

Learning to Trade
via Direct Reinforcement
by John Moody and Matthew Saffell

presented by Dustin Boswell

April 23, 2003

Reinforcement Learning Methods

Value Learning

- Learns a value function $V(state, action)$
- Optimal action is implicit: $a^* = \arg \max_a V(state, a)$
- Ex: Q-Learning, TD-Learning, Advantage Updating
- Suitable in environments where rewards are not immediate:
Grid World, Checkers

Reinforcement Learning Methods

Value Learning

- Learns $V(state, action)$
- Optimal action is implicit.
- Ex: Q-Learning
- Suitable for delayed rewards

Direct Reinforcement Learning

- No value function
- Learns action policy directly
- Ex: Recurrent Reinforcement Learning (RRL) given by Moody
- Suitable in environments where rewards are immediate: Control problems, stock markets

Computational Finance Methods

Typical Approach

- 1) Learn to predict future prices
- 2) Take into account transaction costs, risk, etc...
- 3) Make trading decision

Direct Approach

- 1) Learn trading strategy

Notation

- Our agent trades fixed quantities of a security z .
- The price series is $\{z_1, z_2, \dots, z_t, \dots, z_T\}$ and corresponding price changes $r_t = z_t - z_{t-1}$
- At each time step, our position is
 $F_t \in \{long, neutral, short\} = \{+1, 0, -1\}$

Notation (cont.)

- We wish to learn a trading strategy $F_t = F(\theta_t; F_{i < t}, I_t)$
 θ is the set of parameters we are learning
 $I_t = \{z_{i \leq t}, \text{etc...}\}$ is our **Information** at time t
- For example, a system with $m + 1$ “autoregressive inputs”:
$$F_t = \text{sign}(uF_{t-1} + v_0r_t + v_1r_{t-1} + \cdots + v_mr_{t-m} + w)$$

$$\theta = \{u, v_i, w\}$$
- Whatever functional form used, $F(\theta)$ must be differentiable ($dF/d\theta$).

Profit & Wealth

- Daily Return: $R_t = r_t \cdot F_{t-1} - \delta|F_t - F_{t-1}|$
 δ is the transaction cost
- Total Profits: $P_T = \sum_{t=1}^T R_t$
i.e. no reinvesting
- Wealth: $W_T = W_0 + P_T$
- Utility: $U_T = U(R_1, \dots, R_T; W_0)$
Ex: $U_T = W_T$ measures profit without regard to risk

Measuring Utility: Sharpe Ratio

- Let U_T be the Sharpe Ratio: $S_T = \frac{\text{Average}(R_t)}{\text{StandardDeviation}(R_t)}$
- To maximize U_T , we will need an expression for dU_T/dR_t .
- But dS_T/dR_t is hard to work with.
- Instead, Moody comes up with an approximation...

Approximating the Sharpe Ratio

- First, let us define “exponential moving estimates”
 $A_t = E(R_t)$ and $B_t = E(R_t^2)$.

$$A_t = A_{t-1} + \eta \Delta A_t \quad \left(\frac{dA_t}{d\eta} = \Delta A_t = R_t - A_{t-1} \right)$$

$$B_t = B_{t-1} + \eta \Delta B_t \quad \left(\frac{dB_t}{d\eta} = \Delta B_t = R_t^2 - B_{t-1} \right)$$

- η^{-1} is the size of the window. (Moody sets $\eta = 0.01$.)
- So now we define the Sharpe Ratio in terms of these estimates.

$$S_t = \frac{A_t}{(B_t - A_t^2)^{1/2}} \approx \frac{\text{Ave}(R_t)}{\text{Std}(R_t)}$$

Approximating the Sharpe Ratio (cont.)

- Now Taylor expand S_t about η (not obvious why, but it helps the algebra).

$$S_t \approx S_{t-1} + \eta \frac{dS_t}{d\eta} \Big|_{\eta=0} + O(\eta^2)$$

- Ignore $O(\eta^2)$. Notice only the **center** term depends on R_t .

This lets us say $\frac{dS_t}{dR_t} \approx \frac{d}{dR_t} \eta D_t$

$$D_t \equiv \frac{dS_t}{d\eta} = \frac{d}{d\eta} \frac{A_t}{(B_t - A_t^2)^{1/2}}$$

$$D_t \equiv \frac{B_{t-1} \Delta A_t - \frac{1}{2} A_{t-1} \Delta B_t}{(B_{t-1} - A_{t-1}^2)^{3/2}}$$

Measuring Utility: Sharpe Ratio (cont.)

- Finally, we arrive at

$$\frac{dD_t}{dR_t} = \frac{B_{t-1} - A_{t-1}R_t}{(B_{t-1} - A_{t-1}^2)^{3/2}}$$

- Hence we've gotten our expression :

$$\frac{dU_t}{dR_t} = \frac{dS_t}{dR_t} \approx \eta \frac{dD_t}{dR_t}$$

- Notice the largest improvement to U_t is when

$$R_t^* = B_{t-1}/A_{t-1}$$

- So the Sharpe Ratio penalizes gains larger than R_t^* !

