Inferring Traders' Price Expectations from Time Series Data with POMDPs

Aini Putkonen[1] **(aini.putkonen@aalto.fi)**
Sandra Andraszewicz[2] (sandraszewicz@ethz.ch)
Christoph Hölscher[2] (choelsch@ethz.ch)

[1] Department of Information and Communications Engineering, Aalto University, Espoo, Finland
[2] Chair of Cognitive Science, ETH Zurich, Zürich, Switzerland

Abstract

Price expectations are an important driver of traders' buy, hold, and sell decisions. They are often estimated by surveying investors; however, these verbal accounts may differ from latent expectations. In this paper, we propose how to infer traders' price expectations from trading data instead. We assume traders' goal is to maximize earnings at the end of the trading period by sequentially buying, holding, or selling shares. Due to sequentiality, trading is represented as a Partially Observable Markov Decision Process (POMDP) solved with Deep Reinforcement Learning (DRL). This trading model follows an approximately optimal trading policy with respect to the price paths used in training. Meanwhile, we assume that traders choose optimal trading actions given their price expectations. Thus, price paths used in training are assumed to approximate expectations, and we describe them with separate segments, each characterized by two parameters (trend and volatility). We then infer which values of these parameters produce a trading policy close to that of actual traders. While the presented approach achieves a good model fit when applied to the worst-performing traders in an empirical study, our results are more ambiguous in the case of top traders, suggesting their trading strategy is directed by more elaborate mechanisms.

Keywords: trading; expectations; POMDP

Introduction

Survey-based methods are commonly used in understanding traders' price expectations. These surveys include questions like estimating the percentage increase of the Dow Jones index in the coming year or the probability of a catastrophic crash (Shiller, 2000). Another common way to query expectations involves asking a trader whether they believe the market to be 'bearish', 'bullish', or neutral within a time window (Fisher & Statman, 2000; Clarke & Statman, 1998). Examples of surveys that collect investor sentiments employing the outlined questions include, for instance, the American Association of Individual Investors Sentiment Survey and Yale School of Management Stock Market Confidence Indices. While there is some evidence that stated expectations follow the non-observable true expectations (Greenwood & Shleifer, 2014), concerns have been expressed about the reliability of survey data due to factors like linguistic interpretation and noise (Cochrane, 2011). Similar observations are also prevalent in other domains, for instance, in comparing stated and revealed preferences (Frey, Pedroni, Mata, Rieskamp, & Hertwig, 2017). Additionally, it may be infeasible to survey investors ahead of trading to estimate their expectations. Thus, developing techniques to understand expectations independent of self-reported measures is paramount.

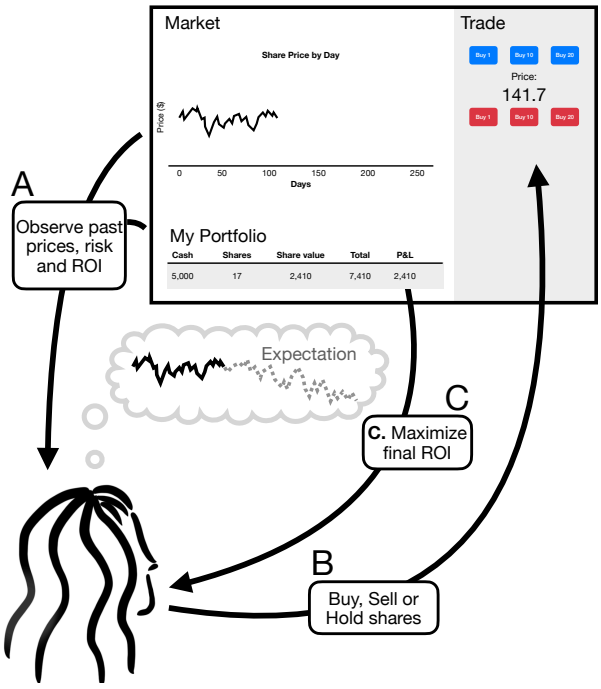

Figure 1: The trading task. The trader observes a price path until time t and information about portfolio composition at time t (A). They use this information to decide whether to Buy, Hold, or Sell shares (B) in an optimal sequence to maximize return on investment (ROI) at the end of the trading period (C). System illustration drawn based on the Zurich Trading Simulator (Andraszewicz, Friedman, et al., 2023).

