Market Self-Learning of Signals, Impact and Optimal Trading: Invisible Hand Inference with Free Energy

(or, How We Learned to Stop Worrying and Love Bounded Rationality)

Igor Halperin and Ilya Feldshteyn

NYU Tandon School of Engineering

e-mail: igor.halperin@nyu.edu, if@q-rd.com

May 16, 2018

Abstract:

We present a simple model of a non-equilibrium self-organizing market where asset prices are partially driven by investment decisions of a bounded-rational agent. The agent acts in a stochastic market environment driven by various exogenous "alpha" signals, agent's own actions (via market impact), and noise. Unlike traditional agent-based models, our agent aggregates *all* traders in the market, rather than being a *representative* agent. Therefore, it can be identified with a bounded-rational component of the market *itself*, providing a particular implementation of an Invisible Hand market mechanism. In such setting, market dynamics are modeled as a fictitious *self-play* of such bounded-rational market-agent in its adversarial stochastic environment. As rewards obtained by such self-playing market agent are not observed from market data, we formulate and solve a simple model of such market dynamics based on a neuroscience-inspired Bounded Rational Information Theoretic Inverse Reinforcement Learning (BRIT-IRL). This results in effective asset price dynamics with a non-linear mean reversion - which in our model is generated *dynamically*, rather than being postulated. We argue that our model can be used in a similar way to the Black-Litterman model. In particular, it represents, in a simple modeling framework, market views of common predictive signals, market impacts and implied optimal dynamic portfolio allocations, and can be used to assess values of private signals. Moreover, it allows one to quantify a "market-implied" optimal investment strategy, along with a measure of market rationality. Our approach is numerically light, and can be implemented using standard off-the-shelf software such as TensorFlow.

We would like to thank Ernest Baver, Eric Berger, Jean-Philippe Bouchaud, Peter Carr, Sergei Esipov, Andrey Itkin, Vivek Kapoor, Dan Nudelman and Nikolai Zaitsev for stimulating remarks and discussions. All errors are ours.

1 Introduction

This paper presents a simple 'structural' model of price dynamics in a financial market. Though based on concepts not commonly used in Finance (Reinforcement Learning, Information Theory, Physics etc. see below), the model we suggest is mathematically rather simple *at the end* (see Eq.(109)), after getting through a 'story' behind its structure. It is designed as both a *practical tool* for market practitioners, and a *theoretical model* of a financial market that can be explored further using simulations and/or analytical methods. For definitiveness, we focus in this paper on stock markets, though the same approach can be applied to other markets in the same way.

In a way, the main idea of a model presented below can be formulated as a *dynamic* and *data-driven* extension of an approach to modeling excess returns that was suggested in the seminal Black-Litterman (BL) model [8]. As will be shown below, a structural asset return model arising in our solution to this problem has some interesting properties such as the presence of *mean reversion* in stock prices, which in our framework appears as a result of joint actions of all traders in the market that *dynamically* implement Markowitz-type mean-variance portfolio strategies.

In essence, the BL model flips the Markowitz optimal portfolio theory [32] on its head, and considers an *inverse* optimization problem. Namely, it starts with an observation that a *market portfolio* (as typically represented by the S&P500 index) is, by definition, the optimal "market-implied" portfolio. Therefore, if we consider such a given market portfolio as an *optimal* portfolio, then we can *invert* the portfolio optimization problem, and ask what is the optimal asset allocation *policy* that corresponds to this optimal market portfolio. Within the framework of a single-period Markowitz mean-variance optimization [32], this translates into market-implied values of expected returns and covariances of returns.

Respectively, this framework was suggested by Black and Litterman as a way to assess values of private "alpha" signals in generating excess returns. The BL model was explicitly re-interpreted as an *inverse* portfolio optimization problem by Bertsimas *et. al.* [7], along with proposing some extensions such as robust inverse optimization. Note that the inverse optimization in [7] is still performed in a single-period (one time step) setting, the same as in the original BL model [8] and in the Markowitz mean-variance portfolio model [32].

A model suggested in this paper extends such inverse optimization view of the market portfolio to a *dynamic*, multi-period setting. While this requires some new mathematical tools, the *outputs* of the model can be used in essentially the same way as the outputs of the BL model: to assess the value of private "alpha" signals, and design trading strategies according to own assessments of joint effects of signals and market impacts from trades on expected excess returns.

An important difference of our model from a majority of market models used in both the industry and the academia is that our model does **not** assume a competitive market equilibrium. As discussed at length by Duffie [14], this paradigm underlies three cornerstone Nobel prize-winning theories of modern Finance, which are used by many practitioners on both the sell and buy sides. On the other hand, George Soros, a famous guru of financial markets, called this paradigm an "absurd postulate"[1].

Our model can be interpreted as an attempt to reconcile such opposite views. Our suggested answer is that both sides are right in their own ways, but we offer a practical and easily computable *unifying* framework. This allows us to quantify Soros' critique and propose a simple model that can be used to describe markets in three different states: disequilibrium, quasi-

[1] "Economics ended up with the theory of *rational expectations*, which maintains that there is a single optimum view of the future, that which corresponds to it, and eventually all the market participants will converge around that view. This postulate is *absurd*, but it is needed in order to allow economic theory to model itself on *Newtonian Physics*." (G. Soros). We thank Vivek Kapoor for this reference.

equilibrium, and a perfect thermal equilibrium. The latter scenario may only occur if there is no inflow of information in a market - hardly a realistic scenario.

The last case of a perfect thermal equilibrium corresponds to assumptions of the competitive market equilibrium paradigm. While we believe that for financial markets the last limit is in a way 'non-physical'[2], it is the limit described by competitive market equilibrium models such as the Modigliniani-Miller's capital structure irrelevance for the market value of a corporation, the Capital Asset Pricing Model (CAPM) of William Sharpe (1964), and the Black-Scholes model of option pricing[3][14].

These models consider market dynamics as equilibrium fluctuations around a perfectly thermodynamically equilibrium market state. Therefore, they implicitly assume that there is no inflow/outflow of money and information in a market as a whole, and the market is in a state of a maximum entropy. This may be a reasonable approximation in a steady/slow market, which may explain why these models work reasonably well under 'normal' market conditions.

