www.joim.com

Journal Of Investment Management, Vol. 22, No. 4, (2024), pp. 26–40

TRAINING MACHINES TO TRADE STOCKS*

Dilip B. Madan[a] *and King Wang*[b]

Machines are trained to trade stocks by developing an investment policy for stock investment in a Markovian context. Importantly, the investment actions impact just the immediate reward and not the state transitions. The policies are designed to maximize a nonlinear expectation of the undiscounted sum of future rewards using the methods of Now Decision Theory. The nonlinear expectations, unlike expectations, are rendered risk sensitive using a distortion of probabilities. The distortions employed need not be concave and display regions of convexity making them volatility desiring at low volatility levels. The technology is illustrated by trading 589 stocks over 15 years using the policy function.

1 Introduction

Many researchers working with a variety of technologies applied to the problem of predicting movements in stock prices contemplate the possibility of market inefficiencies in the sense that actual prices may at trade times be different from their intrinsic or equilibrium values to which they will converge in the near future. By way of examples we cite Khoa and Huynh (2021), Chavarnakul and Enke (2008), Gao *et al.* (2017), and Wu *et al.* (2020). The efficient markets hypothesis (Fama, 1965a, 1965b, 1970; Shleifer, 2000) asserts that actual prices in efficient markets are good approximations of their intrinsic values and there is no point in trying to use information of various types available to market participants to predict price changes. The efficient markets theory more formally asserts that price changes are independent random variables with no information about their future distributions available in past or currently available data.

Let us first comment on what exactly can be meant by such assertions and what import do they have on the activities of investors if any.

- It is stated (Cootner, 1962) in discussions of market efficiency that under such efficiency current prices are conditional expectations of future prices, conditional on all available information. But this cannot be the case as the expected price change must then be zero and there will then not be a positive mean rate of return, which in general there is.
- It is also asserted on occasion (Fama, 1965b) that the available information may not be

[a]Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, USA.
E-mail: dbm@umd.edu
[b]Derivative Product Strats, Morgan Stanley, 1585 Broadway, 5th Floor, New York, NY 10036, USA.

*This paper is the private opinion of the authors and does not necessarily reflect the policy and views of Morgan Stanley.

such as to enhance expected gains. From today's standpoint this may be interpreted as the impossibility of an arbitrage strategy. We now understand the absence of arbitrage to mean that the price is a conditional expectation under a change of measure (Harrison and Kreps, 1979; Harrison and Pliska, 1981; Delbaen and Schachermayer, 1994). However, this says very little for if prices may both rise and fall then the current price is an expectation under a change of measure (Jacod and Shiryaev, 1998). Furthermore, an arbitrage is an extreme situation of enhancing gains. The acceptability index introduced in Cherny and Madan (2009) measures trading performance as an approximation to arbitrage. An arbitrage has an acceptability index of infinity while most trading strategies fail to reach an acceptability index of unity. There are to be found many degrees for the acceptability index reached by many different trading strategies in possibly efficient markets and it is not clear what performance level is being excluded by efficiency.

- Departing from arbitrage considerations and considering equilibrium prices instead, the view is taken that returns from trading strategies in efficient markets are not able to earn abnormal risk-adjusted returns. In this interpretation of efficiency the expected excess returns from employing trading strategies are consistent with the covariation of returns with the return on the change of measure density (Back, 1991). However, the set of acceptable risks on this interpretation is very large and includes all return outcomes that have a positive covariation with the measure change density. Generally the measure change density is not unique as markets are not complete (Harrison and Pliska, 1983). The collection of acceptable risks may be reduced by modeling risk acceptability as a positive covariation with an entire family of measure change densities. The scenario tests of the Federal Reserve, post the crisis of 2008, are a movement in this direction.
- With regard to an import for investor decisions the random walk hypothesis suggests (Fama, 1965a) that investors should focus their attention on spreading their risks across the different return distributions available to them and not be concerned about predicting or managing their responses to movements in these distributions. In particular, they need not answer how much to invest in the market and how this may change but just to make decisions on allocations or portfolio theory. Yet, investors do need to make decisions on how much to invest and the mathematics of investment should guide both humans and machines in answering such questions.

Consider now the efficient markets hypothesis in the context of a single risk asset with a return distribution, possibly independent across periods. Appeals to the central limit theorem (Fama, 1965a) led researchers to postulate a normal distribution. However, empirical research early on provided evidence for the distributions being leptokurtic (Kendall, 1953; Moore, 1962; Alexander, 1961). It is now well understood that the normal distribution is not the only limit distribution and the entire class of limit distributions is given by the self-decomposable laws (Lévy, 1937; Khintchine, 1938; Sato, 1999) many of which have leptokurtic distributions with finite variance. A particularly simple and empirically adequate class of such distributions is provided by the four parameter class of bilateral gamma densities (Küchler and Tappe, 2008) that adequately capture the first four moments of drift, volatility, skewness and kurtosis of the continuously compounded return defined by the log price relative. The investor's interest, however, is in the rate of return over a horizon. Let X denote the continuously compounded return with a bilateral gamma density of $f(x)$. The investor's interest is in the return R

where

$$R = e^X - 1.$$

The bilateral gamma process is an infinitely divisible pure jump process with Lévy density $k(x)$ that announces the arrival rate of jumps of size xdx as $k(x)dx$. If the stock price process is also pure jump it is shown in Madan and Wang (2024) that the mean rate of return μ on the stock is given by

$$\begin{aligned}\mu &= E[R] \\ &= \int_{-\infty}^{\infty} (e^x - 1)k(x)dx.\end{aligned}$$

The mean rate of return is thus predicted by the structure of the shocks to the price process and it may be inferred from the levels of volatility, skewness, and kurtosis. The higher moments of distributions are known to have some degree of predictability as evidenced by the literature on stochastic volatility, see for example Shepard (2005). Clearly in periods with a substantially positive mean one should take a small long position in the stock and a small short position when the mean is substantially negative. However, given that the risk is larger for sizable positions there should be an optimal level of investment mathematically recommended in the light of the mathematical objective guiding the investment. The mathematics of investment thus goes beyond portfolio theory, market efficiency notwithstanding. The present paper is a contribution in this direction and brings together the issues to be addressed and resolved in thus developing mathematical investment theory.