In this work, we propose a parameter inference-based approach to understand traders' price expectations via a process that mimics belief formation through sense-making. We first build a computational model of trading and train this model using different price paths representing expectations. Consider a trading task (Figure 1) of choosing a preferred portfolio composition from cash and shares of one stock for a trading period. A trader observes a price path and information about the current portfolio composition, which they use to make a decision to buy, sell, or hold shares. We assume that the trader's goal is to perform as well as possible at the end of the trading period. Thus, underlying our approach is

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

the assumption that a trader's strategy is optimal (i.e., maximizes Return on Investment, ROI) with respect to what they expect to happen. For instance, if the trader expects the stock price to go up eternally, it makes sense to buy shares. On the other hand, if the trader holds an expectation that stock prices will decline, it is optimal to sell. Therefore, the goal is to understand under what price path traders' observed behavior is optimal, as it could be indicative of their expectations.

Trading is an exemplary task to achieve this goal as we can make some fairly realistic assumptions about trading behavior. As verbalized in the traditional portfolio theory (Markowitz, 1952), consider that portfolio selection consists of two stages: 1) observation and forming beliefs about future prices and 2) making portfolio selection based on this information. Unlike in portfolio theory, we assume that traders primarily try to maximize discounted ROI at the end of a trading period. Through these assumptions, we model trading as a sequential problem using the notion of a Partially Observable Markov Decision Process (POMDP) solved with Deep Reinforcement Learning (DRL), specifically Proximal Policy Optimization (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017). At each timestep, the agent observes past prices and information about performance, which is used to choose a desirable portfolio composition consisting of cash and shares. These decisions impact ROI at the end of the trading period, which is used as the reward.

What about beliefs? This approach allows us to represent expectations without an explicit belief model. As we use DRL, the agent learns an optimal trading policy with respect to the price paths it is exposed to during *training*. Then, when making *predictions*, we use these price paths to approximate expectations. Specifically, Geometric Brownian Motion (GBM) with two parameters (trend and volatility) is used to generate price paths and thus to approximate price expectations, following earlier work in finance (Hull, 2017). We use a modified GBM process, where we assume expectations to be formed of multiple GBM segments described by separate trend and volatility parameters.

The key components of the proposed parameter inference approach are therefore: 1) a model that exhibits optimal trading behavior, and 2) a price model that is described by some parameters that capture a wide range of different price patterns (e.g., random walk with a drift). In our case, we use GBM with prescribed trend and volatility (μ and σ, respectively). First, we consider parameters $\mu_i, \sigma_i \in \boldsymbol{\theta_i}$ and generate GBM price paths, consisting of separate segments, stored in the state of the POMDP (Figure 2). Solving this POMDP using DRL results in a policy that can be compared to observational data. The goal is to find parameter values $\boldsymbol{\theta}^*$ that lead to a policy π^* closely resembling the behavior of actual traders. The inferred trend and volatility parameters are then used as a proxy for price expectations.

We use data collected with the Zurich Trading Simulator (ZTS) to demonstrate our approach (Andraszewicz, Kaszás, Zeisberger, & Hölscher, 2023). The dataset contains trades