But this assumption of competitive market equilibrium also suggests that these models should behave progressively worse during periods of market instabilities, crises and market crashes - an observation that seems to be widely recognized in the literature.

The general reason for such model failures when they are needed most is that in all these cases, a view of a market as an equilibrium fluctuation around a stationary state where entropy is already maximized becomes inadequate to describe market dynamics, see also below on analogies with self-organizing systems and living organisms.

The above remarks concern with potential theoretical implications of our framework. Irrespective, our model also attempts to address the needs of market *practitioners* that want to make a profit rather than do a theoretical research into the dynamics of the markets.

To this end, in addition to providing a multi-period extension of the Black-Litterman model[4], our model produces "market-implied" values of market impact parameters and risk aversion parameter of an agent that *dynamically maintains* such market-optimal portfolio, as well as a "market-implied" optimal investment strategy, which can be viewed for monitoring of the market or individual players in the market (see below). Given an explicit formula produced by our model for a market-implied optimal strategy, expressions like 'a strategy that beats the market' can now be probably given a more quantitative meaning *ex-ante* rather than *ex-post*.

Finally, one more interesting insight may be provided by the fact that one of parameters estimated by the model from market data is a parameter β that describes a degree of rationality of the market (another name for β is the "inverse temperature" of the market). This suggests that market-implied value of β can be used as a monitoring tool and possibly a predictive signal to have an aggregative view of a market, a specific exchange, or even a specific dealer[5].

[2]It is non-physical in the sense that it contradicts the very existence of markets where market makers generate liquidity and speculators make profits by digesting new information - neither should exist in competitive market equilibrium models. This is because a perfect equilibrium is only possible for a closed system that does not exchange information with an outside world. Therefore, competitive market equilibrium models do not try to answer the question *why* markets exist, but rather simply postulate first-order optimality (equilibrium) conditions, and then explore the consequences [46]. In physics, a perfect thermodynamic equilibrium is achieved in the thermodynamic limit of a closed system, and corresponds to a state of a 'heat death of the Universe' [29].

[3]The Black-Scholes model relies on a weaker form of competitive market equilibrium paradigm known as the no-arbitrage principle [14].

[4]Note that because the latter is a one-period model, a question of a market in equilibrium vs non-equilibrium cannot even be formulated in this framework.

[5]The latter case corresponds to a possible application of the model developed in this paper for an individual investor rather than for the market as a whole, see also below.

1.1 Outlook of our approach

The main intuitive idea behind the model can be introduced as follows. While the real market dynamics are highly complex as they are driven, to a large extent, by a very large number of individual rational or bounded-rational market participants, is is commonly known that market players exhibit a strong tendency to a herding behavior: when markets are up, everyone is buying, and when markets are down, everyone is selling.

This suggests a concept of a *representative investor* whose objective is to optimize a given investment portfolio given some objective function. Such representative investor is otherwise known in the literature as an *agent*. In this view of the world, an environment, i.e. the market, is clearly *external* to the agent.

But what if we take an *inverse* optimization view of this problem, as in Black-Litterman [8]? In this approach, the optimal portfolio is already *known*, it is the market portfolio itself. But then, who is an *agent* that *dynamically* maintains (rebalances) such market-optimal portfolio?

We can identify such agent with a 'collective mode' of *all* individual traders involved in the market, that are guided in their decisions by a commonly agreed set of predictors $\mathbf{z}_t$ which may include news, other market indicators and/or indexes, variables describing the current state of the limit order book, etc. Therefore, the first difference of our framework from conventional utility-based models is that our agent is a *sum* of all investors, rather than their *average*, i.e. a 'representative' investor.

Because such agent aggregates actions of a partly *homogeneous* and partly *heterogeneous* crowd of individual investors, it can not be a fully *rational* agent, but rather should be represented as an agent with *bounded rationality*. Bounded rationality, which will be explained in more details below, is a second key difference of our framework from a classical agent-based approach.

Furthermore, because jointly all individual trades by *all* market participants amount to *actual* market moves that dynamically re-adjust the market-optimal portfolio, such agent can then be identified with a bounded-rational component of the market *itself*.

If we adopt such view, the actual dynamics of market prices can now be mathematically modeled as a sequential decision-making process of such bounded-rational agent who is engaged in *self-learning via self-play* in a partly controllable and partly uncontrollable environment which is identified with the *rest* of the market. The first component identified with a RL agent can then be thought of as a 'mind' of such *self-organizing* market, that learns about its environment and itself via self-play in such open environment.

Our agent embodies an 'Invisible Hand' of the market, which is *goal-oriented* in our framework, as will be made more clear below. The Invisible Hand is implemented in our model as a fictitious self-play of a bounded-rational RL agent. Agent's self-play amounts to mimicking a risk-averse investor seeking a dynamic Markowitz-optimal portfolio, while actions of this investor are randomized by entropy. As will be shown below in Sect. 4.6, this is mathematically equivalent to portfolio optimization in a two-party game with an *adversarial* player, such that the original agent and its imaginary adversary form a Nash equilibrium. As a result, the agent simultaneously mimicks *all* traders as a bounded-rational 'mind' of a self-organizing market[6].

This produces a dynamic model of market price dynamics, where a *total* price impact of all traders in the market, who try to construct Markowitz-optimal portfolios, amounts to a *dynamically* generated mean reversion in market-observed asset returns. The resulting model

[6]Equivalence between self-organization in dynamic systems and sequential decision making was emphasized by Yukalov and Sornette in [57]. A similar approach in neuroscience is a unified free-energy model of the brain of Friston [19], see also [39] for recent applications of the free-energy principle to living systems. In short, this approach suggests that "all biological systems instantiate a hierarchical generative model of the world that implicitly minimizes its internal entropy by minimizing free energy" [39].

can be interpreted as a Geometric Mean Reversion model with external signals, where mean reversion arises dynamically, rather than being introduced by hands, as is done in descriptive models of market dynamics. The resulting model can be viewed as a non-linear factor model for returns that can be estimated using standard methods of statistics such as Maximum Likelihood.