In developing such a mathematical investment theory many researchers focus attention on the prediction of price changes. Recent examples include Zhang *et al.* (2021) and Brogaard and Zareei (2023). The transformation of a price prediction into a trading strategy requires the reversal of trades at a horizon matching that of the return prediction and this is typically a short horizon subject to high transaction costs. The approach taken here shifts attention toward a longer-term attention on the economic value of the asset in place. This could be considered to be akin to a fundamental of intrinsic value. However, concepts of intrinsic value are unduly influenced by assumptions of the law of one price and markets can be fickle and change their view on the valuation of economic activities. The objectives we employ are then related to concepts of conservative market valuation as they are now embedded in the theory of nonlinear expectations.

The problem of training a machine to trade stocks is thus addressed and a solution is designed and implemented on trading 589 stocks over the period January 3, 2008 to March 31, 2022. Different trading objectives are employed in the design of the trading policy and their relative performances are compared. The general structure of the trading policy is the construction of a deterministic function of statistics related to the stock at any time that one may term the state of the stock at the particular time with the function recommending the number of dollars to be invested in the stock at this time. Let the vector $s = (s_k, k = 1, \ldots, K)$ denote the statistics to be employed in the investment policy decision. The policy function to be constructed is the function $\pi(s)$ defining the number of dollars to be invested in the stock with the statistics s.

The policy is implemented on daily data for the 589 stocks by investing in stock i on day t the amount a_{it} that may be positive or negative where

$$a_{it} = \pi(s_{it}), \tag{1}$$

and s_{it} are the statistics for stock i at time t. With a view toward eventually unwinding all open positions, a fixed percentage of the aggregate net legacy positions on account of previous decisions for all stocks are reversed each day. The present paper reports results for a five percent daily unwind of legacy positions.

The immediate return to the investment of a_{it} dollars in stock i on day t is

$$R_{it} = a_{it}\left(\frac{S_{it+1} - S_{it}}{S_{it}}\right), \tag{2}$$

where S_{it} is the closing price of stock i on day t.

In addition to this immediate return there is a state transition for stock i to the state $s_{i,t+1}$. The investor is presumed to have an interest in the sequence of returns over the long term as opposed to just the immediate return. In this longer-term perspective the investor will have to evaluate the sequence of states, actions, rewards, and next states or the sequence

$$s_{it,}a_{it}, R_{it}, s_{it+1}, a_{it+1} \ldots . \tag{3}$$

It is then quite natural to try and apply the methods of Reinforcement Learning (Sutton and Barto, 2018) to build simultaneously a state valuation function $V(s)$ along with a policy function $\pi(s)$, whereby the value of a state is the value of the objective under an optimal policy and the policy is optimal for the resulting value function. In this formulation there is a presumed transition probability $P(s' \mid s, a)$ and for the maximization of expected discounted rewards, with discount rate γ, the value function solves the fixed point problem

$$V(s) = E[aR + \gamma V(s')]. \tag{4}$$

There are, however, three issues with such an application addressed in this paper.

- The first is the maximization of expected rewards as it is not risk sensitive and is thereby not an appropriate financial objective.
- The second is the presence of discounting. In principle one may have a long sequence of rewards over a short time span calling into question the validity of a discount rate as time is no longer an issue. We propose to implement the principles of "Now Decision Theory," as expounded in Madan *et al.* (2023) that avoids discounting.
- The third and probably the most critical is the recognition that in financial markets the actions of individual market participants have nothing to do with state transitions. These are determined by economic circumstances quite divorced from one's own actions. Hence the relevance of the state transition functions defined by $P(s'|s, a)$ is called into question. A solution addressing all three concerns is proposed and implemented here.

It is helpful to fix ideas and work with a concrete example for the state space being employed in the policy construction. For this purpose consider four state variables defined as weighted averages of past returns and squared returns. This choice of state variables is just for the purpose of illustrating all the steps to be taken and implemented in the design of a policy function. It is a very simplistic choice and not one that is expected to yield a strong performance but serves the purpose of the illustration. On the data of 589 stock prices from January 3, 2007 to March 31, 2022 comprising a period of 3839 days, let r_{it} be the continuously compounded return on stock i between day t and $t - 1$. Now define $y_{it,k}$, $v_{it,k}$, for $k = 1, 2$ by

$$y_{it,k} = \Delta_{jk} \sum_{l=0}^{\infty} \rho_1^{(t-l)} (t-l)^{k-1} r_{i(t-l)} \tag{5}$$

$$v_{it,jk} = \Delta_{jk} \sum_{l=0}^{\infty} \rho_2^{(t-l)} (t-l)^{k-1} r_{i(t-l)}^2 \tag{6}$$

where $\rho_1 = 0.8187$, $\rho_2 = 0.6065$, and Δ_{jk} are normalizing constants. The use of both geometric and power weighting of lagged returns is motivated by recent research into the relevance of Tempered Fractional processes as studied in Madan and Wang (2022).

The state vector is then four dimensional with

$$s_{it} = (y_{it,11}, y_{it,12}, v_{it,21}, v_{it,22}). \qquad (7)$$

The solution developed employs three feed-forward neural nets. Reinforcement Learning methodologies proposing neural net candidates for value and policy functions are adopted. In addition, a multioutput neural net is built for the state transition specification. The value and policy functions are constructed on a quantized sample of states and then extended to the full space via neural nets as opposed to adjusting their parameters using a gradient descent. The policy functions employ a full optimization on the quantized training sample. The policy functions are retrained every 63 days on data for the last 63 days and the most recently trained policy function is called upon to define the new trading positions at any date.