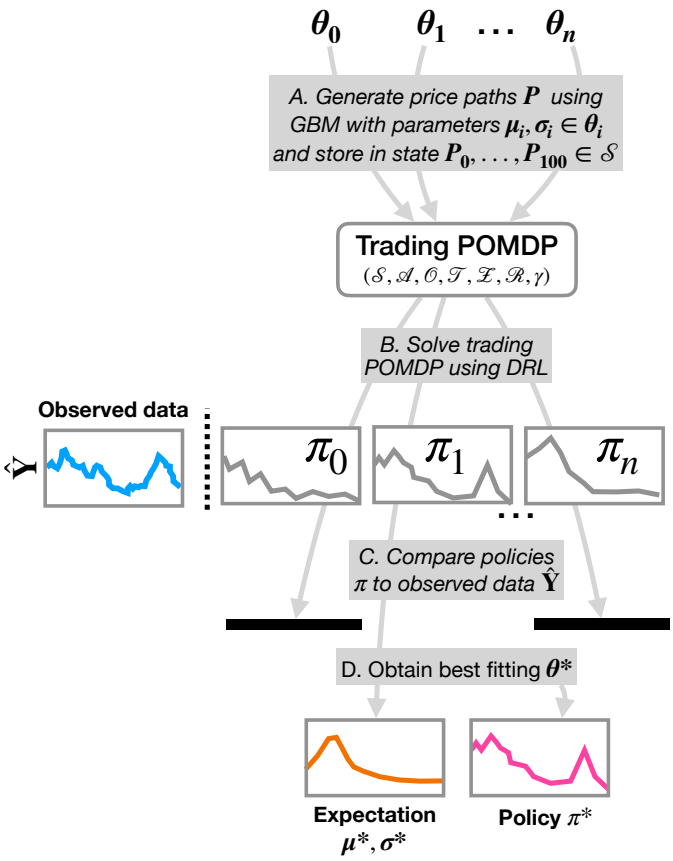

Figure 2: Price expectations via parameter inference, the goal of which is to find parameters $\boldsymbol{\mu_i}, \boldsymbol{\sigma_i} \in \boldsymbol{\theta^*}$ of the price model that produce a policy π^* that predicts actions closely resembling observations $\hat{\boldsymbol{Y}}$. Parameters $\boldsymbol{\mu^*}, \boldsymbol{\sigma^*}$ then describe price expectations. Additionally, we fit the discount factor γ.

from approximately 800 experienced traders in a simulated environment corresponding to the task of buying, selling, or holding shares of one stock (Figure 1). We demonstrate our approach by inferring expectations of the top and bottom decile of traders as measured by performance. Our results reflect a primarily upward-trending expectation held by the worst decile of traders, who display various trading strategies that, on average, start with a period of buying shares. In contrast, the behavior of top traders is more difficult to capture with our approach, reflecting potential trading strategies omitted in the trading model, such as trend-following.

The main contributions of this paper are as follows:

- We present an approximately optimal model that simulates traders behaving under given price expectations.
- This trader model is then used to infer parameters of traders' price expectations in empirical data.

We conclude by discussing the implications of the proposed approach for studying expectations and modeling trading.

Related work

Price expectations

It is often assumed that those participating in the stock market have the same information on which they form expectations about price movements, yet this information is often interpreted differently, resulting in divergent expectations (Davis, Pagano, & Schwartz, 2009). Existing studies eliciting these expectations via stock market forecasts largely focus on stock allocations (Fisher & Statman, 2000) or verbal accounts that state price expectations, such as newsletters (Fisher & Statman, 2000; Clarke & Statman, 1998), survey data (Fisher & Statman, 2000), or controlled studies (Haruvy, Lahav, & Noussair, 2007). Key observations from these studies include that past prices impact expectations of price movements and that investor expectations are somewhat indicative of future returns (Haruvy et al., 2007). Unlike with risk preferences (Frey et al., 2017), fewer studies focus on eliciting stock market expectations from behavioral data.

POMDPs in computational cognitive modeling

Partially Observable Markov Decision Process (POMDP) is a flexible framework for modeling sequential decision-making, originally stemming from engineering (Åström, Karl Johan, 1965). POMDP is an extension of a Markov Decision Process (MDP) which models interaction between an agent and an environment through the notions of state, action, transition, and reward. In contrast to MDPs, POMDPs assume the state is not directly observable, but the agent has to use sensory inputs via observations. Solving a POMDP then refers to finding an optimal policy, a mapping between states and actions, that maximizes cumulative rewards. Several approaches have been used, while recently Deep Reinforcement Learning (DRL) has become popular. Approaches based on POMDPs have been introduced to cognitive science (Littman, 2009), specifically in tasks including visual search (J. P. Jokinen, Wang, Sarcar, Oulasvirta, & Ren, 2020), planning (Callaway et al., 2022), decision-making (Chen, Chang, & Howes, 2021), and multi-tasking (J. P. P. Jokinen, Kujala, & Oulasvirta, 2021).

Reinforcement learning in trading

Most work using POMDPs solved with Reinforcement Learning (RL) in the context of trading focuses on quantitative trading, referring to the automated identification of investment opportunities. RL models have achieved good performance in tasks like algorithmic trading and portfolio management (Sun, Wang, & An, 2023). Often, these models are trained with historical prices. A typical goal for using RL in trading has been automation, rather than modeling human-like trading behavior, for instance by building agents that outperform indices or a traditional mean-variance approach to stock trading (Yang, Liu, Zhong, & Walid, 2021).

Method

Trading task

The trading task we consider is based on previous empirical work conducted using ZTS (Andraszewicz, Kaszás, et al., 2023). In this task, the trader chooses an optimal portfolio composition, where the assets are cash and shares of one stock. The trader can buy, sell, or hold shares using quick trade buttons at every timestep t until the end of the trading period T. The trader is displayed information about their current portfolio composition and performance (Figure 1). Traders' actions do not impact share prices. The traders, who have risk-neutral preferences, try to maximize earnings at the end of the trading period. The only information about the market the traders have access to is historical share prices.