More than that, because our resulting asset return model is a *structural* model, mean reversion in our model has a 'story' behind it. In our approach it is produced by a total bounded-rational action of all (bounded-rational or rational) agents in the market, that dynamically optimize their investment portfolios following mean-variance strategies. This can be compared with a mechanism for mean reversion due to zero-intelligence 'noise traders' suggested in 1988 by Poterba and Summers [44]. We have only one agent, but it is bounded-rational, unlike many noise traders with a zero total rationality/intelligence in the model of [44].

1.2 Possible insights from the model

Because we formulate the problem in a setting of inverse, rather than direct portfolio optimization, the objective of a bounded-rational agent can be viewed as the problem of rebalancing its own fictitious "shadow" portfolio, such that it is kept as close as possible to the market portfolio in such continuous self-play. Note that except for a bounded rationality of an agent assumed in such framework, it resembles the classical pole balancing problem of Reinforcement Learning (see e.g. [49]), where now it is the market portfolio that serves the role of a pole, and we *invert* the problem.

Our model is also quite similar to an index tracking problem, except we set it as an *inverse* optimization problem to infer market views of its own dynamics, instead of solving a *forward* optimization problem of finding a good tracking portfolio for an index. Note that data for such model formulation is readily available as level-1 limit order book (LOB) data (level-2 LOB data can be incorporated in the model via a set of external predictors $\mathbf{z_t}$, see below).

This is unlike a (mathematically identical) portfolio optimization problem for an *individual* trader, that would require trader's *proprietary* execution data for model estimation. If, however, such trader's proprietary data *are* available, our framework can be used in the same way to construct a probabilistic model of the trader. This could be used, in particular, by regulators for monitoring activities of exchanges or individual traders.

Note that in a single-period setting, our problem formulation brings us back to the BL model, where instead of multi-period trading strategies, we have just single-period optimal portfolio allocations.

On the other hand, in a multi-period formulation, it extends the setting of the BL model in multiple ways, including a *probabilistic* model of observed actions, that takes into account effects that are absent in a single-period settings, such as dynamic market impacts and dynamically changing predictors. As our model is probabilistic, i.e. *generative*, it can be used for *forward* simulation of dynamics.

Also note that in a multi-period setting, it is a combination of non-linearity induced by market impacts and dynamical exogenous predictors $\mathbf{z_t}$ that may produce potentially very rich dynamics that would be driven by a combinators of external signals $\mathbf{z}_t$, non-linear system feedback via agent's trades, and incontrollable noise. As we will show below, our model is tractable in a quasi-equilibrium setting using conventional tools of constrained convex optimization, due to its simple structure with a *quadratic* non-linearity of dynamics.

On the other hand, external signals $\mathbf{z}_t$ have their own dynamics, and might operate at different frequencies from typical times of market responses to news, events, and other changes in predictors $\mathbf{z}_t$. Therefore, due to its non-linearity, and depending on a relation between character-

istic times of market responses and signal changes, the model can describe both an equilibrium and non-equilibrium settings with such non-linear dynamics. A combination of non-linearity of dynamics with particular patterns of external signals $\mathbf{z}_t$ whose changes provide new information to the agent, can lead to potentially very rich dynamics.

We will leave an exploration into *generative* properties of our model for a future research. The focus of the present paper is rather of a *batch-mode* (off-line) learning from *past* data. Such learning can be done using model-free or model-based Reinforcement Learning (see e.g. [49]) when rewards are observable, or Inverse Reinforcement Learning (IRL) when they are not. As in our case rewards (either of a single investor, or 'market-implied' rewards) are *not* observable, we rely on an IRL-based approach for learning in such setting.

Our model, where rewards are not observed but rather *inferred* from data, belongs in a class of model-based IRL approaches with a parametrized reward function and dynamics. The objective of modeling in this approach is to infer the *reward function* and *action policy* from data by tuning model parameters. As in our case we solve the dynamic inverse portfolio optimization problem for a *market-optimal* portfolio, our IRL approach infers *market-implied* reward function and optimal action policy.

Note that in typical applications of RL for financial decision making, an agent is typically a (representative or particular) trader or a financial institution who is *external* to the market. In contrast, in our approach, an agent is the bounded-rational component of the market *itself*, as it is now *inseparable* from the market, so long as it *maintains* the market-optimal portfolio.

Therefore, our model is a dynamic model of the market itself, rather than a model of an external representative investor in such market. Our model is inspired by IRL, Information Theory, statistical physics and neuroscience, yet it is based on a simple parametric specification of a one-step reward, and a simple specification of dynamics.

The model is tractable as a non-linearity of dynamics is 'only' quadratic. Furthermore, because we use a simple low-dimensional parametric specification of the 'actual' reward of the agent, the data requirements for the model are modest. The model does not need tens, hundreds, or thousands years of training data, even though both the state and action spaces in our problem are very high-dimensional.

Computationally, the model amounts to a simple and transparent scheme, rather than being a black-box model in the spirit of Deep Reinforcement Learning. This is because a simple parametric specification of the model enables proceeding without sophisticated function approximations that are typically implemented in Deep Reinforcement Learning by deep neural networks. The main computational tool employed by the model is (an iterative version of) the conventional Maximum Likelihood estimation method available via standard off-the-shelf numerical optimization software. This can be conveniently done with TensorFlow using its automatic differentiation functionality.

The paper is organized as follows. In Sect. 2, we review related work, and simultaneously provide further high-level details of our framework. In Sect. 3 we introduce our notation and describe an investment portfolio of stocks. In Sect. 4, we present a RL formulation of the model. Sect. 5 re-formulates the model in an IRL setting, and presents our solution to the problem of finding an optimal policy and reward function for a single investor case. The IRL problem for the market as a whole is addressed in Sect. 6. The same section introduces an effective market dynamics model that is obtained as a by-product of our IRL solution. Experiments are presented in Sect. 7. Sect. 8 discusses our results and outlines future directions. A brief summary is given in Sect. 9.

2 Related work

Our model builds on several threads developed separately by the Quantitative Finance, Reinforcement Learning, Information Theory, Physics and Neuroscience communities. Here we provide a brief overview of related work in these different fields that have a close overlap with the model developed here, as well as explain their relation with our approach.

2.1 What kind of equilibrium holds for markets?

Quoting Duffie, "while there are important alternatives, a current basic paradigm for valuation, in both academia and in practice, is that of competitive market equilibrium" [14]. While this was said in 1997, this assessment remains true to this day.