Measuring Utility: Sterling Ratio

- $Sterling_T = \frac{\text{Average Yearly Return}}{\text{Maximum Draw Down}}$
- $MDD = \max_{i < j} (z_i - z_j)$
where $j - i \leq 1$ year
- Useful for mutual fund managers wanting to minimize displeased customers
- Unfortunately, difficult to deal with analytically

Measuring Utility: Double Deviation Ratio

- $DDR_T = \frac{\text{Average}(R_t)}{DD_T}$
- $DD_T = \left(\frac{1}{T} \sum_{t=1}^T \min\{R_t, 0\}^2 \right)^{1/2}$
- DDR rewards large average positive returns,
penalizes “risky” (large negative) returns

Learning Parameters through Gradient Ascent

- Given $F_t(\theta)$ we want to adjust θ to maximize U_T :

$$\frac{dU_T(\theta)}{d\theta} = \sum_{t=1}^T \frac{dU_T}{dR_t} \left\{ \frac{dR_t}{dF_t} \frac{dF_t}{d\theta} + \frac{dR_t}{dF_{t-1}} \frac{dF_{t-1}}{d\theta} \right\}$$

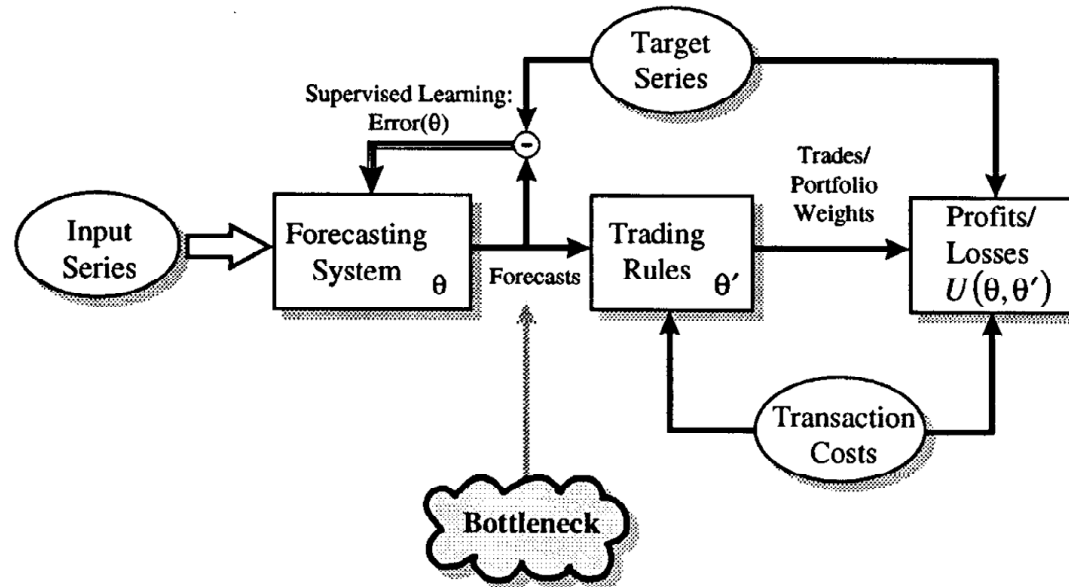
$$\Delta\theta = \rho \frac{dU_T(\theta)}{d\theta} \quad (\rho \text{ is the learning rate})$$

- $\frac{dU_T}{dR_t}$: we saw this from before

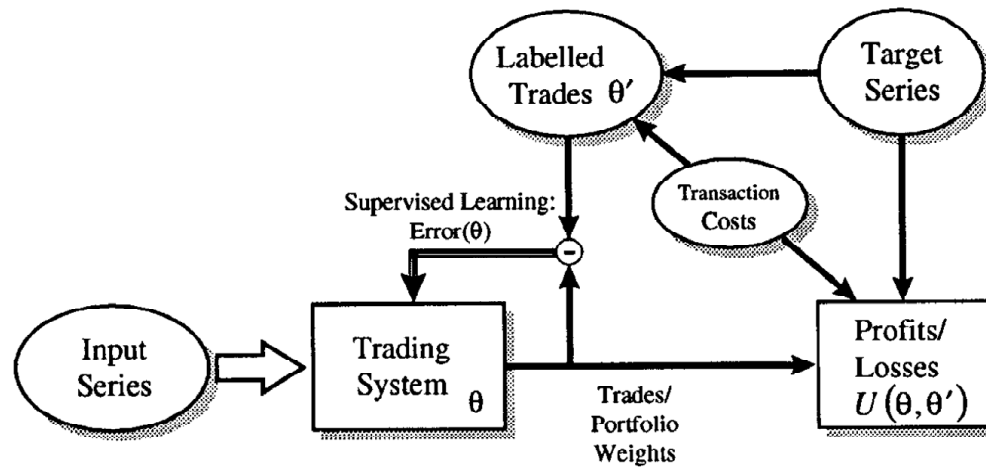
- $\frac{dR_t}{dF_t}$: easy to determine