The outline of the rest of this paper is as follows. Section 2 addresses the adequacy of the state specification. Section 3 presents the risk sensitivity solution adopted in the objective. Section 4 reviews the Now Decision Theory resolution for the absence of discounting. Section 5 describes the underlying state transition mechanism. Section 6 reports on the implementation of the state transition designs for the states defined by Equation (7). The policy function constructions are reported in Section 7. The specific probability distortions employed are described in Section 8. Section 9 provides results on trading the policy functions. Section 10 discusses new research directions. Section 11 concludes.

2 Adequacy of State Specification

With a view toward commenting on the adequacy of the selected state variables for making investment decisions on the stocks we regressed the forward return on the stocks over the next day on the selected state variables. The data on state variables and forward returns was pooled across the 589 stocks and the 3,818 days for a full sample of 2, 248, 802 observations. The result of the regression is presented in Table 1.

Table 1 Regression of forward returns on state variables.

Variable	Coefficient	t-Stat
Constant	0.0532	46.19
y_1	−0.1696	−73.43
y_2	0.0918	27.37
v_1	0.0084	118.6
v_2	−0.0084	−107.9
R^2	0.694%	

The R^2 is low as expected and all the variables are significant. The set of illustrative state variables are related to future returns and are thereby a potential set of state variables to consider as inputs of a trading policy.

3 Risk Sensitivity of Objectives

The maximization of expected accumulated returns, discounted or otherwise, ignores the risk embedded in stock returns. It is well understood that for two assets with the same expected return and different risk levels, there is a general preference in markets for the asset with lower risk level. The valuation or selection of investment opportunities must pay attention to risk levels in addition to the expected returns. Traditionally, and in a substantial part of the literature, this has been done by employing penalties for variance in a mean–variance objective, or by the maximization of expected utility.

Objections to both these objectives are described in Madan *et al.* (2022). Mean–variance analysis measures rewards in dollars and risks in squared dollars with the latter eventually and arbitrarily dominating the former to produce finite solutions that are then questionable. Expected utility theory on the other hand makes current wealth, unrealistically, a state variable and forces one to model the dissatisfaction with getting rich as embedded in the diminishing utility of money. We wish to explicitly exclude current wealth as a relevant state variable and it is excluded from our illustrative example. For these reasons, neither of these traditional objectives is appropriate for the construction of a trading policy.

Instead, we consider the theory of nonlinear valuation presented in Madan and Schoutens (2016, 2022) based on the definition of acceptable risk as developed in Artzner *et al.* (1999). The theory of acceptable risk explicitly distinguishes the treatments of gains from losses by defining all nonnegative outcomes as acceptable with no nonpositive outcomes being acceptable. Unlike the use of variance as a risk measure in mean–variance theory and locally in expected utility theory, risks are recognized as inherently asymmetric with respect to the sign of the outcome. The resulting nonlinear valuation is seen to be an infimum of expectations taken with respect to a family of probability measures. Implementation of such nonlinear valuations is significantly hindered, however, by the necessity of defining the full probability space and the set of alternative measures on this probability space over which the infimum is to be conducted.

Our ability to define complete probability spaces in practical situations, let alone a family of measures on it, is very limited and unlikely to deliver an implementable solution. What one may hope to learn from experience is the distribution of outcomes with little knowledge of their relationship with underlying events. The question of risk acceptability can then shift from the acceptability of random variables on a probability space, to that of acceptable distributions of real-valued return outcomes. The latter is a considerably more manageable problem with simpler solutions being possible. Equivalently, one may ask the question about the acceptability of inverse distribution functions that are real valued nondecreasing functions on the unit interval. For two such functions of the same uniform variate we may additionally ask for the level of acceptability to be additive. Under such assumptions Kusuoka (2001) showed that acceptability reduces to a positive distorted expectation using a concave distortion.

More exactly, let X be a random variable with distribution function $F_X(x)$. Further let $\Psi(u)$, $0 \leq u \leq 1$ be a concave distribution function define on the unit interval. The acceptability of X requires that the distorted expectation $\mathcal{E}(X)$ be nonnegative, where

$$\mathcal{E}(X) = -\int_{-\infty}^{0} \Psi(F_X(x))dx + \int_{0}^{\infty} (1 - \Psi(F_X(x)))dx. \quad (8)$$

The distorted expectation is easily seen to be an expectation of X conducted with respect to the distorted distribution $\Psi(F_X(x))$. It follows from the domination of $F_X(x)$ by $\Psi(F_X(x))$ that the distorted expectation lies below the expectation. One may write the distorted expectation more symmetrically on defining $\widehat{F}_X(x) = 1 - F_X(x)$ and $\widehat{\Psi}(u) = 1 - \Psi(1 - u)$ as

$$\mathcal{E}(X) = -\int_{-\infty}^{0} \Psi(F_X(x))dx + \int_{0}^{\infty} \widehat{\Psi}(\widehat{F}_X(x))dx. \quad (9)$$

For reasons outlined in Madan and Wang (2022a, 2022b) one may consider for the design of policy functions distortions that are not concave but have convex regions near zero, unity, or both. They are, however, nondecreasing functions that are zero at zero and one at unity. One may also allow for some regions with a negative derivative by considering

$$\widetilde{\Psi}(u) = (1 + \lambda)\Psi(u) - \lambda u, \tag{10}$$

where Ψ is nondecreasing.

Risk sensitivity is incorporated into the objective by replacing the expectation operator in Equation (4) by a suitably selected distorted expectation operator.

4 Now Decision Theory and the Absence of Discounting

Madan *et al.* (2023) introduce the subject of Now Decision Theory in which the near future is folded into current time with the consequence that in the fixed point equation (4) the value of the future state given by $V(s')$ is no longer discounting when it is added to the immediate reward R. It is observed then that the value function is only determined up shift and scale. Any value function may be arbitrarily shifted and scaled and it remains a value function. The invariance under shift and scale is an important property to have for otherwise one is claiming to have determined uniquely the units of measurement for the value function. Such unique determination is akin to being able to assert that temperature is to be measured in Fahrenheit degrees and the use of the Celsius scale is not permissible or somehow incorrect. The unique determination of value independent of shift and scale may itself be seen as possibly erroneous.