Of course, deficiencies of standard financial models based on competitive market equilibrium and/or no-arbitrage paradigm were not left unnoticed both within the financial community, and among researchers in other disciplines, most notably physics and computer science. The latter disciplines contributed a number of interesting and fresh ideas to financial modeling [9], [46]. In particular, agent-based models may provide interesting insight into how financial markets can operate when viewed as *evolving complex systems*, see e.g. [1].

The main challenge with agent-based models is that while they are capable of explaining some stylized facts of the market, they can hardly be turned, at least at the current stage, into practically useful tools - in part, due to their high computational complexity. While models such as CAPM or the Black-Scholes model may miss some important features of real markets, they also work reasonably well under certain market and trade conditions, and they are fast.

Yet, to better model effects such as market liquidity, Amihud *et. al.* suggested that, instead of assuming competitive market equilibrium, researchers should assume an "equilibrium level of disequilibrium" [3]. In physics, this is normally referred to as a non-equilibrium steady state.

Viewing markets as an 'equilibrium disequilibrium' is beneficial if we are willing to consider them as evolving and self-organizing systems that may bear some similarities to living organisms. Boltzmann and Schödinger have emphasizes that activities of living organisms are impossible in thermal equilibrium, and necessarily depend on harnessing a pre-existing disequilibrium. In other words, as a consequence of the Second Law of thermodynamics, living organisms can only exist as processes *on the way* to a state of maximum entropy describing a thermal equilibrium, but *not* in this state itself [51], [34].

A demand-based option pricing model that does not rely on no-arbitrage assumptions was proposed by Garleanu *et. al.* [20]. A Reinforcement Learning based option pricing model that similarly does not rely on no-arbitrage but uses instead a model-free and data-driven Q-learning approach was proposed by one of the authors in [26][7]. Residual inefficiencies of markets resulting from multi-step strategies and market impact were studied by Esipov [16].

2.2 Optimal portfolio execution

A close analog of a setting of our model is a problem of optimal execution in stock trading, one of the classical problems of Quantitative Finance. The problem amounts to designing an optimal strategy (policy) for partitioning a large trade order to buy or sell a large block of a stock of some company into smaller chunks, and buy these chunks sequentially so that a potential market impact would be minimized, and respectively the total cost of implementing the trade will be

[7] If so desired, the latter model can also be constructed as an arbitrage-free model, by using a suitable utility function, instead of a quadratic utility [26].

minimized as a result. This is a problem solved many thousands times a day by brokers, as well as those asset managers and hedge funds that execute such trades themselves instead of passing trade orders to brokers.

The classical way to address such (forward) optimization problem is to start with building and calibrating models for stock dynamics and price impact. Provided this is done, the next step is to define a *cost function* that specifies loss that will be observed upon taking certain actions in certain states. If we focus for now on execution strategies that involve only market orders but not limit orders, then these market orders will be our *actions* $\mathbf{a}_t$[8].

Assume that the trade order is to sell N shares of a given stock within time T. *Optimal* actions $\mathbf{a}_t^\star$ are obtained from (forward) optimization of total cumulative costs of execution, as determined by a *policy* $\pi_t = \pi_t(\mathbf{y}_t)$. Here t is current time and $\mathbf{y}_t$ is a *state* vector of the system that includes the current mid-price of the stock S_t, the number of stocks n_t currently held, and values of external signals $\mathbf{z}_t$ that may include, in particular, predictors derived from properties of the limit order book (LOB).

If $\pi_t^\star(\mathbf{y}_t)$ is a (deterministic) optimal policy, then the optimal action $\mathbf{a}_t^\star$ is simply the value $\mathbf{a}_t^\star = \pi_t^\star(\mathbf{y}_t)$. The classical multi-period optimal execution problem was formulated in the dynamic programming (DP) setting by Bertsimas and Lo [6] for a risk-neutral investor, and then extended by Almrgen and Chriss [2] to a risk-averse investor.

2.3 Inverse portfolio optimization

In this paper, we consider three (related) modifications to the direct optimization problem described above. *First*, we take the view of dynamic *inverse* optimization, in the spirit of the Black-Litterman model [8] and its reformulation in [7], and assume that such optimization problem was already *solved* by the market itself. Respectively, we look for market-implied optimal trading policies/strategies rather than trading/execution strategies of an individual investor. However, our market-wise aggregate trader-agent does the same thing as nearly all traders in the market do, i.e. it dynamically optimizes its own investment portfolio.

2.4 Dynamic portfolio management with constrained convex optimization

In our specification of single-step rewards, or negative costs of trading, we follow a large literature on multi-period mean-variance optimization. An accessible review of a version of such mean-variance optimization is given by Boyd *et. al.* [10]. We largely adopt the notation and assumption of the portfolio model suggested by Boyd *et. al.*, while in addition we explicitly introduce predictors and market impacts effects not considered in [10]. Quadratic objective functions for multi-period portfolio optimization discussed at length in [10] are formulated within the conventional DP approach that assumes a known model, including a known risk aversion parameter.

2.5 Stochastic policies

The *second* modification we make to the classical formulation of the optimal execution problem is that we consider *stochastic* (probabilistic), rather than *deterministic* policies π. A stochastic

[8] This is sufficient if we look at aggregate actions of *all* traders, i.e. the market itself, which is the main setting of our model in this paper. If the model is applied to an *individual* investor, restricting a model to modeling only market orders may be a reasonable approximation for liquid stocks, while for stocks with limited liquidity optimal strategies may involve combinations of market and limit orders. Extensions of our framework to such setting of mixed market and limit orders for individual investors will be provided elsewhere.

policy $\pi_t(\mathbf{y}_t)$ describes a probability distribution, so that action $\mathbf{a}_t$ becomes a sample from this distribution, $\mathbf{a}_t \sim \pi(\mathbf{y}_t)$, rather than a fixed number. Respectively, an *optimal* action would be a sample from an *optimal* policy, $\mathbf{a}_t^\star \sim \pi^\star(\mathbf{y}_t)$. Deterministic policies can now be viewed as a special case of stochastic policies, where the action distribution is a Dirac delta-function $\pi(\mathbf{a}_t|\mathbf{y}_t) = \delta\left(\mathbf{a}_t - \mathbf{a}_t^\star(\mathbf{y}_t)\right)$ where $\mathbf{a}_t^\star(\mathbf{y}_t)$ is an optimal action for state $\mathbf{y}_t$, which corresponds to a deterministic policy setting of the classical DP approach.