Trading POMDP

We formulate trading as a POMDP, that is, a tuple $(\mathcal{S}, \mathcal{A}, O, \mathcal{T}, \mathcal{Z}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, O is the observation space, $\mathcal{R}$ is rewards, and $\mathcal{T}$ and $\mathcal{Z}$ are transitions, defined below. The agent's goal is to follow a policy π maximizing expected reward R for a discount factor γ until horizon T:

$$\arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R_t\right] \tag{1}$$

The discount factor $\gamma \in \boldsymbol{\theta}$ is considered a parameter of the trader model.

State The state $s_t \in \mathcal{S}$ is inaccessible to the agent. The current number of shares (n), cash (b), and ROI are stored in the state. Also, the state includes price paths. These paths follow Geometric Brownian Motion (GBM), as is typical in finance (Hull, 2017). We assume that a price path $\boldsymbol{P}$ consists of three segments with independent parameters, simulated as

$$p_t = p_0 \exp\left(\left(\mu_k - \frac{\sigma_k^2}{2}\right)t + \sigma W_t\right) \tag{2}$$

where μ_k is a trend (also known as drift) and σ_k a volatility parameter, $k \in [0,1,2]$ is the segment index, and $W_t \sim \mathcal{N}(0,1)$. For the first segment, we assume $p_0 = 141.7$ (price at $t = 0$ in the empirical price path). The trading day where an 'inflection' occurs is randomly sampled from $d_i \in [i \cdot \lfloor \frac{T}{2} \rfloor, (i+1) \cdot \lfloor \frac{T}{2} \rfloor]$, where $i \in [0,1]$. The state S consists of 100 price paths altogether. Each path is sampled from GBM given parameters μ_k and σ_k, where $k \in [0,1,2]$. For each episode, a price path $\boldsymbol{P} \in \mathcal{S}$ is randomly sampled. This price model is used to approximate expectations. The trend and volatility $\sigma_k, \mu_k \in \boldsymbol{\theta}$ are treated as the parameters of the price model. Thus, $s_t = [n_t, b_t, \mathrm{ROI}_t, \boldsymbol{P_0}, ..., \boldsymbol{P}_{100}]$.

Reward At the end of each trading round, the agent is rewarded with its return on investment (ROI). In training, we use nominal returns so each reward is calculated as $R_T = \mathrm{ROI} \times b_0$, where b_0 is the initial budget of the agent.

Action At each timestep t, the agent can either hold, buy, or sell shares, constituting actions $a_t \in \mathcal{A}$. The actions available to the agent at any time t are $a_t \in [0, 1, 10, 20, -1, -10, -20]$, with negative values corresponding to selling, positive values to buying, and zero to holding shares.

Observation The agent's observation consists of risk (percentage of shares in the portfolio, r), observed share prices (p), ROI, and executed actions (a) until timestep t. That is, $o_t = [r_0, ..., r_t, p_0, ..., p_t, \text{ROI}_0, ..., \text{ROI}_t, a_0, ..., a_t]$.

Transitions After choosing an action a_t, the state changes according to the distribution $s_t \sim \mathcal{T}(s_{t+1}|s_t, a_t)$, whereby the number of shares n_t, cash b_t, and ROI are updated. Note that the trading environment is limited such that the maximum number of shares that can be sold at time t is b_{t-1}, and the maximum number that can be bought is cash divided by share price $\frac{c_{t-1}}{p_t}$. When $t > T$, a new price path is sampled. Observations are updated according to $o_t \sim \mathcal{Z}(o_{t+1}|o_t, a_t)$, where r_t, p_t, ROI_t, and a_t are revised.

Implementation

Python 3.12 was used in implementing the trading model. Proximal Policy Optimization (PPO) from the Stable Baselines 3 library is used to solve the trading POMDP (Hill et al., 2018). The custom trading environment was implemented using the Gymnasium API (Towers et al., 2024).

Parameter inference

Traders try to take an optimal sequence of buy, hold, and sell actions given what they expect to happen in the stock market. Similarly, the trained trading model (POMDP solved with DRL) chooses actions according to an optimal policy $\pi(a_t|o_t)$. Recall that observation o_t depends on the state space $\mathcal{S}$ storing a set of price paths described by parameters $\mu_k, \sigma_k \in \boldsymbol{\theta}$ where $k \in [0, 1, 2]$. Given we solve the POMDP using DRL, its behavior is approximately optimal with respect to O. Therefore, we approximate expectations implicitly by representing them in the state. The problem then is to find parameters $\boldsymbol{\theta}$ that lead to an agent that produces predictions closest to the empirical data. We use empirical time series data from a previous study conducted with ZTS (Andraszewicz, Kaszás, et al., 2023), where trades are collected in one round using a price path with a flat trend, based on scaled historical prices in the Swiss Market Index.

