frac{1}{N}\sum_i \frac{dU_{\mathfrak{t}}(\boldsymbol{\theta}_{\mathfrak{t},i})}{d\boldsymbol{\theta}_{\mathfrak{t},i}}, \quad i \in \mathcal{E} \tag{13}$$

$$\widehat{\boldsymbol{\theta}}_{\mathfrak{t}+1} = \widehat{\boldsymbol{\theta}}_{\mathfrak{t}} + \rho\Delta\widehat{\boldsymbol{\theta}}_{\mathfrak{t}}. \tag{14}$$

According to the literature, some researchers use the sim-trader which produces the highest Sharpe ratio in an evaluation period from these sim-traders for out-of-sample trading. We refer the sim-trader as to the 'elitist' trader in this study.

*2) The 'multiple elitist' scheme:* In this section, we introduce our proposed 'multiple elitist' scheme which we develop for RRL where equity returns are highly correlated or are jointly affected by certain economic factors. We argue that taking into account the collective behavior of correlated stock returns may facilitate the prediction of price changes. For example, there is an equity group $\mathcal{M}$ which consists of a number of $M$ highly correlated stocks. In the training period, we use the return series of these highly correlated stocks to adjust the parameters of the RRL trading systems which we

set up for trading these stocks. The gradient vector of a stock member $m$ in the out-of-sample period is defined as follows:

$$\Delta\breve{\boldsymbol{\theta}}_{\mathrm{t}}^{m} = \left(\psi\Delta\widehat{\boldsymbol{\theta}}_{\mathrm{t}}^{m} + (1-\psi)\frac{1}{M-1}\sum_{j\neq m}\Delta\widehat{\boldsymbol{\theta}}_{\mathrm{t}}^{j}\right), \quad j \in \mathcal{M}, \tag{15}$$

where $\psi$ is a parameter controlling the impact of gradients from its peers on $\Delta\breve{\boldsymbol{\theta}}_{\mathrm{t}}^{m}$. In other words, the update relies not only on the gradients from its own system, but also on the gradients from the others in the group.

We use Eq. (12) to initialize the signal parameters $\breve{\boldsymbol{\theta}}_{\mathrm{t_o}}^{m}$. The update of signal parameters in the out-of-sample trading period follows:

$$\breve{\boldsymbol{\theta}}_{\mathrm{t}+1}^{m} = \breve{\boldsymbol{\theta}}_{\mathrm{t}}^{m} + \rho\Delta\breve{\boldsymbol{\theta}}_{\mathrm{t}}^{m}. \tag{16}$$

Industry sectors, statistical measures (e.g. Pearson's correlation, Spearman's correlation or Kendall's tau), and data mining skills such as clustering techniques can be used to construct equity groups.

## III. Results of the Experiment

### A. Data sets

The companies selected for this research are S&P 500 American companies traded on the New York Stock Exchange and on Nasdaq. We download the daily prices of the 500 companies from Bloomberg (1st January 2009 – 3rd December 2012, 980 observations for a single series). The 980 observations are partitioned into: an initial training set consists of the first 750 samples ($T_{train} = 750$) which also covers an evaluation period comprised of the last 250 samples in the training set ($T_{eva} = 250$); and an out-of-sample period consists of the following 230 observations ($T_{trade} = 230$).

In this paper, we use the correlation coefficient to identify highly correlated stocks. A correlation coefficient matrix is constructed based on the first 750 observations of the 500 stocks. We are able to identify four groups (covering 18 stocks), given a correlation coefficient of any pairs in the group greater than 0.7.

### B. Parameter settings

Each of the RRL trading systems consists of 100 sim-traders. The signal parameters of these sim-traders are initialized by using random numbers from a Gaussian distribution with a mean of 0 and a standard deviation of 0.05. The trading systems are then tuned by using the RRL technique with the first 750 observations. The top 5% performing sim-traders from the 100 sim-traders are selected as elitist members in the set $\mathcal{E}$, according to their Sharpe ratio rankings in the evaluation period, i.e. the last 250 trades in the training set.

Based on preliminary tests, the following parameters were found to be suitable settings for the daily equity trading problem: the number of shares traded $\nu = 1$; the learning rate $\rho = 0.15$; the adaption rate $\eta = 0.05$. We expect that new information will be reflected in stock prices in a maximum period of two weeks, therefore we use a value of $l = 10$. The transaction cost $\delta$ has a value of 3 bps.