What would be the meaning of such probabilistic modeling of execution orders, given that at the end they amount to specific numbers, rather than probabilities? Such choice can be justified for both the direct and inverse problems of optimal execution.

Let's start with an argument why stochastic policies can be useful for *direct* optimization. Given that parameters defining optimal strategies are estimated from data, the resulting policy is always stochastic *de-facto*, even though this is *not* explicitly recognized in deterministic policy execution models such Bertsimas and Lo [6] and Almgren and Chriss [2] models.

Adapting stochastic policies as a principal modeling tool allows one to explicitly *control uncertainty* around an optimal action in each state of the world. The latter can be identified with a mode of the policy distribution, while uncertainty around this value would be specified by properties of this distribution, and measured, in a simplest case, by the variance of a predicted optimal action. This argument is rather similar to an argument for stochastic, rather than deterministic, portfolio allocations for a one-period Markowitz-like portfolio optimization problem, which was put forward by Marschinski *et. al.* [33].

In the setting of *inverse* portfolio optimization adopted in this paper, the usefulness of stochastic policies becomes even more evident. In this case, stochastic policies are needed in order to account for a possible *sub-optimality* of a policy used in generated data. Such events would be incompatible with an assumption of a strict optimality of *each* action in the data, leading to vanishing probabilities of observed execution paths. Reliance on stochastic rather than deterministic policies allows one to cope with possible sub-optimality of historical data.

2.6 Reinforcement Learning

Deterministic policy optimization problem in a dynamic mean-variance optimization setting similar to Boyd *et. al.* [10] was reformulated in a data-driven Reinforcement Learning (RL) way by Ritter [25]. Ritter considers the classical *on-line* Q-learning for the problem of multi-period portfolio optimization from data, using a quadratic risk-adjusted cost function. This translates the problem into a *data-driven* forward optimization that can can be solved, given enough training data, by the famous Q-Learning of Watkins and Dayan [56].

The difference between our approach and Ritter's is that we consider an *off-line* (batch-mode) learning, and we do *not* observe one step costs (or, equivalently, negative rewards). Therefore, our setting is that of IRL, while Ritter [25] considers an on-line RL formulation. Also, unlike [25] that uses a discretized state space, we use a continuous-state formulation. Furthermore, Ritter considers a general optimal portfolio investment problem for a given (representative?) investor, while here we focus on modeling an agent that represents a bounded-rational component of the market as a whole. This transforms our approach into a *market* model, unlike the case considered by Ritter, which is a *trader* model.

Quadratic risk-adjusted objective functions were considered in an apparently different problem of optimal option pricing and hedging using a model-free, data-driven approach in the work by one of the authors [26, 27]. The approach used in this work assumes *off-line*, batch-mode learning, that enables using data-efficient batch RL methods such as Fitted Q Iteration [15, 37].

We use entropy-regularized Reinforcement Learning in the form suggested by Tishby and

co-workers under the name of G-learning [18], as a way to do Reinforcement Learning in a noisy environment. While [18] assumed a tabulated discrete-state/discrete-action setting, in our case both the state and action spaces are high-dimensional continuous spaces. For a tutorial-style introduction to Information-constrained Markov Decision Processes, see Larsson *el. al.* [30].

2.7 Inverse Reinforcement Learning

A *third* modification we introduce to the classical portfolio optimization scheme is that we assume that some critical model parameters are *unknown*. Note that forward optimization using Dynamic Programming always assumes that dynamics and model parameters are *known*, or estimated using independent models. In particular, market impact parameters or risk aversion parameters are not easy to mark down without using additional models to estimate them *before* using them in a direct execution optimization method. Moreover, traders do not necessarily even *think* in terms of *any* utility function, and respectively may not even know their *own* risk-aversion parameter λ.

Unlike such DP approach, in our model we treat these parameters as *unknown*, and estimate them *simultaneously* with estimating optimal policy from historical trading data. What we obtain with such procedure can be interpreted as *implied* market impact and risk aversion parameters, similar to how implied volatilities are used to price and hedge options in option markets. In particular, even if traders may not *think* in terms of a quadratic utility function with some pre-determined value of λ, their *observed behavior* might be consistent with such simple utility function, with some *data-implied* risk aversion rate $\lambda = \lambda_{imp}$.

Note that when risk aversion λ and parameters determining market impact are unknown, it also means that one-step *costs* (see below) are unknown as well. Our data therefore consist of sequences of states and actions, but it does not reveal *costs* incurred by following these actions. Such problems of estimating costs (or rewards) from an observed behavior are solved using methods of Inverse Optimal Control (IOC) when dynamics are *known*, or using Inverse Reinforcement Learning (IRL) when dynamics are *unknown*.

In this paper, we address this problem using model-based IRL. Our framework relies on a model for specification of one-step costs, market impact, and risk metrics (we will use quadratic risk measures going forward). On the side of IRL literature, our approach is based on Maximum Entropy IRL developed in [58], and extended to continuous-space formulation in [36]. A closely related method is Iterative Quadratic-Gaussian Regulator (IQGR) of Todorov and Li [53].

2.8 Neuroscience and biology

Our approach is similar to a Free-Energy Principle (FEP) approach to living systems and the brain function developed by Friston and collaborators in [19, 39]. Under this formalism, "for an organism to resist dissipation and persist as an adaptive system that is a part of, coupled with, and yet statistically independent from, the larger system in which it is embedded, it must embody a probabilistic model of the statistical interdependencies and regularity of its environment" [39].

Our model applies similar approach, based on ideas from statistical thermodynamics, to the market as a dynamic persistent and adaptive system that embodies a bounded-rational RL agent that imitates a 'mind' of the market as a *goal-directed* 'living organism' in an adversarial environment. We implement the above requirement that the agent should embody a probabilistic model of its environment by formulating this problem as Inverse Reinforcement Learning. The free energy arises in this approach either as a way to regularize (Inverse) Reinforcement Learning in a noisy environment by entropy, as in G-learning [18], or as a way to model bounded-rational decision-making of the agent by imposing constraints on information processing costs [40, 52, 42],

or equivalently as a way to account for an adversarial character of the environment, see the next two sections.