That is, assume there is this ground truth data $\hat{\boldsymbol{Y}}$ of buy, sell, and hold actions in one price path. We can use a trained trader model to obtain predictions $\boldsymbol{Y}$ with policy π of buy, sell, and hold actions in the same price path the ground truth data was collected in. The task is to find the optimal parameters $\boldsymbol{\theta}^*$ that minimize the discrepancy d between the empirical data $\hat{\boldsymbol{Y}}$ and predicted trade actions $\boldsymbol{Y}$, where d is the mean squared error (MSE) between a moving average of the observed and predicted number of shares the agent or traders hold (window of 10 trading days). MSE was used as it is common in forecasting tasks (Shcherbakov et al., 2013). Due to the intractability of the model, we use Bayesian optimization to infer parameters; specifically, the Agent Forge (Junior & Oulasvirta, 2025) framework. We fix the PPO hyperparameters based on an initial round of model fitting. The optimization (300 iterations per model fitting) was run on two NVIDIA GeForce RTX 4090 GPUs.

Results

We report on results of fitting expectations to trading data collected with ZTS. Initial analysis with aggregate data suggested that differently performing traders have distinct strategies. Thus, in what follows, we compare results of the top and bottom decile (by ROI) of the traders.

Worst-performing traders

The model fitting results (MSE=7) suggest that the 1st decile of traders hold an expectation of a broadly upward-downward facing trend, in contrast to the flat trend that the true price path follows (Figure 3, pane A, and Table 1). The first and second segments have an upward trend ($\mu_0 = 0.279$ and $\mu_1 = 0.290$), and the third segment has a downward trend ($\mu_2 = -0.234$). The volatility varies across the segments ($\sigma_0 = 0.0078$, $\sigma_1 = 0.021$, and $\sigma_2 = 0.175$). Upon visual inspection, the performance (Figure 3, pane B) and the number of shares the trained policy and the traders hold on each timestep are similar (Figure 3, pane C). The mean number of shares held by traders is slightly higher than the trained policy predicts (29.19 vs. 28.08), while the median number of shares is higher for traders and the maximum number of shares is lower (35 vs. 24 and 73 vs. 74, respectively), see Table 2. The correlation between the ground truth number of shares and predictions is strong ($R = 0.91$, see Figure 5).

Top-performing traders

The model fitting results are worse for the top decile of traders. The best-fitting trading behavior from the top traders (MSE=225) emerges when the model is trained with price paths that have a combination of a broadly flat and downward trend (Figure 4, pane A, and Table 1). The first segment follows a flat upward trend ($\mu_0 = 0.052$), followed by a relatively steeper downward trend in the second segment ($\mu_1 = -0.226$). The final segment also follows a downward trend ($\mu_2 = -0.061$). The first segment has the smallest volatility ($\sigma_0 = 0.028$), followed by larger volatilities ($\sigma_1 = 0.049$ and $\sigma_2 = 0.035$). A clear difference is visible in the median and maximum number of shares held by humans and the model

	μ_0	σ_0	μ_1	σ_1	μ_2	σ_2
Worst	0.279	0.0078	0.290	0.021	-0.234	0.175
Best	0.052	0.028	-0.226	0.049	-0.061	0.035

Table 1: Trend (μ) and volatility (σ) parameters fitted to bottom (Worst) and top (Best) deciles of traders for the three price path segments. Red highlight refers to a bullish and blue to a bearish trend expectation.

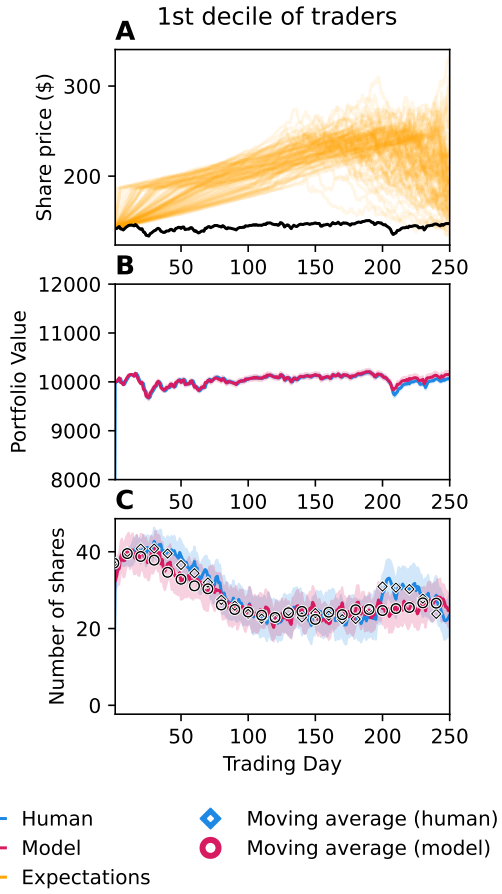

Figure 3: Empirical data vs. predictions (100 episodes) for the 1st decile of traders. The number of shares held by the traders is similar to model prediction (pane C), while the approximated expectation differs from the true prices (pane A).