The issue of fixing the shift and scale is addressed in Madan *et al.* (2023) by fixing the value function at two selected states *a priori*. In the applications of this paper we set the value at asset return percentiles of 25 and 75 at -100 and 100, respectively. Once this is done, the problem turns to determining the value function at the other states. This problem is shown in Madan *et al.* (2023) to be a discounted problem with discount rates that are both state contingent and endogenously defined in terms of the probabilities of transitioning to the states where the valuations have been fixed.

In the applications presented we update policy functions given a value function, then define an updated value function from the updated policy function which is shifted and scaled to preserve the valuations of -100 and 100 at the initially selected points. The updated value function is then used to update another policy function until the successive policy functions are observed to have converged. The convergence is anticipated from the nature of the problem as a discounted problem with state contingent discounting. The final policy function is taken as the recommended policy.

5 State Transitions

For many applications it is envisaged that one will have time series data on both the state descriptions for all stocks on all days along with the return on the next day. State transition laws must define $P(s'|s, x)$, where s' is the next state, x a prospective return and s the current state. For our selected state specification the dimensions of s and s' are four and x is univariate. Consider a training time t_n with the training to be based on current state observations s_{it}, for all stocks indexed by $i = 1, \ldots, M$, and $t = t_n - j$, for say $j = 0, 1, \ldots, L$ where L is the number of lags in the training set. We also have available for the training the set of returns x_{it}, $t = t_{n+1} - j$, $j = 0, 1, \ldots, L$, for all $i = 1, \ldots, M$. In addition, we have observations on s'_{it}, $t = t_{n+1} - j$, $j = 0, 1, \ldots, L$, and all i.

The first step in building the state transition is to build a multioutput feedforward neural net trained on this data to predict the forward state of dimension equal to that of s' from data on s and r. This multioutput net we denote by

$$\widehat{s'} = \Phi(s, x). \tag{11}$$

In addition, one may also observe the residuals

$$u_{it} = s'_{it} - \Phi(s_{it}, x_{it}) \tag{12}$$

Unlike a regression, the means of these residuals are not zero and we define the mean vector

$$m = \frac{1}{M(L+1)} \sum_{it} u_{it}, \tag{13}$$

along with the mean corrected predictions

$$\widehat{s}'' = \Phi(s, x) + m, \tag{14}$$

and zero mean residuals

$$\widetilde{u}_{it} = u_{it} - m. \tag{15}$$

The data on the zero mean residuals $\widetilde{u}_{it}$ may be employed to build the marginal residual distributions for the components of the vectors $\widetilde{u}$. We denote these distributions $G_k(u)$ for $k = 1, \ldots, K$, where K is the dimension of the state vector which is four in our example. One may go further to build joint distributions for $\widetilde{u}$ using a variety of copulas and their estimation. Here we employ the fact that the maximal entropy joint distribution given marginals is the product of the marginals or the joint distribution under independence and proceed under independence. We may then simulate from component distributions under independence to define a draw $\widetilde{u}' = \left(\widetilde{u}'_k, k = 1, \ldots, K\right)$ and then define

$$s' = \widehat{s}'' + \widetilde{u}'. \tag{16}$$

These steps define a data trained transition law $P(s'|s, x)$.

6 Sample State Transition

This section describes the sample state transition law trained on October 6, 2015 for data on the 589 stocks for 63 days starting July 9, 2015 for a total of 37,107 observations. The first step was the construction of a multioutput feedforward net for the prediction of the state the next day as function of the current state and the next return. There were then five inputs into the net which had four outputs. Two hidden layers with 50 and 25 variables

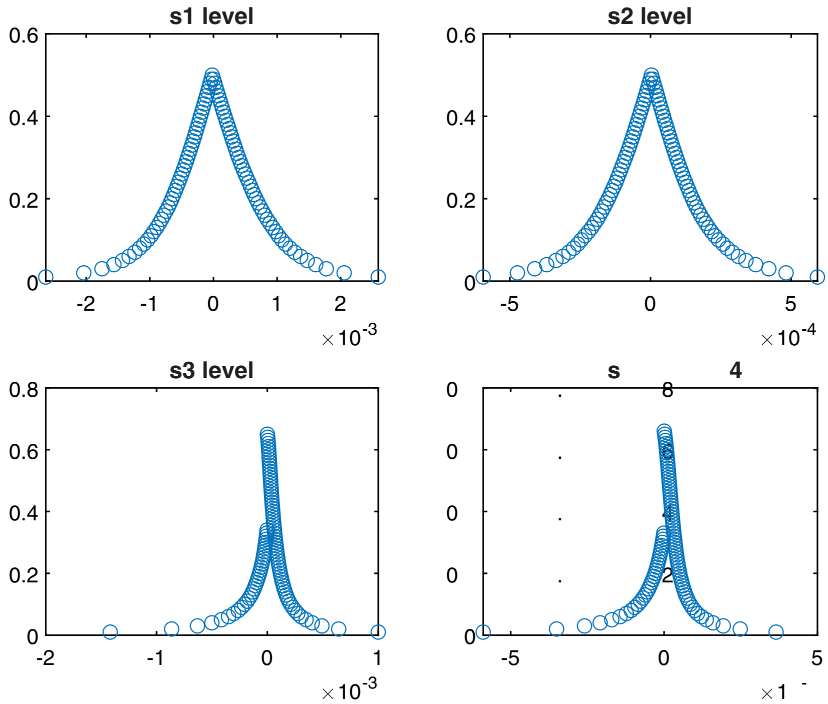

Figure 1 State Transition Residuals for training on October 6, 2015.