2.9 Thermodynamics, Bounded Rationality and Information Theory

Another, and mathematically equivalent way to introduce entropy and free energy into the problem of sequential decision-making, was formulated within an information-theoretic and physics-inspired approach in [40, 52, 42]. In particular, Ortega *et. al.* [40, 42] emphasize that a regularization 'inverse temperature' parameter β that corresponds to a cost of information processing in a system, can also be interpreted as a *degree of rationality* of an agent that dynamically maximizes its free energy (i.e. an entropy-regularized value function).

This interpretation is provided by noting that parameter β determines complexity of a search for a better policy starting from a given prior policy [40, 42]. Agents with large $\beta \to \infty$ can afford a highly complex (costly) search for a better policy, and therefore are more rational than agents that live in a world with a small value $\beta \to 0$. In this regime, an agent cannot afford to change from a prior policy, and therefore behaves as an irrational (entropy-dominated) agent. The information-theoretic approach thus provides a quantitative and tractable framework for a bounded-rational agent of Simon [45].

2.10 Self-play, adversarial learning, and the free energy optimization

An adversarial interpretation of Information Theoretic Bounded Rationality was suggested in [41] where it was shown that a single-agent free energy optimization is equivalent to a fictitious game between an agent and an imaginary adversary. In our model, we have a similar setting, where an agent representing a bounded-rational component of the market optimizes its free energy. The optimization amounts to a dynamical optimization of agent's portfolio in a stochastic market environment with information processing costs. The latter are expressed as an entropy regularization of a value function, see below. As will be shown in Sect. 4.6, using the method of [41], such self-play can be equivalently viewed as *adversarial* learning in a fictitious *two-party* game with an adversarial opponent.

2.11 Bounded Rational Information Theoretic IRL (BRIT-IRL)

Our approach integrates ideas of Maximum Entropy IRL with a Bounded Rational Information-Theoretic interpretation of the process of learning, and applies them to make inferences of an 'Invisible Hand', in the spirit of the Black-Litterman model. In splitting the market into its bounded-rational self and the rest, the model also has strong similarities with the free-energy approach to the brain and biological systems [19, 39]. In our approach, such view is applied to a financial market as a *dynamic self-organizing* system, with a focus on *inverse* rather than direct learning.

As our setting is of *inverse* learning, instead of assuming some value of degree of rationality β, we *infer* such parameter implied by the market data within our model. This produces a dynamic "market-implied" index of rationality β_t that can be used as a simple monitoring statistic, or possibly as a predictor of future events in the market. If the model is applied to an *individual* investor, provided corresponding proprietary trading data are available, it can produce an implied 'amount of rationality' of that particular trader.

3 Investment portfolio

We adopt the notation and assumption of the portfolio model suggested by Boyd *et. al.* [10]. In this model, dollar values of positions in n assets $i = 1, \ldots, n$ are denoted as a vector $\mathbf{x}_t$ with components $(x_t)_i$ for a dollar value of asset i at the beginning of period t. In addition to assets $\mathbf{x}_t$, an investment portfolio includes a risk-free bank cash account b_t with a risk-free interest rate r_f. A short position in any asset i then corresponds to a negative value $(x_t)_i < 0$. The vector of mean of bid and ask prices of assets at the beginning of period t is denoted as $\mathbf{p}_t$, with $(p_t)_i > 0$ being the price of asset i. Trades $\mathbf{u}_t$ are made at the beginning of interval t, so that asset values $\mathbf{x}_t^+$ immediately after trades are deterministic:

$$\mathbf{x}_t^+ = \mathbf{x}_t + \mathbf{u}_t \tag{1}$$

The total portfolio value is

$$v_t = \mathbf{1}^T \mathbf{x}_t + b_t \tag{2}$$

where $\mathbf{1}$ is a vector of ones. The post-trade portfolio is therefore

$$v_t^+ = \mathbf{1}^T \mathbf{x}_t + b_t^+ = \mathbf{1}^T (\mathbf{x}_t + \mathbf{u}_t) + b_t^+ = v_t + \mathbf{1}^T \mathbf{u}_t + b_t^+ - b_t \tag{3}$$

We assume that all re-balancing of stock positions are financed from the bank cash account (additional cash cost related to the trade will be introduced below). This imposes the following 'self-financing' constraint:

$$\mathbf{1}^T \mathbf{u}_t + b_t^+ - b_t = 0 \tag{4}$$

which simply means that the portfolio value remains unchanged upon an instantaneous re-shuffle of the wealth between the stock and cash:

$$v_t^+ = v_t \tag{5}$$

The post-trade portfolio v_t^+ and cash are invested at the beginning of period t until the beginning of the next period. The return of asset i over period t is defined as

$$(r_t)_i = \frac{(p_{t+1})_i - (p_t)_i}{(p_t)_i}, \quad i = 1, \ldots, n \tag{6}$$

Asset positions at the next time period are then given by

$$\mathbf{x}_{t+1} = \mathbf{x}_t^+ + \mathbf{r}_t \circ \mathbf{x}_t^+ \tag{7}$$

where $\circ$ stands for an element-wise (Hadamard) product, and $\mathbf{r}_t \in \mathbb{R}^n$ is the vector of asset returns from period t to period $t + 1$. The next-period portfolio value is then obtained as follows:

$$v_{t+1} = \mathbf{1}^T \mathbf{x}_{t+1} = (1 + \mathbf{r}_t)^T \mathbf{x}_t^+ = (1 + \mathbf{r}_t)^T (\mathbf{x}_t + \mathbf{u}_t) \tag{8}$$

Given a vector of returns $\mathbf{r}_t$ in period t, the change of the portfolio value in excess of a risk-free growth is

$$\begin{aligned} \Delta v_t &\equiv v_{t+1} - (1 + r_f) v_t = (\mathbf{1} + \mathbf{r}_t)^T (\mathbf{x}_t + \mathbf{u}_t) + (1 + r_f) b_t^+ - (1 + r_f) \mathbf{1}^T \mathbf{x}_t - (1 + r_f) b_t \\ &= (\mathbf{r}_r - r_f \mathbf{1})^T (\mathbf{x}_t + \mathbf{u}_t) \end{aligned} \tag{9}$$

where in the second equation we used Eq.(4).

