Reinforcement Learning for Trading Systems and Portfolios: Immediate vs Future Rewards

John Moody*, Matthew Saffell, Yuansong Liao and Lizhong Wu
Oregon Graduate Institute, CSE Dept.
P.O. Box 91000, Portland, OR 97291–1000
{moody, saffell, liao, lwu}@cse.ogi.edu

Abstract

We propose to train trading systems and portfolios by optimizing financial objective functions via reinforcement learning. The performance functions that we consider as value functions are profit or wealth, the Sharpe ratio and our recently proposed *differential Sharpe ratio* for online learning. In Moody & Wu (1997), we presented empirical results in controlled experiments that demonstrated the efficacy of some of our methods for optimizing trading systems. Here we extend our previous work to the use of Q-Learning, a reinforcement learning technique that uses approximated future rewards to choose actions, and compare its performance to that of our previous systems which are trained to maximize immediate reward. We also provide new simulation results that demonstrate the presence of predictability in the monthly S&P 500 Stock Index for the 25 year period 1970 through 1994.

1 Introduction: Reinforcement Learning for Trading

The investor's or trader's ultimate goal is to optimize some relevant measure of trading system performance, such as profit, economic utility or risk-adjusted return. In this paper, we propose to use reinforcement learning to directly optimize such trading system performance functions, and we compare two different reinforcement learning methods. The first uses immediate rewards to train the trading systems, while the second (Q-Learning (Watkins 1989)) approximates discounted future rewards. These

* John Moody is also with Nonlinear Prediction Systems.

A.-P.N. Refenes et al. (eds.), Decision Technologies for Computational Finance, 129–140.

methodologies can be applied to optimizing systems designed to trade a single security or to trade a portfolio of securities. In addition, we propose a novel value function for risk adjusted return suitable for online learning: the *differential Sharpe ratio.*

Trading system profits depend upon sequences of interdependent decisions, and are thus path-dependent. Optimal trading decisions when the effects of transactions costs, market impact and taxes are included require knowledge of the current system state. Reinforcement learning provides a more elegant means for training trading systems when state-dependent transaction costs are included, than do more standard supervised approaches (Moody, Wu, Liao & Saffell 1998). The reinforcement learning algorithms used here include maximizing immediate reward and Q-Learning (Watkins 1989).

Though much theoretical progress has been made in recent years in the area of reinforcement learning, there have been relatively few successful, practical applications of the techniques. Notable examples include Neuro-gammon (Tesauro 1989), the asset trader of Neuneier (1996), an elevator scheduler (Crites & Barto 1996) and a space-shuttle payload scheduler (Zhang & Dietterich 1996). In this paper we present results for reinforcement learning trading systems that outperform the S&P 500 Stock Index over a 25-year test period, thus demonstrating the presence of predictable structure in US stock prices.

2 Structure of Trading Systems and Portfolios

2.1 Traders: Single Asset with Discrete Position Size

In this section, we consider performance functions for systems that trade a single security with price series z_t. The trader is assumed to take only long, neutral or short positions $F_t \in \{-1, 0, 1\}$ of constant magnitude. The constant magnitude assumption can be easily relaxed to enable better risk control. The position F_t is established or maintained at the end of each time interval t, and is re-assessed at the end of period $t + 1$. A trade is thus possible at the end of each time period, although nonzero trading costs will discourage excessive trading. A trading system return R_t is realized at the end of the time interval $(t - 1, t]$ and includes the profit or loss resulting from the position F_{t-1} held during that interval and any transaction cost incurred at time t due to a difference in the positions F_{t-1} and F_t.

In order to properly incorporate the effects of transactions costs, market impact and taxes in a trader's decision making, the trader must have internal state information and must therefore be recurrent. An example of a single asset trading system that could take into account transactions costs and market impact would be one with the following decision function: $F_t = F(\theta_t; F_{t-1}, I_t)$ with $I_t = \{z_t, z_{t-1}, z_{t-2}, \ldots; y_t, y_{t-1}, y_{t-2}, \ldots\}$