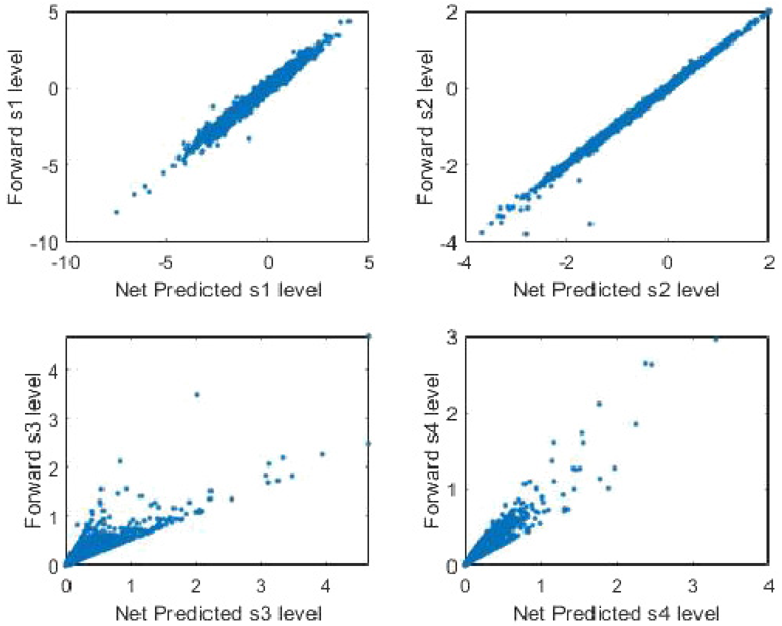

Figure 2 State predictions graphed against the forward states.

each were employed. The four mean residuals in basis points were −0.0633, −0.0409, −0.0826, and −0.0026. They are close to zero and were employed to form zero mean residuals. Figure 1 displays the four residual tail probabilities.

The residual correlation matrix C is as follows. The correlations are not large with differing signs. The joint law employed was that of independence.

$$C = \begin{pmatrix} 1 & -0.2249 & 0.1794 & -0.0564 \\ -0.2249 & 1 & -0.0362 & 0.2568 \\ 0.1794 & -0.0362 & 1 & -0.2855 \\ -0.0564 & 0.2568 & -0.2855 & 1 \end{pmatrix}$$

Figure 2 presents graphs of the four forward state levels plotted against their neural net predictions.

7 Policy Function Constructions

At each policy function training date the full data set of states experienced by all stocks in the recent past of 63 days for our example with 37,107 cases in all is reduced by quantization to the smaller set of 1,024 cases. This is done to reduce the number of optimizations involved. For each state four optimizations are involved as there are four value policy iterations being conducted for the policy convergence. The initial value function was an estimate of the expected return in the state assuming a single dollar investment in the stock. Given $V_n(s)$ and a distorted expectation operator $\mathcal{D}_x$ acting on the uncertainty of the next return x, the policy function is constructed on the quantized state space with entries s_m, $m = 1, \ldots, 1024$, to solve for

$$p_n(s_m) = \arg\max_p \mathcal{D}_x(p(e^x - 1) + V_n(s') - V_n(s_m)), \tag{17}$$

where s' is drawn from the state transition law $P(s' \,|\, s, x)$ as constructed in Section 5. On the quantized space the next value function is defined by

$$W_{n+1}(s_m) = \mathcal{D}_x(p_n(s_m)(e^x - 1) + V_n(s') - V_n(s_m)). \tag{18}$$

The value functions were rescaled and shifted to be -100 and 100 at the states s_0 and s_1 by defining

$$V_{n+1}(s) = -100 + \frac{W_{n+1}(s) - W_{n+1}(s_0)}{W_{n+1}(s_1) - W_{n+1}(s_0)} 200. \quad (19)$$

The implementation of the updates (17) and (18) requires a specification of the distribution of the uncertainty on x. The values for x were taken on a grid starting at -0.3 to 0.3 in steps of 0.005. The probabilities for these outcomes were defined by combining a bilateral gamma distribution fitted to data on past returns of just the S&P 500 index over the immediate past 252 days with a weight of 10%. In addition, a 90% weight was placed on the empirical distribution of residuals obtained on regressing the forward returns in the data set on the set of current states. These were just two reference return distributions for the distribution of x and are realistic in comparison with the use of a uniform probability on the grid. The final quantized policy function was taken as $p_4(s_m)$, $m = 1, \ldots, 1024$.

The value functions at each of the four iterations on a single training data and the final policy function was extended to the whole space by fitting a feedforward neural net to the outputs obtained on the quantized sets. The neural net value functions were needed to compute $V_n(s')$ in the two update Equations (17) and (18). The neural net policy functions are used in trading the policies into the future. The training on a single date took half an hour and the policy functions for 60 training dates were computed in parallel on an 18-core machine in a few hours.

8 Probability Distortions

Risk acceptability was defined in Artzner *et al.* (1999) as a convex cone containing the nonnegative random variables. Probability distortions deliver such convex cones via nonnegative distorted expectations when the distortion functions are concave distribution functions defined on the unit interval. Cherny and Madan (2009) introduced a one-parameter family of distortions $\Psi^{\gamma}(u)$ termed $minmaxvar$ for parameter γ defined by

$$\Psi^{\gamma}(u) = 1 - (1 - u^{\frac{1}{1+\gamma}})^{1+\gamma}. \quad (20)$$

The derivative of this distortion tends to infinity near zero and zero near unity, thereby reweighting large losses up toward infinity and large gains down to zero when a distorted expectation is seen as a reweighted expectation. The distortion is concave and the degree of concavity rises with the parameter γ.

Madan and Wang (2022a, 2022b) employ the two-parameter distortion termed $minmaxvar2$ with parameters γ_1, γ_2 defined by

$$\Psi^{\gamma_1,\gamma_2}(u) = 1 - (1 - u^{\frac{1}{1+\gamma_1}})^{1+\gamma_2}. \quad (21)$$

This distortion can be convex near zero for $\gamma_1 < 0$ and convex near unity for $\gamma_2 < 0$. It is convex when both γ_1 and γ_2 are negative. Madan and Wang (2022) show that distortions that are partially convex display preferences that are positive with respect to volatility and they deliver profitable trading strategies as compared to concave ones that lose money when followed. The lower convex distortions are termed LRT for loss risk taker, while the upper convex distortions are termed GRT for gain risk taker in Madan and Wang (2022).