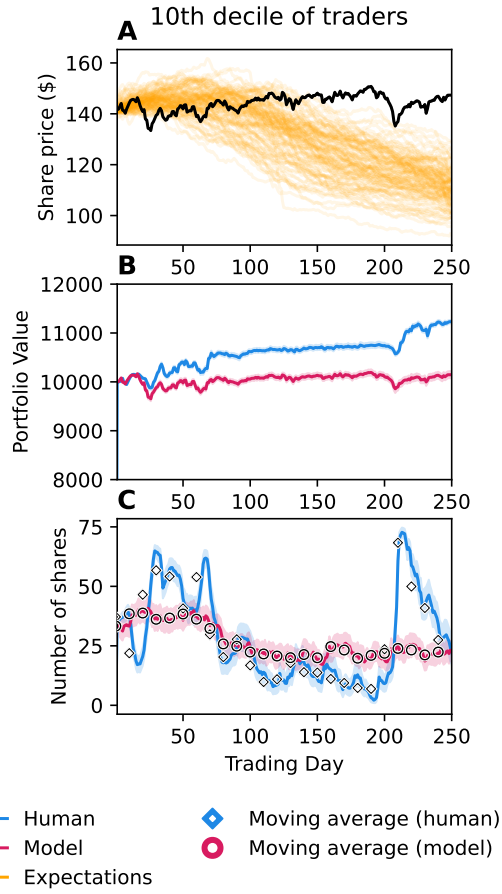

Figure 4: Empirical data vs. predictions (100 episodes) for the 10th decile of traders. The traders buy and sell shares in larger quantities on average than the model predicts (pane C), indicating a poorer model fit than for the worst traders.

(20 vs. 21 and 82 vs. 78, respectively), and in the standard deviation (29.40 vs. 23.30). The correlation between the ground truth and predicted number of shares is only moderate ($R = 0.55$), and visual inspection reveals that the model fails to capture the top range of held shares, especially (Figure 6). Inspection of individual traders' behavior suggests a considerable amount of noise in the case of the worst traders, while the top traders follow similar strategies.

		Mean	Median	Min	Max	SD
Worst	**H**	29.19	35	0	73	23.70
	M	28.08	24	0	74	23.04
Best	**H**	28.77	20	0	82	29.40
	M	26.76	21	0	78	23.30

Table 2: Number of shares (n) held by traders (H) and the model (M) in bottom (Worst) and top (Best) deciles.

Discussion

This work contributes an approach to understanding traders' price expectations using trading data. We implement this by representing the trading task as a POMDP solved with Deep Reinforcement Learning, wherein the state contains price path segments described by trend and volatility. While prior studies have examined traders' price expectations and used POMDPs for trading models, we introduce a novel approach by representing expectations within the state of a POMDP.

An advantage of the presented approach is that it allows investigating price expectations without verbal accounts or an explicit belief model. Verbal accounts, prone to errors due to factors like linguistic interpretation, are expensive to collect. Developing methods similar to the one proposed allows automating analyses of expectations, which could help mitigate potential biases associated with self-selected survey responses. Also, a benefit of modeling expectations through a model like Geometric Brownian Motion (GBM), instead of the commonly used categories (bearish, bullish, neutral), is

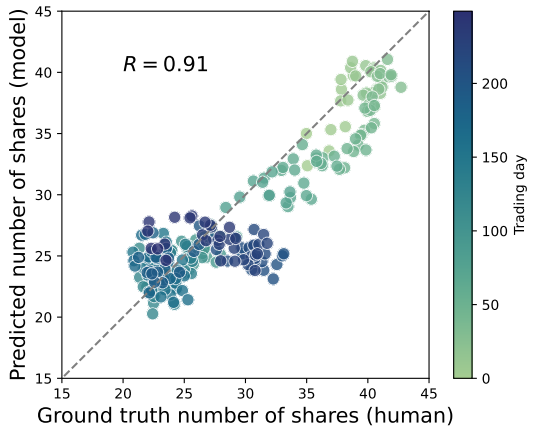

Figure 5: Correlation between ground truth and predicted number of shares held at each trading day for the worst performing traders in the dataset. The correlation is strong pointing to a better model fit for the worst-performing traders.