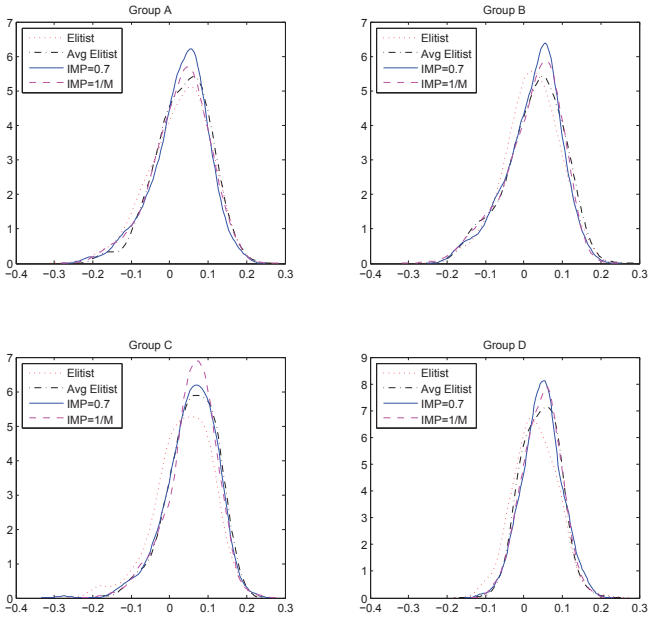


Fig. 2. Distribution of aggregated Sharpe ratios

Regarding the 'multiple elitist' scheme, we consider three values of $\psi$ ($\psi = 1$, $\psi = 0.7$, and $\psi = 1/M$). It should be noted that, given a value of $\psi = 1$, the 'multiple elitist' scheme reduces to the 'average elitist' scheme. When $\psi$ has a value of $\frac{1}{M}$, the 'multiple elitist' scheme is a 'one size fits all' approach, because the gradients $\Delta\breve{\boldsymbol{\theta}}_{\mathrm{t}}^{m}$ to the group members are all the same.

### C. Out-of-sample trading performance

Profitability and stability are probably the most two important factors when traders assess a financial trading system. In this study, we use the daily Sharpe ratio to measure the profitability. Stability refers to the consistency of the Sharpe ratios generated from independent restarts of the trading system with different initial signal parameters, as trading performance relates directly to the starting values of the signal parameters.

To assess the stability, we restart the RRL trading system 100 times. We save the Sharpe ratio of each stock which is produced by using the RRL trading system with the three different values of $\psi$ in each trial. For comparison purposes, we also save the Sharpe ratio on each stock which is produced by the elitist trader, i.e. the one producing the highest Sharpe ratio from the 100 sim-traders in the evaluation period.

As we are interested in studying the Sharpe ratios of stocks which are highly correlated, we put the Sharpe ratios of the stocks together in the same group. The probability density of the aggregated Sharpe ratios of each group are shown in Figure 2. Generally speaking, the Sharpe ratios which are produced by using the 'elitist' trader are lower, comparing to that of the 'average elitist' and the 'multiple elitist' schemes.

It seems that the Sharpe ratio density curves which are produced using the 'multiple elitist' RRL trading system with different values of $\psi$ are very close to each other. Table I provides the Sharpe ratio means and standard deviations. It is

TABLE I. STATISTICS OF THE DAILY SHARPE RATIOS

| | | Elitist | Avg. E. | $\phi$=0.7 | $\phi$=1/M |
|---|---|---|---|---|---|
| Group A | Mean | 0.0246 | 0.0348 | 0.0288 | 0.0290 |
| | Std | 0.0777 | 0.0702 | 0.0715 | 0.0739 |
| Group B | Mean | 0.0193 | 0.0275 | 0.0241 | 0.0224 |
| | Std | 0.0712 | 0.0753 | 0.0710 | 0.0747 |
| Group C | Mean | 0.0371 | 0.0607 | 0.0572 | 0.0577 |
| | Std | 0.0718 | 0.0625 | 0.0650 | 0.0630 |
| Group D | Mean | 0.0296 | 0.0427 | 0.0465 | 0.0455 |
| | Std | 0.0564 | 0.0487 | 0.0504 | 0.0483 |