In addition, we introduce here distortions that are both convex near zero and unity and below the identity while being concave and above the identity near the median. The intuition is based on recognizing that near the median we have considerable experience and trust the probability calculations while in the two tails the probabilities are questionable. Hence we take risk aversion near the median but are risk takers in both tails. Figure 3 displays such a distortion.

The specific formula for the distortion is

$$\begin{aligned}\Psi(u) = {} & \left(\frac{\alpha}{2} + \frac{\theta}{1+\theta}\frac{\alpha^{(1+\theta)/\theta}}{\beta^{1/\theta}} + \gamma\right)u \\ & + \frac{\beta}{(1+\theta)(2+\theta)}|u-0.5|^{2+\theta} \\ & - \alpha\frac{u^2}{2} - \frac{\beta}{(1+\theta)(2+\theta)2^{2+\theta}}, \end{aligned} \tag{22}$$

and Figure 3 displays the distortion for the stated parameter values. Policy functions are constructed for a variety of these distortions and the trading results are reported.

In addition to nondecreasing distortions, Madan and Wang (2022) introduced distortions bounded below by $-\lambda$ on defining

$$\widetilde{\Psi}(u) = (1+\lambda)\Psi(u) - \lambda u, \tag{23}$$

where Ψ is nondecreasing. A negative value for λ in Equation (23) yields distortions with a positive lower bound for the derivative.

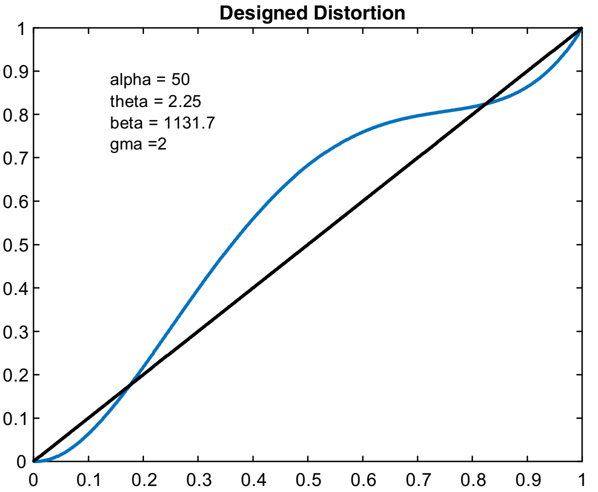

Figure 3 A distortion designed to be convex in the two tails.

9 Trading the Policy Functions

The policy functions constructed for a variety of distortions were employed to trade the 589 stocks from January 3, 2008 to the March 31, 2022. Each day and for each stock the new position taken in the stock is the recommendation resulting from applying the most recently trained policy function evaluated at the current state. In addition five percent of aggregated net legacy positions are reversed each day. A sample of the cumulated returns, defined by the daily change in the marked to market value of holdings plus net daily cash flows divided by the sum of the absolute value of all positions as at the end of the previous day, is displayed in Figures 4 and 5. The sample of eight distortions covers four cases for the use of different $minmaxvar2$ distortions, two distortions with two convex or concave regions near zero and unity and two distortions combine these with a straight expectation to yield distortions with a negative and positive lower bounds for the distortion derivative. We also present tables for the quartiles of three associated performance statistics. These are the Sharpe ratio, the Acceptability Index, and the Maximum Drawdown.

Sharpe Ratios for minmaxvar2 distortions

	0.1, 0.1	0.3, 0.3	0.15, −0.1	−0.075, 0.125
$Q1$	−0.2274	−1.8964	0.0113	0.0682
$Q2$	0.9942	−0.6273	1.2504	1.2908
$Q3$	2.0897	0.7973	2.3435	2.3528

Acceptability Index for minmaxvar2 distortions

	0.1, 0.1	0.3, 0.3	0.15, −0.1	−0.075, 0.125
$Q1$	−0.008	−0.0661	0.0005	0.0025
$Q2$	0.0361	−0.0228	0.0453	0.0466
$Q3$	0.0766	0.0288	0.0863	0.0864

Max Draw Down for minmaxvar2 distortions

	0.1, 0.1	0.33, 0.33	0.15, −0.1	−0.075, 0.125
$Q1$	0.0385	0.0430	0.0400	0.0400
$Q2$	0.0594	0.0640	0.0603	0.0603
$Q3$	0.1038	0.1072	0.1010	0.0991

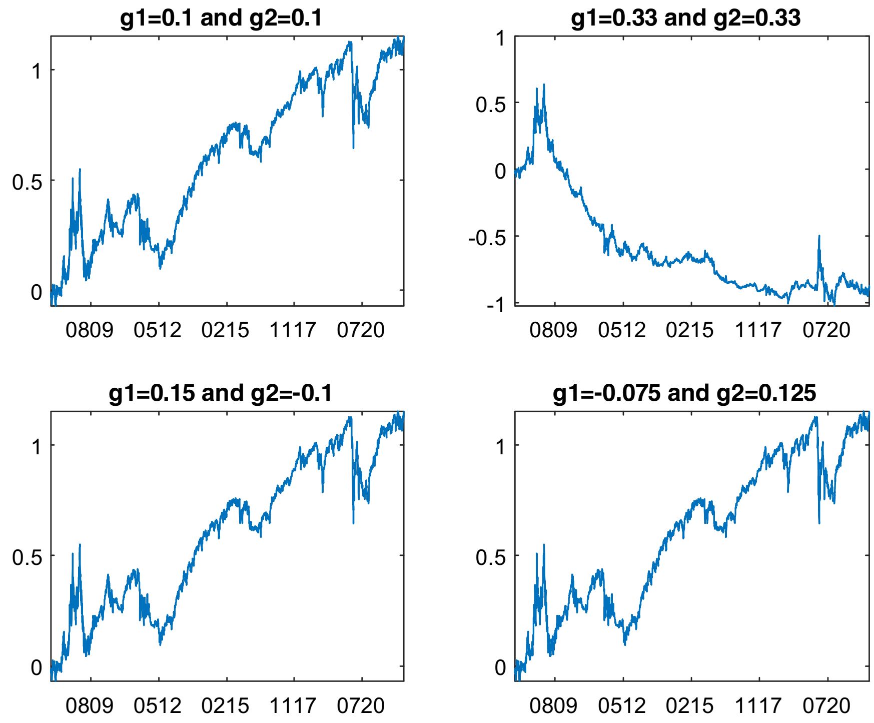

Figure 4 Cumulated returns from minvaxvar2 distortions.