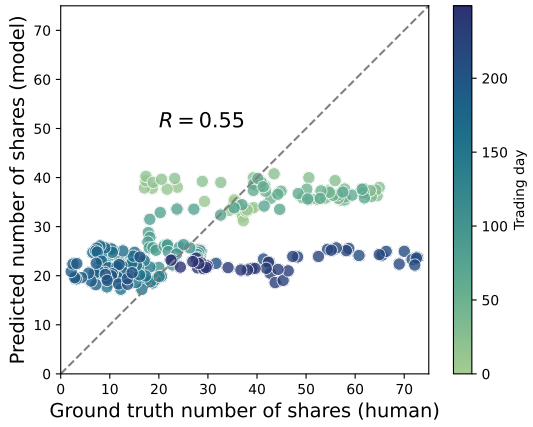

Figure 6: Correlation between ground truth and predicted number of shares held at each trading day for the best performing traders in the dataset. The model does not capture particularly the top-end of held shares.

expressing more complicated expectations than trend alone. This could be significant when trading at high frequency.

Our results suggest that traders who perform the worst have upward-downward trending price expectations. The observation that the worst traders enter the experiment with positive expectations aligns with previous empirical studies (Caginalp, Porter, & Smith, 2000). A caveat of this analysis is the poor model fit in the top traders, pointing to a limitation of the trading model. Notably, the worst traders adopt a broad range of trading strategies in the empirical data, in contrast to the best traders who seem to be potential trend-followers. It is also possible that all traders hold similar beliefs, and by chance, the best traders encounter superior strategies. That said, the similarity of strategies that the best traders adopt suggests that their performance is not purely by chance. Therefore, capturing the systematic strategies of the top traders would require extending the model to consider further factors, like the aforementioned trend-following.

The notion of expectations is related to mental models, that is, small-scale models of the world (Craik, 1967). In this work, this model is approximated as GBM with two parameters: trend and volatility. This broadly corresponds to a belief that the stock market moves according to a random walk with a drift. As we use data from an experiment with a single asset and a controlled trading environment, making this assumption of the traders' psychology is plausible: in the absence of other information, expectations of price movements could drive behavior, in conjunction with discounting of future rewards. Indeed, while the worst traders' performance seems to potentially reflect expectations described by trend and volatility, this might not hold for the top performers. Traders' behavior may also be driven by psychological factors like willingness to take risks and attention, which our model currently does not account for. For instance, the larger quantities of traded shares and trend-following in the best performers could speak to the relevance of these attributes in their trading behavior. Additionally, in reality, traders have access to more information than just past prices. Therefore, it is plausible that traders hold mental models of price movements, other than the employed GBM, that incorporate this information. Future work should investigate different models than GBM to describe price expectations.

Our results also point to a need for additional validations; most pertinently, potential issues with identifiability need examination through parameter recovery studies (i.e., to ensure that different trend and volatility parameters lead to distinguishable policies). To this end, different discrepancy metrics, in addition to MSE, should be considered. Specifically, we observe that different expectation patterns (e.g., downward vs. upward) may lead to similar MSE values and thus predicted policies. Similarly, the trained policy displays some noise, which means the stability of the results should be investigated, alongside the impact of the PPO hyperparameters.

Notwithstanding these limitations, the current work could be used to inform our understanding of stock markets more broadly. As expectations influence trader behavior and impact markets, generating testable predictions of them may provide valuable insights. The presented approach could be used for this purpose. Future work could aim to validate such predictions in trading simulators by eliciting expectations, for instance, through exposing traders to specific price paths. The strength of the expectations in different groups could then be assessed by measuring trading performance across various trials: consistent behavior regardless of the shown price path could indicate the traders' actions are driven by the elicited expectations. These predictions could also be tested in more realistic settings by observing the consistency of traders' actions in relation to expectations inferred with our approach.

Acknowledgments

This work was supported by the Research Council of Finland flagship program FCAI (grants 328400, 345604, 341763) and the project Subjective Functions (grant 357578). Department of Information and Communications Engineering at Aalto University School of Electrical Engineering and the Chair of Cognitive Science at ETH Zurich also supported this work. We thank our colleagues for their comments, in particular, Antti Oulasvirta for feedback on the manuscript.

References

Andraszewicz, S., Friedman, J., Kaszás, D., & Hölscher, C. (2023). Zurich trading simulator (zts) — a dynamic trading experimental tool for otree. *Journal of Behavioral and Experimental Finance*, *37*, 100762. doi: https://doi.org/10.1016/j.jbef.2022.100762

Andraszewicz, S., Kaszás, D., Zeisberger, S., & Hölscher, C. (2023). The influence of upward social comparison on retail trading behaviour. *Scientific Reports*, *13*(1), 22713.