- $\frac{dF_t}{d\theta}$: $\frac{dF_t}{d\theta} = \frac{\partial F_t}{\partial \theta} + \frac{\partial F_t}{\partial F_{t-1}} \frac{dF_{t-1}}{d\theta}$ recursively ...

Training to make forecasts



Training with labeled data (example trades).



Empirical Results: Data Sets Used

DataSet	Goal
1) Artificial Time Series	Show the system can learn
2) Foreign Exchange Data	Show the system can learn a profitable strategy on real data
3) S&P and Treasury Bill	Show RRL is better than Q-Learning

1) Artificial Data

- Data is designed to have a “tradeable structure”
- They generate log-normal random walks, but with autoregressive trends:

$$\text{Trend variable: } \beta(t) = \alpha\beta(t-1) + \nu(t)$$

$$\text{Log price: } p(t) = p(t-1) + \beta(t-1) + \epsilon(t)$$

- $\epsilon(t)$, $\nu(t)$ are “noise” terms with zero mean, unit variance
- $\alpha < 1$ sets how the “autoregressiveness”
- $z_t = \exp(p(t))$ is used to generate 10,000-point price series.

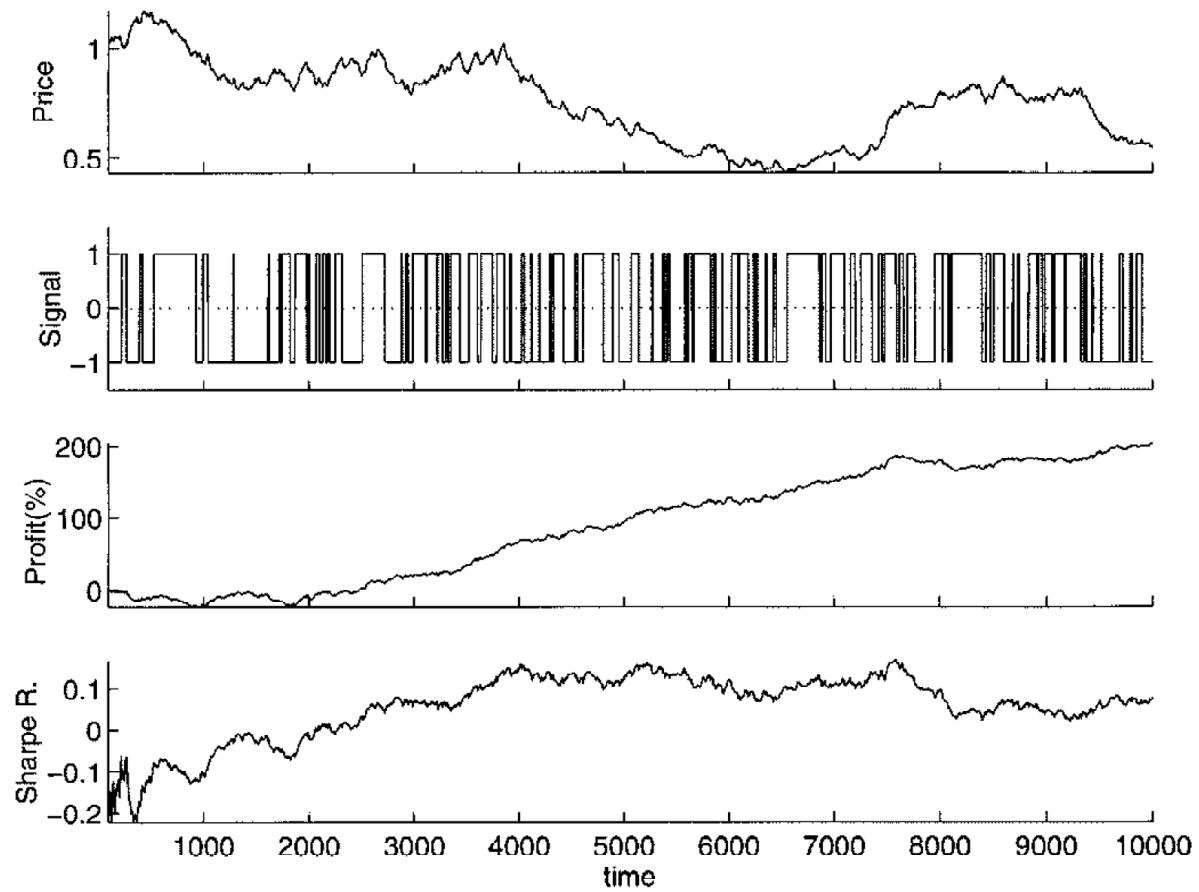
1) Artificial Data - System Details

- The trading function has autoregressive inputs (matches the data):