Figure 5 presents cumulated returns for distortions that are convex or concave near zero or unity. This is followed by performance measure quartiles for these strategies.

Sharpe Ratios for distortions convex in both tails

	$Convex$	$Concave$	$NegativeLB$	$PositiveLB$
$Q1$	0.0706	−0.4153	−0.1054	0.0468
$Q2$	1.3269	0.9172	1.2954	1.2745
$Q3$	2.3608	2.1832	2.3733	2.3475

Acceptability Index for distortions convex in both tails

	$Convex$	$Concave$	$NegativeLB$	$PositiveLB$
$Q1$	0.0025	−0.0147	−0.0038	0.0017
$Q2$	0.0481	0.0330	0.0467	0.0465
$Q3$	0.0867	0.0803	0.0869	0.0864

Max Draw Down for distortions convex in both tails

	$Convex$	$Concave$	$NegativeLB$	$PositiveLB$
$Q1$	0.0388	0.0336	0.0379	0.0400
$Q2$	0.0591	0.0568	0.0560	0.0596
$Q3$	0.0895	0.0897	0.0878	0.0918

It may be observed that the different distortions deliver in some cases quite diverse trading experiences while a number of them are somewhat similar with comparable performance statistics. The highest median Sharpe ratio and Acceptability Index results from the distortion that is convex in both tails while the lowest median Max Drawdown is provided by the distortion convex in both tails with a negative lower bound for the reweighting probability.

10 Further Research Directions

The strategies presented were machine-trained solutions to the particular problem of deciding the investment level in a stock. The input variables used in the decision were a simple set of weighted averages of past returns and squared returns. One could instead employ other more sophisticated summaries of past returns. But, apart from the selection of explanatory or decision variables,

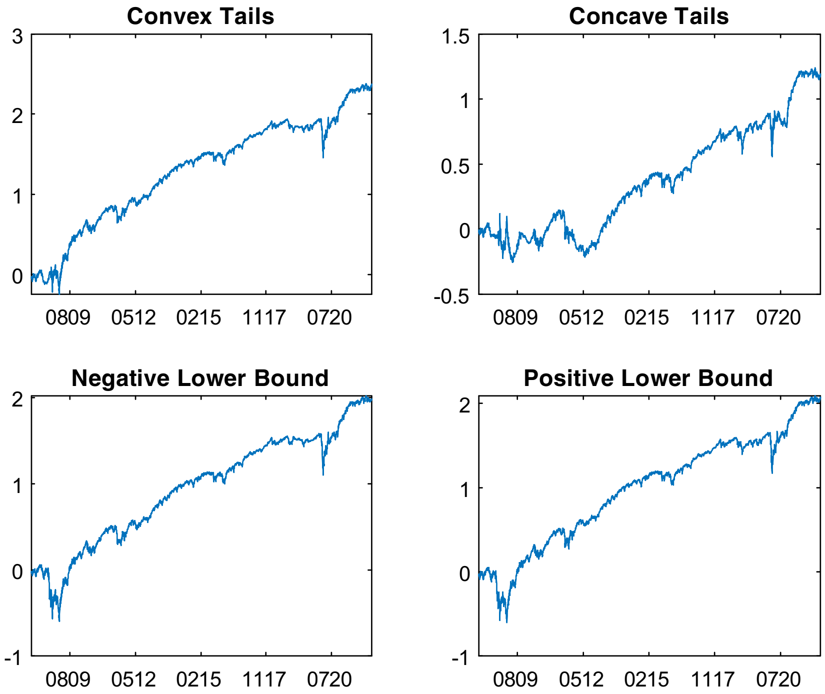

Figure 5 Cumulated returns using distortions that are convex or concave in both tails.

the problem being solved can be reformulated to better reflect what the strategy is to be designed to accomplish.

For example, if the interest is in market neutrality for instance, the positions taken could be financed by simultaneously taking a short position in a selected market index, like the S&P 500 index. The investment is then in a return differential and the transition law could be built around the return differential or the two returns taken separately. Still more generally the short or financing position could be taken in a set of indices. The policy is then multidimensional in its output but the design principles are the same.

A dynamic portfolio management context could be entertained where the investment sought is into a fixed set of asset classes with positions constrained to sum to unity with a possibly negative lower bound for the position in each asset class. The construction of relevant state variables for this context is a related research agenda.

Yet, another area of application that has seen some work is the design of dynamic hedging strategies for options on a single underlier. The state of the market is to be described by parameters that synthesize the option surface. The immediate return is that of the underlying asset. The state transition is to be learned by an appropriate multioutput neural net. The hedging state describes in addition the state of the option book described by, for example, a vector of vega buckets by strike and maturity. The hedging assets are preselected options by moneyness and maturity. The problem is to describe the hedging policy given the market and option book state, for a trained state transition.

11 Conclusion

A stylized problem in a stylized context of training a machine to trade a stock is addressed. The general structure of applying Reinforcement Learning methods to this context is modified for underlying financial conditions recognizing that rewards are related to uncertainties and actions while future states reflect the uncertainties but are divorced from our actions. In addition, objectives are taken to be risk sensitive using probability distortions and the sequence of future rewards is not discounted. The absence of discounting reflects the principles of Now Decision Theory set out in Madan *et al.* (2023).

The strategies are illustrated by trading policy functions applied to state variables for 589 stocks over the period January 3, 2008 to March 31, 2022. Feedforward neural nets are used to summarize both the policy functions and the underlying state valuation functions trained on a quantized subset of visited points. In addition, multioutput feedforward neural nets are employed to summarize state transitions.