Caginalp, G., Porter, D., & Smith, V. L. (2000). Overreaction, momentum, liquidity, and price bubbles in laboratory and field asset markets. *Journal of Psychology and Financial Markets*(1), 24–48.

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, *6*(8), 1112–1125.

Chen, H., Chang, H. J., & Howes, A. (2021). Apparently irrational choice as optimal sequential decision making. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 792–800).

Clarke, R. G., & Statman, M. (1998). Bullish or bearish? *Financial Analysts Journal*, *54*(3), 63–72. doi: 10.2469/faj.v54.n3.2182

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, *66*(4), 1047-1108. doi: https://doi.org/10.1111/j.1540-6261.2011.01671.x

Craik, K. J. W. (1967). *The nature of explanation* (Vol. 445). CUP Archive.

Davis, P., Pagano, M., & Schwartz, R. (2009). Divergent expectations. In *Technology and regulation: How are they driving our markets?* (pp. 85–100).

Fisher, K. L., & Statman, M. (2000). Investor sentiment and stock returns. *Financial Analysts Journal*, *56*(2), 16–23. doi: 10.2469/faj.v56.n2.2340

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*(10), e1701381. doi: 10.1126/sciadv.1701381

Greenwood, R., & Shleifer, A. (2014, 01). Expectations of returns and expected returns. *The Review of Financial Studies*, *27*(3), 714-746. Retrieved from https://doi.org/10.1093/rfs/hht082 doi: 10.1093/rfs/hht082

Haruvy, E., Lahav, Y., & Noussair, C. N. (2007, December). Traders' expectations in asset markets: Experimental evidence. *American Economic Review*, *97*(5), 1901–1920. doi: 10.1257/aer.97.5.1901

Hill, A., Raffin, A., Ernestus, M., Gleave, A., Kanervisto, A., Traore, R., ... Wu, Y. (2018). *Stable baselines.* https://github.com/hill-a/stable-baselines. GitHub.

Hull, J. (2017). *Options, futures, and other derivatives, global edition.* (9th ed. ed.). Harlow, United Kingdom: Pearson Education Limited.

Jokinen, J. P., Wang, Z., Sarcar, S., Oulasvirta, A., & Ren, X. (2020). Adaptive feature guidance: Modelling visual search with graphical layouts. *International Journal of Human-Computer Studies*, *136*, 102376.

Jokinen, J. P. P., Kujala, T., & Oulasvirta, A. (2021). Multitasking in driving as optimal adaptation under uncertainty. *Human Factors*, *63*(8), 1324-1341. Retrieved from https://doi.org/10.1177/0018720820927687 (PMID: 32731763) doi: 10.1177/0018720820927687

Junior, F. E. F., & Oulasvirta, A. (2025). *Agentforge: A flexible low-code platform for reinforcement learning agent design.* Retrieved from https://arxiv.org/abs/2410.19528

Littman, M. L. (2009). A tutorial on partially observable markov decision processes. *Journal of Mathematical Psychology*, *53*(3), 119-125. (Special Issue: Dynamic Decision Making) doi: https://doi.org/10.1016/j.jmp.2009.01.005

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91. Retrieved 2025-01-30, from http://www.jstor.org/stable/2975974

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms.*

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., Kamaev, V. A., et al. (2013). A survey of forecast error measures. *World applied sciences journal*, *24*(24), 171–176.

Shiller, R. J. (2000). Measuring bubble expectations and investor confidence. *The Journal of Psychology and Financial Markets*, *1*(1), 49–60.

Sun, S., Wang, R., & An, B. (2023, March). Reinforcement learning for quantitative trading. *ACM Trans. Intell. Syst. Technol.*, *14*(3). Retrieved from https://doi.org/10.1145/3582560 doi: 10.1145/3582560

Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., Cola, G. D., Deleu, T., ... Younis, O. G. (2024). *Gymnasium: A standard interface for reinforcement learning environments.*

Yang, H., Liu, X.-Y., Zhong, S., & Walid, A. (2021). Deep reinforcement learning for automated stock trading: an ensemble strategy. In *Proceedings of the first acm international conference on ai in finance.* New York,

NY, USA: Association for Computing Machinery. doi: 10.1145/3383455.3422540

Åström, Karl Johan. (1965). Optimal Control of Markov Processes with Incomplete State Information I. , *10*, 174–205. doi: 10.1016/0022-247X(65)90154-X