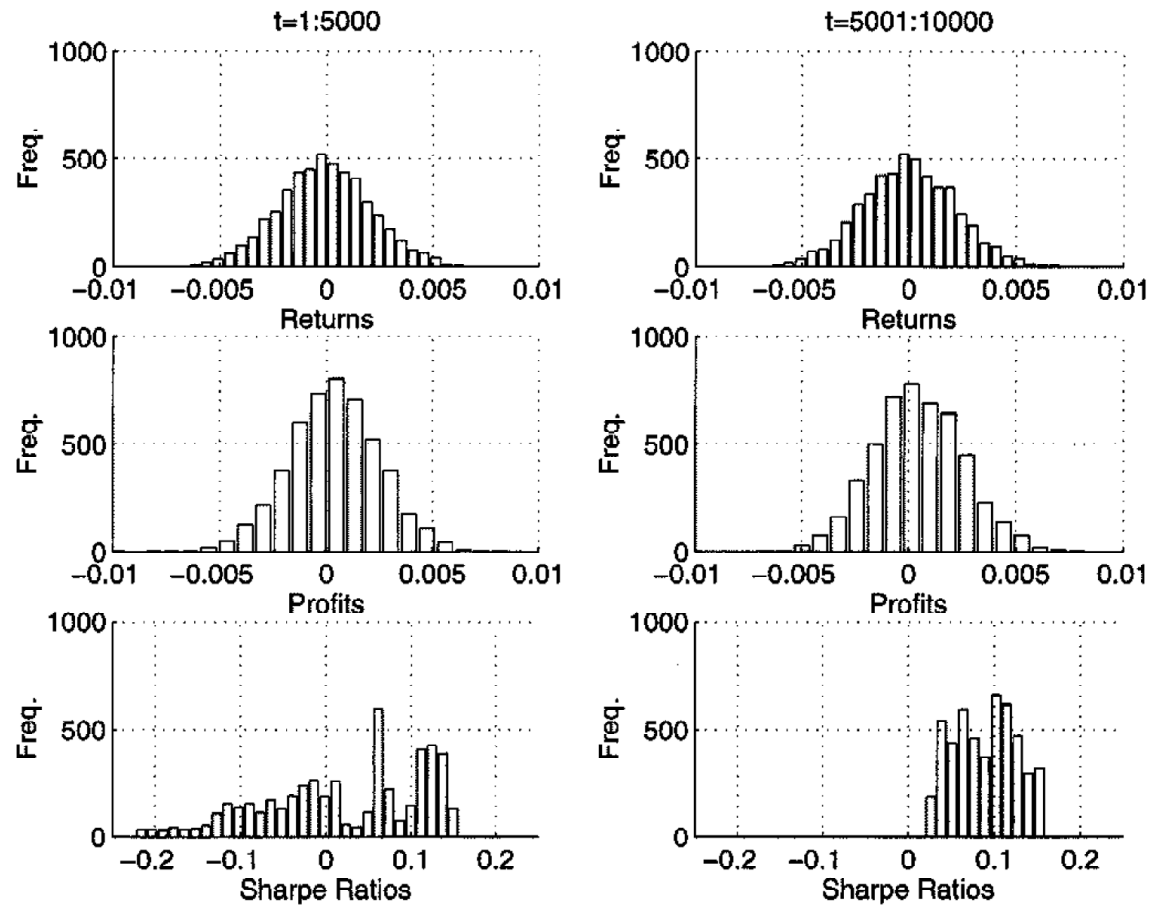
$$F_t = \text{sign}(uF_{t-1} + v_0r_t + v_1r_{t-1} + \cdots v_7r_{t-7} + w)$$

- Transaction cost: $\delta = 0.5\%$
- Learns to maximize the Differential Sharpe Ratio