TABLE II. P-VALUES OF THE TWO SAMPLE KOLMOGOROV-SMIRNOV TEST

| Group A | Elitist | Avg. Elitist | IMP=0.7 | IMP=1/M |
|---|---|---|---|---|
| Elitist | 1 | | | |
| Avg. Elitist | 0.09 | 1 | | |
| IMP=0.7 | 0.10 | 0.28 | 1 | |
| IMP=1/M | 0.43 | 0.52 | 0.62 | 1 |
| **Group B** | **Elitist** | **Avg. Elitist** | **IMP=0.7** | **IMP=1/M** |
| Elitist | 1 | | | |
| Avg. Elitist | 0.03 | 1 | | |
| IMP=0.7 | 0.04 | 0.29 | 1 | |
| IMP=1/M | 0.03 | 0.57 | 0.86 | 1 |
| **Group C** | **Elitist** | **Avg. Elitist** | **IMP=0.7** | **IMP=1/M** |
| Elitist | 1 | | | |
| Avg. Elitist | 0.00 | 1 | | |
| IMP=0.7 | 0.00 | 0.75 | 1 | |
| IMP=1/M | 0.00 | 0.31 | 0.90 | 1 |
| **Group D** | **Elitist** | **Avg. Elitist** | **IMP=0.7** | **IMP=1/M** |
| Elitist | 1 | | | |
| Avg. Elitist | 0.00 | 1 | | |
| IMP=0.7 | 0.00 | 0.27 | 1 | |
| IMP=1/M | 0.00 | 0.51 | 0.90 | 1 |

found that in most cases the 'average elitist' scheme outperforms the 'multiple elitist' scheme in terms of higher Sharpe ratio means, although the differences between the Sharpe ratio means are not statistically significant. We use the two-sample Kolmogorov-Smirnov test to quantify the difference between the Sharpe ratio distributions and Table II reports the $p$-values. We find that the difference between the Sharpe ratios is not statistically significant at the conventional 5% level.

Figures 3 and 4 provide the Sharpe ratio densities of the stocks in Group C and D respectively. For comparison purposes, a random trading strategy is used to benchmark the trading performance of the two RRL-type trading systems. We set up a random trading system for each asset. We restart the random trading system 100 times and save the Sharpe ratios. In addition to the random trading strategy, a buy-and-hold strategy is used as the second benchmark strategy in the out-of-sample study. We use a bootstrap approach to generate artificial returns. The bootstrap approach keeps the dependence structure unchanged among the assets in a group, and bootstraps a $230 \times M$ artificial return matrix from the out-of-sample historical returns in an iteration. We save the Sharpe ratios which are generated by using a bootstrapped iteration number of 100 for each stock. It is found that the random trading strategy and the buy-and-hold strategy can hardly generate any significant profits. In the light of these plots, some of the Sharpe ratio means which are produced by using the two proposed RRL parameter schemes are statistically greater than zero. The two density curves which are produced by using the trading systems based on the 'average elitist' and 'multiple elitist' schemes are very close to each other. In other words, the improvement in the Sharpe ratio by using the 'multiple elitist' scheme is not significant when compared with the Sharpe ratio produced using the 'average elitist' update scheme.

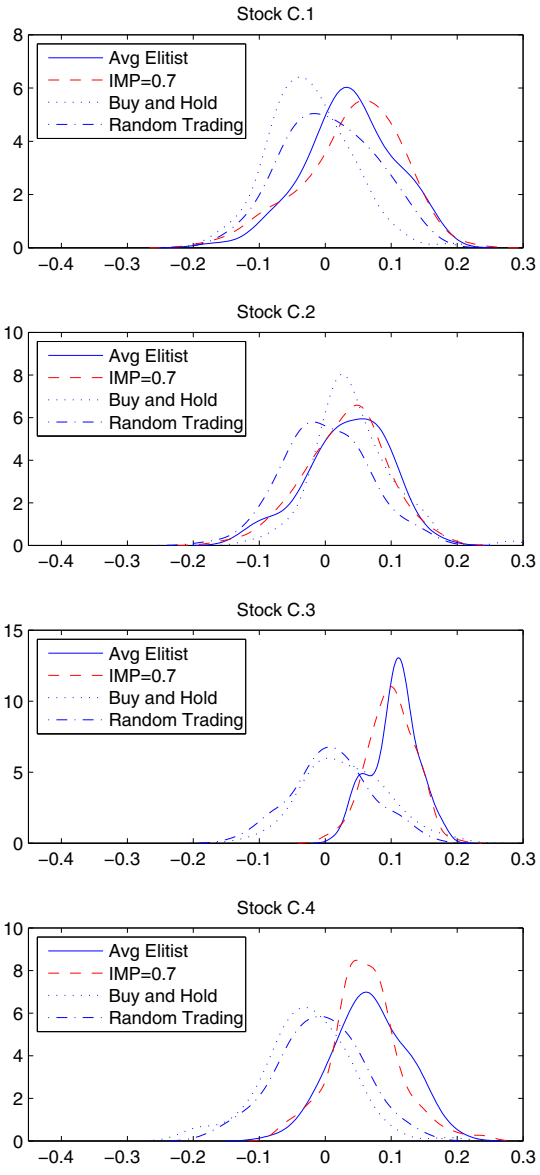


Fig. 3. The Sharpe ratios in Group C

## IV. CONCLUSIONS

In this paper, we introduce two parameter update schemes for recurrent reinforcement learning: the 'average elitist' scheme and the 'multiple elitist' scheme. The purpose of the first scheme is to improve out-of-sample performance. The second scheme aims to exploit serial dependence in stock returns to improve trading performance when trading highly correlated stocks. The distributions of the Sharpe ratios which are produced by using the two new RRL trading systems are examined using the daily data of four groups of S&P stocks for the period January 2009 to December 2012.