3.1 Terminal condition

A terminal condition for the market portfolio is obtained from the requirement that at a planning horizon T, all stock positions should be equal to the actual observed weights of stocks in the market index. This implies that $\mathbf{x}_T = \mathbf{x}_T^M$ where $\mathbf{x}_T^M$ are market cap weights in the S&P 500 index at time T. By Eq.(1), this fixes the action $\mathbf{u}_T$ at the last time step:

$$\mathbf{u}_T = \mathbf{x}_T^M - \mathbf{x}_{T-1} \tag{10}$$

Therefore, action $\mathbf{u}_T$ at the last step is deterministic and is not subject to optimization that should be applied to T remaining actions $\mathbf{u}_{T-1}, \ldots, \mathbf{u}_0$.

If the model is applied to an individual investor, the planning horizon T is an investment horizon for that investor, while the terminal condition (10) can be replaced by a similar terminal condition for the investor portfolio.

3.2 Asset returns model

We assume the following linear specification of one-period excess asset returns:

$$\mathbf{r}_t - r_f \mathbf{1} = \mathbf{W}\mathbf{z}_t - \mathbf{M}^T \mathbf{u}_t + \varepsilon_t \tag{11}$$

where $\mathbf{z}_t$ is a vector of predictors with factor loading matrix $\mathbf{W}$, $\mathbf{M}$ is a matrix of permanent market impacts with a linear impact specification, and ε_t is a vector of residuals with

$$\mathbb{E}\left[\varepsilon_t\right] = 0, \ \mathrm{Var}_t\left[\varepsilon_t\right] = \Sigma_r \tag{12}$$

Equation (11) specifies stochastic returns $\mathbf{r}_t$, or equivalently the next-step stock prices, as driven by external signals $\mathbf{z}_t$, control (action) variables $\mathbf{u}_t$, and uncontrollable noise ε_t.

Though they enter 'symmetrically' in Eq.(11), two drivers of returns $\mathbf{z}_t$ and $\mathbf{u}_t$ play entirely different roles. While signals $\mathbf{z}_t$ are completely *external* for the agent, actions $\mathbf{u}_t$ are *controlled* degrees of freedom. In our approach, we will be looking for *optimal* controls $\mathbf{u}_t$ for the market-wise portfolio. When we set up a proper optimization problem, we solve for an optimal action $\mathbf{u}_t$. As will be shown in this paper, this optimal control turns out to be a linear function of $\mathbf{x}_t$, plus noise. Substituting it back into Eq.(11), this produces effective *dynamically* generated dynamics that involve only stock prices, see Eq.(109) below in Sect. 6.1[9].

3.3 Signal dynamics and state space

For dynamics of signals $\mathbf{z}_t$, similar to [21], we will assume a simple multi-variate mean-reverting Ornstein-Uhlenbeck (OU) process for a K-component vector $\mathbf{z}_t$:

$$\mathbf{z}_{t+1} = (\mathbf{I} - \Phi) \circ \mathbf{z}_t + \varepsilon_t^z \tag{13}$$

where $\varepsilon_t^z \sim \mathcal{N}(0, \Sigma_z)$ is the noise term, and Φ is a diagonal matrix of mean reversion rates.

It is convenient to form an extended state vector $\mathbf{y}_t$ of size $N + K$ by concatenating vectors $\mathbf{x}_t$ and $\mathbf{z}_t$:

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{bmatrix} \tag{14}$$

The extended vector $\mathbf{y}_t$ describes a full state of the system for the agent that has some control of its x-component, but no control of its z-component.

[9] The reader interested only in the final asset return model resulting from our framework but not in its derivation can jump directly to Eq.(109).

3.4 One-period rewards

We first consider an idealized case when there are no costs of taking action $\mathbf{u}_t$ at time step t. An instantaneous random reward received upon taking such action is obtained by substituting Eq.(11) in Eq.(9):

$$R_t^{(0)}(\mathbf{y}_t, \mathbf{u}_t) = \left(\mathbf{W}\mathbf{z}_t - \mathbf{M}^T\mathbf{u}_t + \varepsilon_t\right)^T (\mathbf{x}_t + \mathbf{u}_t) \tag{15}$$

In addition to this reward that would be obtained in an ideal friction-free world, we have to add (negative) rewards received due to instantaneous market impact and transaction fees[10]. Furthermore, we have to include a negative reward due to risk in a newly created portfolio position at time $t+1$. Similar to [10], we choose a simple quadratic measure of such risk penalty, as the variance of the instantaneous reward (15) conditional on the new state $\mathbf{x}_t + \mathbf{u}_t$, multiplied by the risk aversion parameter λ:

$$R_t^{(risk)}(\mathbf{y}_t, \mathbf{u}_t) = -\lambda \mathrm{Var}_t\left[R_t^{(0)}(\mathbf{y}_t, \mathbf{u}_t)\middle|\, \mathbf{x}_t + \mathbf{u}_t\right] = -\lambda(\mathbf{x}_t + \mathbf{u}_t)^T \Sigma_r (\mathbf{x}_t + \mathbf{u}_t) \tag{16}$$

To specify negative rewards (costs) of an instantaneous market impact and transaction costs, it is convenient to represent each action u_{ti} as a difference of two non-negative action variables $u_{ti}^+, u_{ti}^- \geq 0$:

$$u_{ti} = u_{ti}^+ - u_{ti}^-\,, \quad |u_{ti}| = u_{ti}^+ + u_{ti}^-\,, \quad u_{ti}^+, u_{ti}^- \geq 0 \tag{17}$$

so that $u_{ti} = u_{ti}^+$ if $u_{ti} > 0$ and $u_{ti} = -u_{ti}^-$ if $u_{ti} < 0$. The instantaneous market impact and transaction costs are then given by the following expressions:

$$\begin{aligned} R_t^{(impact)}(\mathbf{y}_t, \mathbf{u}_t) &= -\mathbf{x}_t^T \Gamma^+ \mathbf{u}_t^+ - \mathbf{x}_t^T \Gamma^- \mathbf{u}_t^- - \mathbf{x}_t^T \Upsilon \mathbf{z}_t \\ R_t^{(fee)}(\mathbf{y}_t, \mathbf{u}_t) &= -\nu^{+T}\mathbf{u}_t^+ - \nu^{-T}\mathbf{u}_t^- \end{aligned} \tag{18}$$

Here Γ^+, Γ^-, Υ and ν^+, ν^- are, respectively, matrices-valued and vector-valued parameters that in a simplest case can be parametrized in terms of single scalars multiplied by unit vectors or matrices.