Artificial Prices, and Results



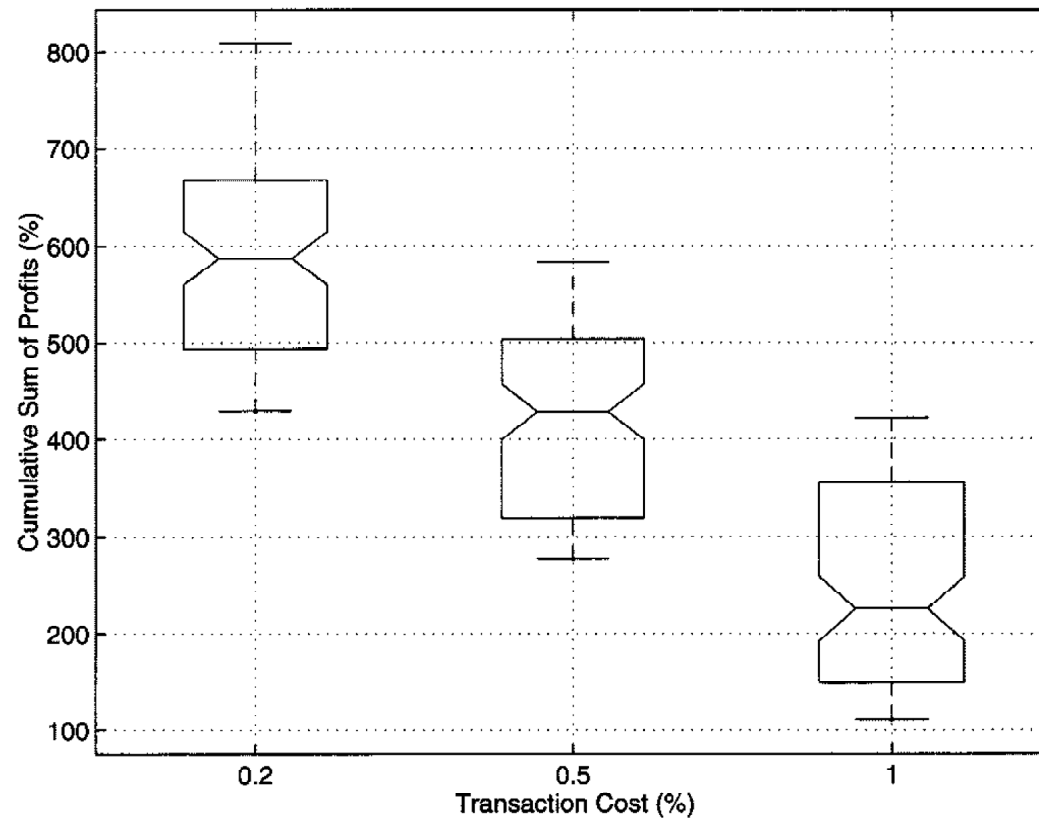
Histograms of Artificial Data and Results



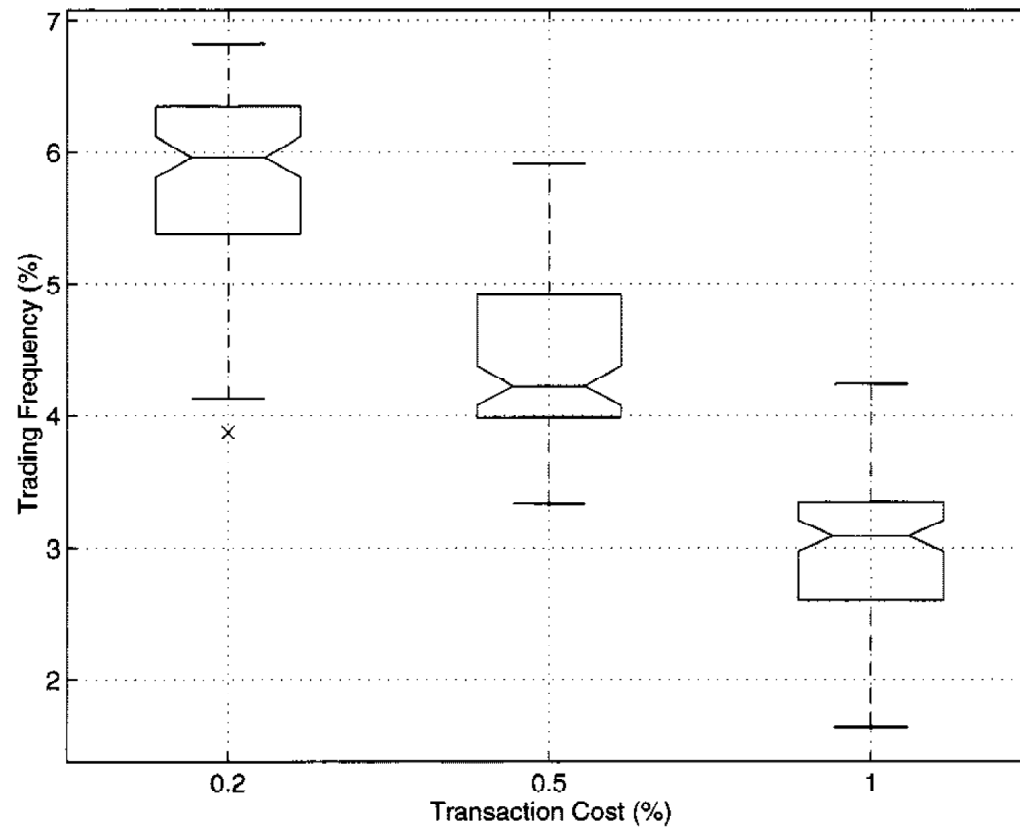
Artificial Data (continued)