It is found that the trading systems developed based on the two proposed RRL techniques outperform the random trading strategy and the buy-and-hold strategy in producing higher Sharpe ratio means. The results of our experiment also show that the 'average elitist' scheme outperforms the 'elitist' scheme in terms of profitability and stability. Although the aggregated Sharpe ratios from trading the stocks in a same group are not significant, it is found that the Sharpe ratio of individual stocks are statistically greater than zero. The improvement in the Sharpe ratios by using the 'multiple elitist' scheme is not significant in the comparison with that produced using the 'average elitist' update scheme.

## ACKNOWLEDGMENT

The authors acknowledge the support provided by the Swiss National Science Foundation (SNSF) grant program 138095.

## REFERENCES

[1] J. Moody, L. Wu, Y. Liao, and M. Saffell, "Performance functions and reinforcement learning for trading systems and portfolios," *Journal of Forecasting*, vol. 17, pp. 441–470, 1998.

[2] C. Gold, "FX trading via recurrent reinforcement learning," in *Proceedings of IEEE international conference on computational intelligence in financial engineering*. Los Alamitos: IEEE Computer Society Press, 2003, pp. 363–370.

[3] M. A. H. Dempster, T. W. Payne, Y. Romahi, and G. W. P. Thompson, "Computational learning techniques for intraday FX trading using popular technical indicators," *IEEE Transactions on Neural Networks*, vol. 12(4), pp. 744–754, 2001.

[4] M. A. H. Dempster and V. Leemans, "An automated FX trading system using adaptive reinforcement learning," *Expert Systems with Applications*, vol. 30, pp. 543–552, 2006.

[5] D. Maringer and T. Ramtohul, "Regime-switching recurrent reinforcement learning for investment decision making," *Computational Management Science*, vol. 9, pp. 89–107, 2012.

[6] J. Zhang and D. Maringer, "Indicator selection for daily equity trading with recurrent reinforcement learning," in *GECCO 2013*, 2013, pp. 1757–1758.

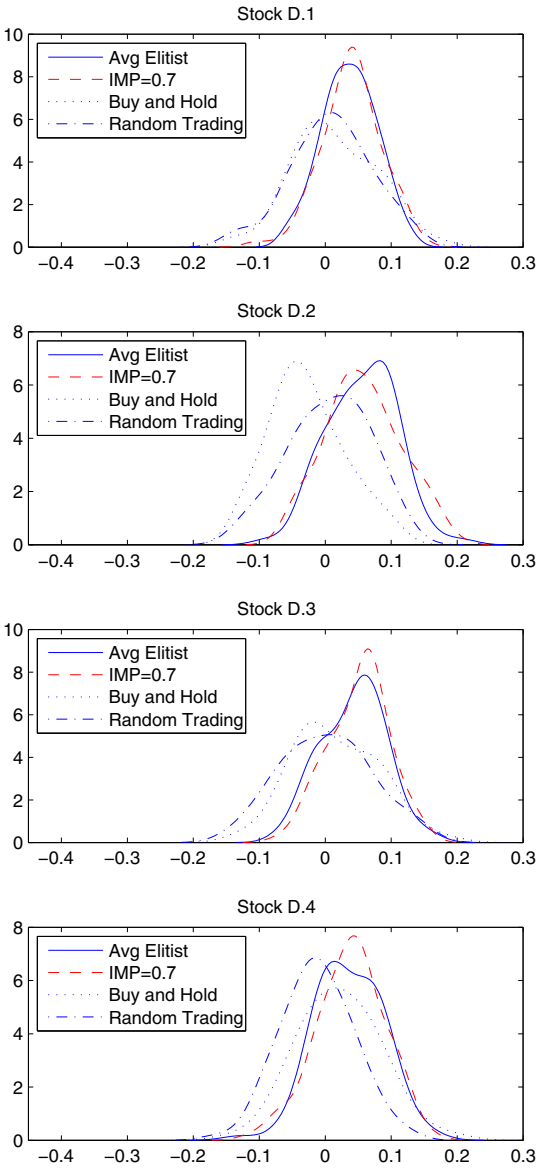


Fig. 4. The Sharpe ratios in Group D