A variety of probability distortions are used in the objectives including concave distortions, S-shaped distortions, distortions that are convex both near zero and unity, or concave at these extremes, along with examples permitting negative distortion derivatives. The best trading performance for Sharpe ratios and Acceptability Indices was delivered by the distortions convex in the two tails. From the perspective of lowering drawdowns the use of distortions with a negative derivative in some regions was useful.

Future research in this direction is discussed from the perspectives of beating benchmarks, managing portfolios dynamically, or hedging option books continuously using option hedges.

References

Alexander, S. S. (1961). “Price Movements in Speculative Markets: Trends or Random Walks,” *Industrial Management Review* **2**, 7–26.

Artzner, P., Delbaen, F., Eber, J. M., and Heath, D. (1999). “Coherent Measures of Risk,” *Mathematical Finance* **9**, 203–228.

Back, K. (1991). “Asset Pricing for General Processes,” *Journal of Mathematical Economics* **20**, 371–395.

Brogaard, J. and Zareei, A. (2023). “Machine Learning and the Stock Market,” *Journal of Financial and Quantitative Analysis* **58**, 1431–1472.

Chavarnakul, T. and Enke, D. (2008). “Intelligent Technical Analysis Based Equivolume Charting for Stock Trading Using Neural Networks,” *Expert Systems with Applications* **34**, 1004–1017.

Cherny, A. and Madan, D. B. (2009). “New Measures for Performance Evaluation,” *Review of Financial Studies* **22**, 2571–2606.

Cootner, P. H. (1962). “Stock Prices: Random versus Systematic Changes,” *Industrial Management* **3**, 24–45.

Delbaen, F. and Schachermayer, W. (1994). “A General Version of the Fundamental Theorem of Asset Pricing,” *Mathematische Annalen* **300**, 463–520.

Fama, E. F. (1965a). “The Behavior of Stock-Market Prices,” *The Journal of Business* **38**, 34–105.

Fama, E. F. (1965b). “Random Walks in Stock Market Prices,” *Financial Analysts Journal* **51**, 75–80.

Fama, E. F. (1970). “Efficient Capital Markets: A Review of Theory and Empirical Work,” *The Journal of Finance* **25**, 383–417.

Gao, T., Chao, Y., and Liu, Y. (2017). “Applying Long-Short Memory Neural Networks for Predicting Stock Closing Price,” in *8th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (Beijing, China), pp. 575–578.

Harrison, J. and Kreps, D. (1979). “Martingales and Arbitrage in Multiperiod Securities Markets,” *Journal of Economic Theory* **20**, 381–408.

Harrison, J. M. and Pliska, S. R. (1981). “Martingales and Stochastic Integrals in the Theory of Continuous Trading,” *Stochastic Processes and Their Applications* **11**, 215–260.

Harrison, J. M. and Pliska, S. R. (1983), “A Stochastic Calculus Model of Continuous Trading: Complete Markets,” *Stochastic Processes and Their Applications* **15**, 313–316.

Jacod, J. and Shiryaev, A. (1998), “Local Martingales and the Fundamental Asset Pricing Theorems in the Discrete-Time Case,” *Finance and Stochastics* **2**, 259–273.

Kendall, M. G. (1953), “The Analysis of Economic Time Series,” *Journal of the Royal Statistical Society* **96**, 11–25.

Khintchine, A. Ya. (1938), “Limit Laws of Sums of Independent Random Variables,” ONTI, Moscow, (Russian).

Khoa, B. T. and Huynh, T. T. (2021). “Is It Possible to Earn Abnormal Return in an Inefficient Market?

An Approach Based on Machine Learning in Stock Trading," *Computational Intelligence and Neuroscience*, https://doi.org/10.1155/2021/2917577.

Küchler, U. and Tappe, S. (2008). "Bilateral Gamma Distributions and Processes in Financial Mathematics," *Stochastic Processes and their Applications* **118**, 261–283.

Kusuoka, S. (2001), "On Law-Invariant Coherent Risk Measures," *Advances in Mathematical Economics* **3**, 83–95.

Lévy, P. (1937). *Théorie de l'Addition des Variables Aléatoires* (Gauthier-Villars, Paris).

Madan, D. B., Schoutens, W., and Wang, K. (2022), "Risk Conscious Investment," SSRN paper no. 4197305.

Madan, D. B., Schoutens, W. and Wang, K. (2023). "Now Decision Theory," *Probability, Uncertainty, and Quantitative Risk* **8**, 391–416.

Madan, D. B. and Schoutens, W. (2016). *Applied Conic Finance* (Cambridge University Press, Cambridge, UK).

Madan, D. B. and Schoutens, W. (2020). *Nonlinear Valuation and Non-Gaussian Risks in Finance* (Cambridge University Press, Cambridge, UK).

Madan, D. B. and Wang, K. (2022a). "Attractive Investment Opportunities: The Irrationality of being Rational," SSRN Paper No. 4330749.

Madan, D. B. and Wang, K. (2022b). "Stationary Increments Reverting to a Tempered Fractional Lévy Process (TFLP)," *Quantitative Finance* **22**, 1391–1404.

Madan, D. B. and Wang, K. (2024), "Investor Determined Dividend Policies," SSRN paper no. 4821692.

Moore, A. (1962). "A Statistical Analysis of Common-Stock Prices," Unpublished PhD. Dissertation, Graduate School of Business, University of Chicago.

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions* (Cambridge Uinversity Press, Cambridge).

Shepard, N. (2005). *Stochastic Volatility* (Oxford University Press, Oxford, UK).

Shleifer, A. (2000). *Inefficient Markets: An Introduction to Behavioral Finance* (Oxford University Press, Oxford, UK).

Sutton R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).

Wu, X., Chen, H., Wang, J., Troiano, L., Loia, V., and Fujita, H. (2020). "Adaptive Stock Trading Strategies with Deep Reinforcement Learning Methods," *Information Sciences* **538**, 142–158.

Zhang, J., Li, L., and Chen, W. (2021). "Predicting Stock Price Using Two-Stage Machine Learning Techniques," *Computational Economics* **57**, 1237–1261.

Keywords: Acceptable risks; distorted expectations; tempered fractional process; now decision theory