- How do transaction costs affect trading performance?
- Repeat the previous experiments 100 times...
try $\delta = 0.2\%, 0.5\%, 1.0\%$
- Hypothesis: lower costs should allow:
 - more trading
 - more profits
 - better Sharpe Ratio

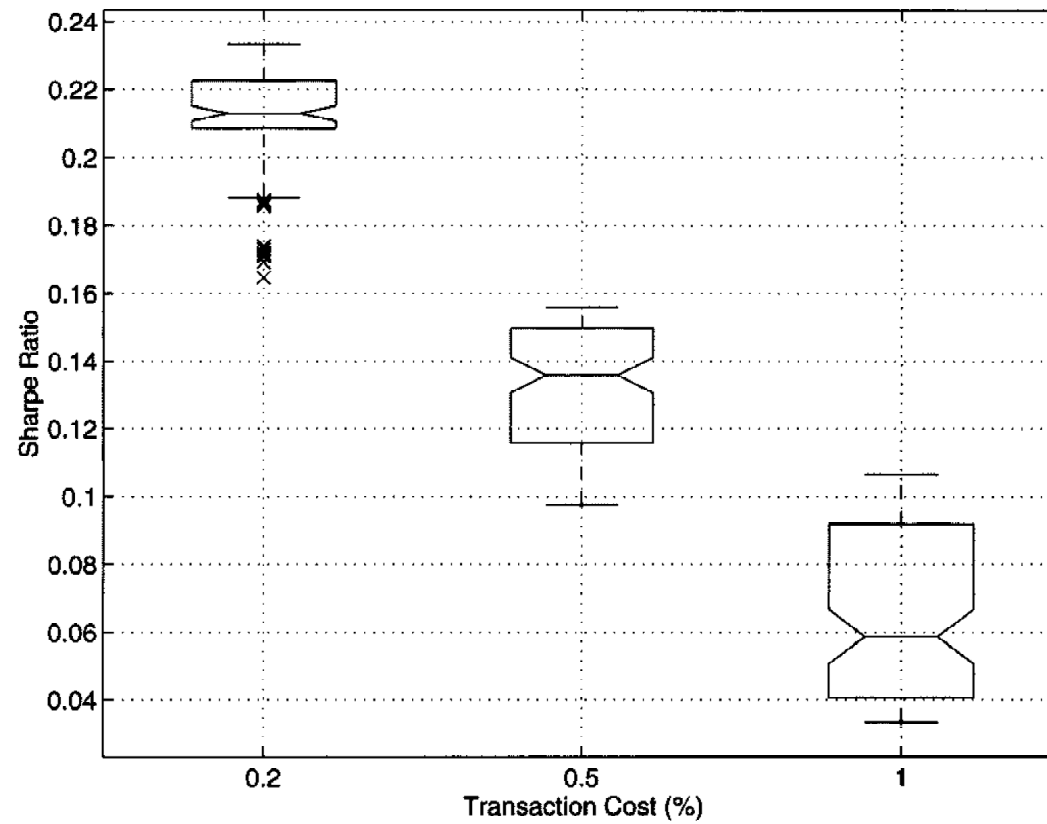
Boxplots of how transaction costs affect **profits**



Boxplots of how transaction costs affect trading frequency



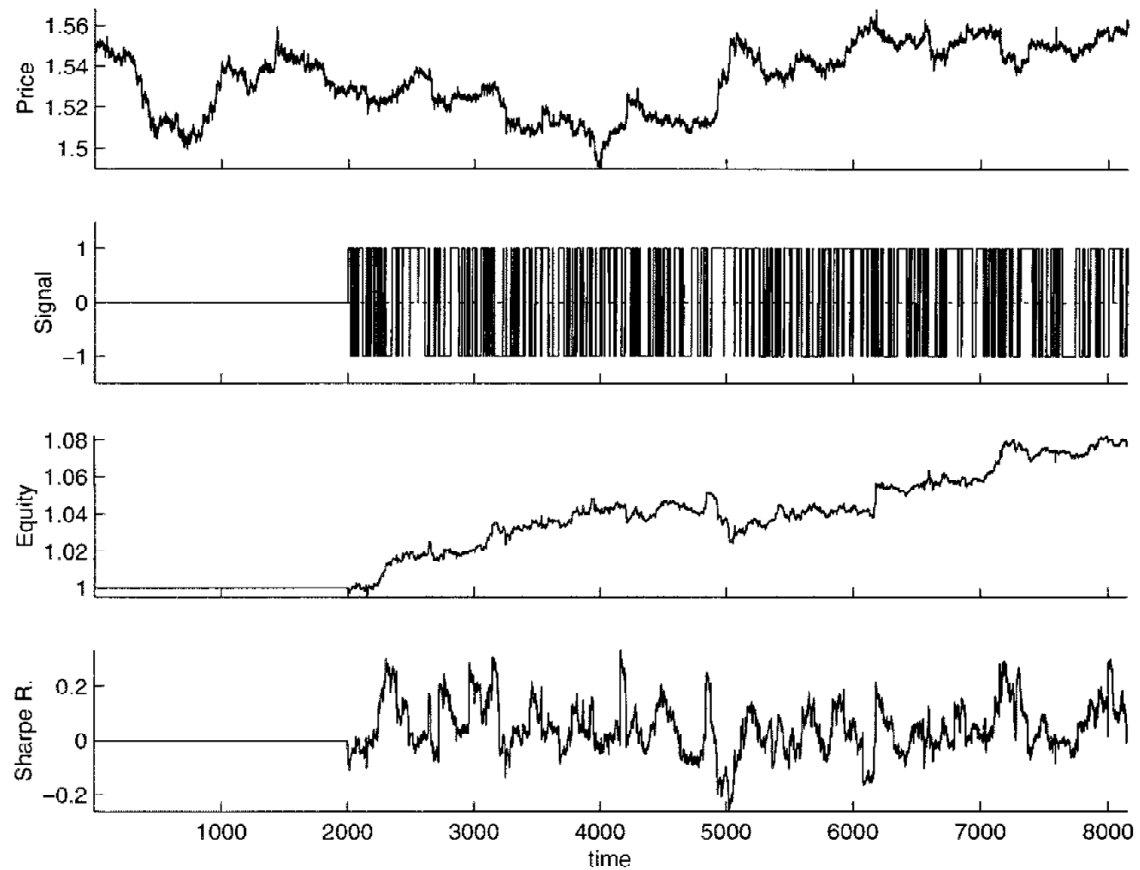
Boxplots of how transaction costs affect **Sharpe Ratio**



2) Foreign Exchange Data

- US Dollar vs. British Pound
- 8 months of data (half-hour quotes) during 1996
- Same autoregressive inputs as Artificial Data experiment?
(the paper was unclear)
- The system is trained to maximize the Downside Deviation Ratio.
- Transaction cost is the bid-ask spread
(which has a typical average but is not fixed).

Foreign Exchange Prices, and Results



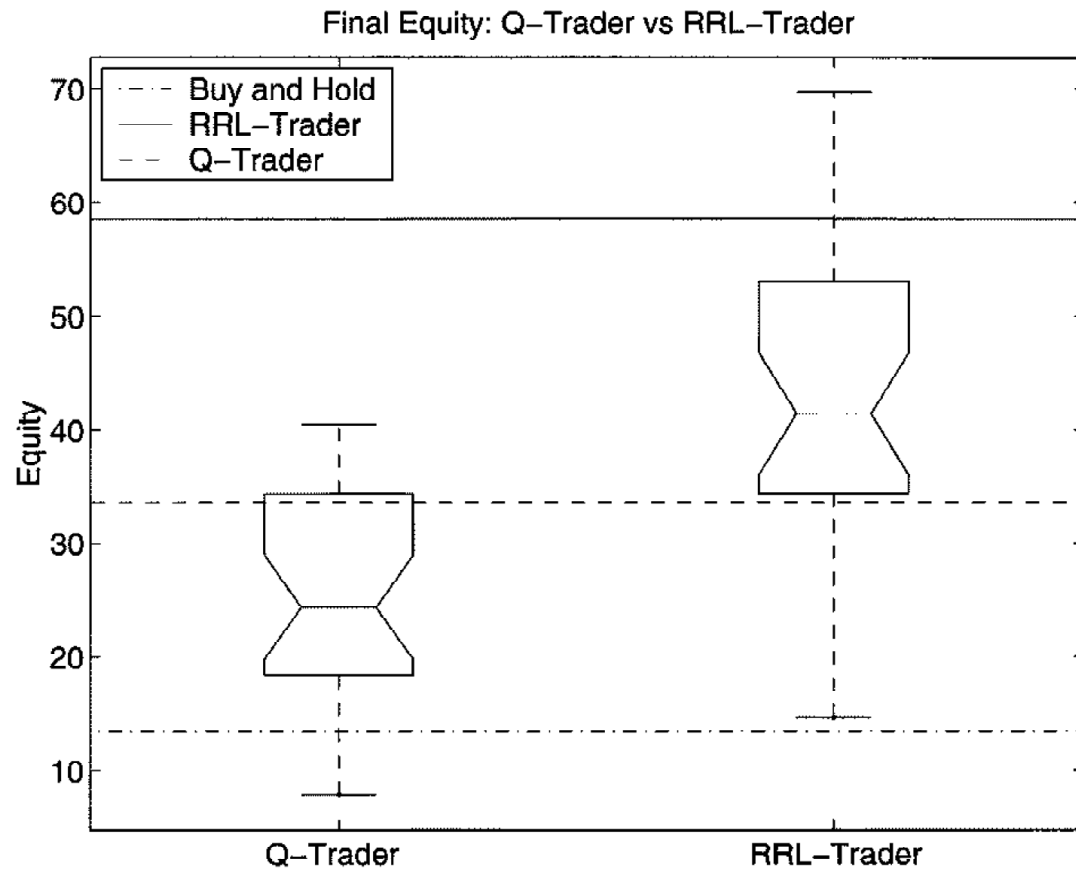
Foreign Exchange Result Summary

- 15% annualized return, Sharpe Ratio of 2.3
- (S&P index gets roughly 15% return, *Sharpe* < 1)
- Trading frequency: trades are made roughly once every 5 hours
- It's difficult to say how well this would have done in real world environment (since you can't simulate all market frictions).

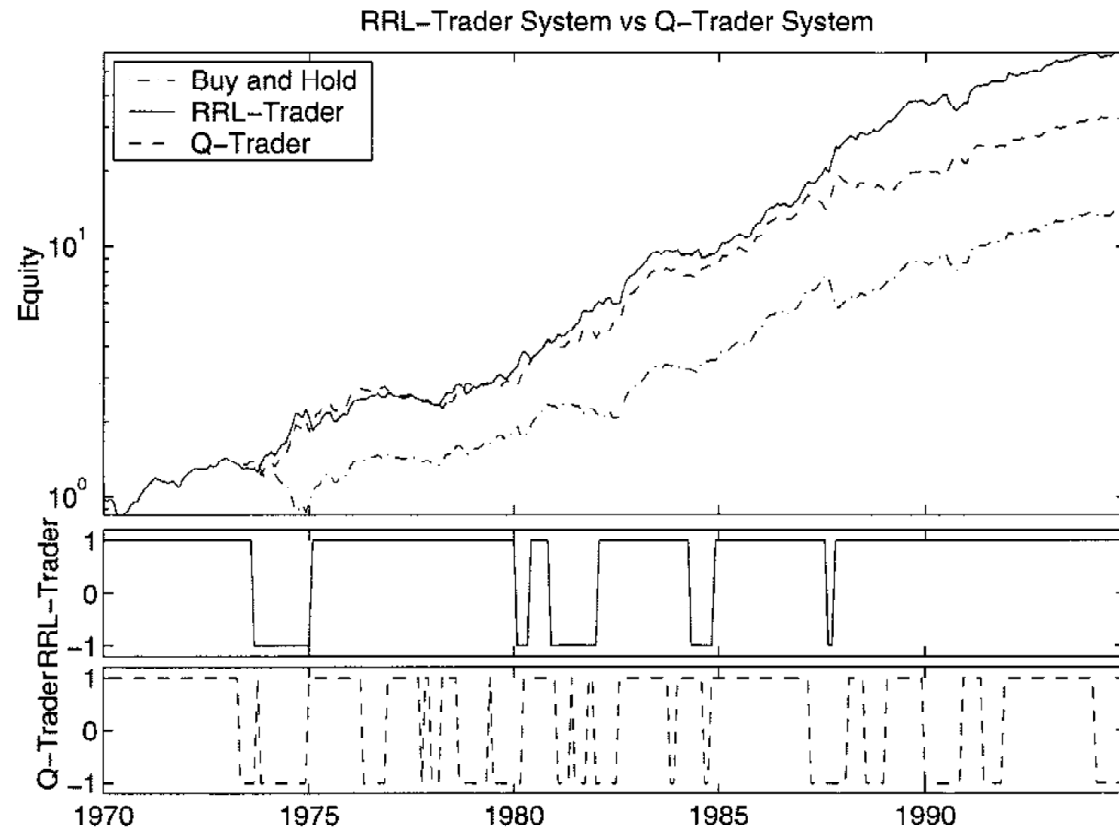
3) US Stock Market Data: Q-Trader against RRL-Trader

- S&P vs. Treasury Bill
- 25 years of data: 1970-1994
- System is trained on previous 20 years (sliding window)
- The Information I_t also includes macroeconomic data
- Also implement Q-Learning (actually, a variant called Advantage Updating) to compare.

Q-Trader vs. RRL-Trader Results



Q-Trader vs. RRL-Trader Results (continued)



Conclusions

- Moody makes the case that RRL is better than Q-Learning for trading since it is a simpler approach.
(Simpler is better - a recurring theme in this class.)
- I'm not sure I agree personally with some of his methods.
- Moody has set up an interesting method for learning directly, but hasn't addressed the problem of choosing a good trading model.