Combining Eqs.(15, (16), (18), we obtain our final specification of a risk- and cost-adjusted instantaneous reward function for the problem of optimal portfolio liquidation:

$$R_t(\mathbf{y}_t, \mathbf{u}_t) = R_t^{(0)}(\mathbf{y}_t, \mathbf{u}_t) + R_t^{(risk)}(\mathbf{y}_t, \mathbf{u}_t) + R_t^{(impact)}(\mathbf{y}_t, \mathbf{u}_t) + R_t^{(fee)}(\mathbf{y}_t, \mathbf{u}_t) \tag{19}$$

The *expected* one-step reward given action $\mathbf{u}_t = \mathbf{u}_t^+ - \mathbf{u}_t^-$ is given by

$$\hat{R}_t(\mathbf{y}_t, \mathbf{u}_t) = \hat{R}_t^{(0)}(\mathbf{y}_t, \mathbf{u}_t) + R_t^{(risk)}(\mathbf{y}_t, \mathbf{u}_t) + R_t^{(impact)}(\mathbf{y}_t, \mathbf{u}_t) + R_t^{(fee)}(\mathbf{y}_t, \mathbf{u}_t) \tag{20}$$

where

$$\hat{R}_t^{(0)}(\mathbf{y}_t, \mathbf{u}_t) = \mathbb{E}_{t,u}\left[R_t^{(0)}(\mathbf{y}_t, \mathbf{u}_t)\right] = \left(\mathbf{W}\mathbf{z}_t - \mathbf{M}^T(\mathbf{u}_t^+ - \mathbf{u}_t^-)\right)^T \left(\mathbf{x}_t + \mathbf{u}_t^+ - \mathbf{u}_t^-\right) \tag{21}$$

where $\mathbb{E}_{t,u}[\cdot] = \mathbb{E}[\cdot|\mathbf{y}_t, \mathbf{u}_t]$ stands for averaging over next-periods realizations of market returns.

Note that the one-step expected reward (20) is a quadratic form of its inputs. We can write it more explicitly using vector notation:

$$\hat{R}(\mathbf{y}_t, \mathbf{a}_t) = \mathbf{y}_t^T \mathbf{R}_{yy} \mathbf{y}_t + \mathbf{a}_t^T \mathbf{R}_{aa} \mathbf{a} + \mathbf{a}_t^T \mathbf{R}_{ay} \mathbf{y}_t + \mathbf{a}_t^T \mathbf{R}_a \tag{22}$$

where

$$\begin{aligned} \mathbf{R}_{aa} &= \begin{bmatrix} -\mathbf{M} - \lambda\Sigma_r & \mathbf{M} + \lambda\Sigma_r \\ \mathbf{M} + \lambda\Sigma_r & -\mathbf{M} - \lambda\Sigma_r \end{bmatrix}, \ \mathbf{R}_{yy} = \begin{bmatrix} -\lambda\Sigma_r & \mathbf{W} - \Upsilon \\ 0 & 0 \end{bmatrix}, \\ \mathbf{R}_{ay} &= \left[\begin{bmatrix} -\mathbf{M} - 2\lambda\Sigma_r - \Gamma^+ \\ \mathbf{M} + 2\lambda\Sigma_r - \Gamma^- \end{bmatrix}, \begin{bmatrix} \mathbf{W} \\ \mathbf{W} \end{bmatrix}\right], \ \mathbf{R}_a = -\begin{bmatrix} \nu^+ \\ \nu^+ \end{bmatrix} \end{aligned} \tag{23}$$

[10] We assume no short sale positions in our setting, and therefore do not include borrowing costs.

3.5 Multi-period portfolio optimization

Multi-period portfolio optimization is equivalently formulated either as maximization of risk- and cost-adjusted returns, as in the Markowitz portfolio model, or as minimization of risk- and cost-adjusted trading costs. The latter specification is usually used in problems of optimal portfolio liquidation.

A multi-period risk- and cost-adjusted reward maximization problem reads

$$\begin{aligned} &\text{maximize } \mathbb{E}_t \left[\sum_{t'=t}^{T-1} \gamma^{t'-t} \hat{R}_{t'}(\mathbf{y}_{t'}, \mathbf{a}_{t'}) \right] \\ &\text{where } \hat{R}_t(\mathbf{y}_t, \mathbf{a}_t) = \mathbf{y}_t^T \mathbf{R}_{yy} \mathbf{y}_t + \mathbf{a}_t^T \mathbf{R}_{aa} \mathbf{a} + \mathbf{a}_t^T \mathbf{R}_{ay} \mathbf{y}_t + \mathbf{a}_t^T \mathbf{R}_a \\ &\text{w.r.t. } \mathbf{a}_t = \begin{pmatrix} \mathbf{u}_t^+ \\ \mathbf{u}_t^- \end{pmatrix} \geq 0, \\ &\text{subject to } \mathbf{x}_t + \mathbf{u}_t^+ - \mathbf{u}_t^- \geq 0 \end{aligned} \tag{24}$$

Here $0 < \gamma \leq 1$ is a discount factor. Note that the sum over future periods $t' = [t, \ldots, T-1]$ does not include the last period $t' = T$, because the last action is fixed by Eq.(10).

An equivalent cost-focused formulation is obtained by flipping the sign of the above problem, and re-phrasing it as minimization of trading costs $\hat{C}_t(\mathbf{y}_t, \mathbf{a}_t) = -\hat{R}_t(\mathbf{y}_t, \mathbf{a}_t)$:

$$\text{minimize } \mathbb{E}_t \left[\sum_{t'=t}^{T-1} \gamma^{t'-t} \hat{C}_{t'}(\mathbf{y}_{t'}, \mathbf{a}_{t'}) \right] \tag{25}$$

$$\text{where } \hat{C}_t(\mathbf{y}_t, \mathbf{a}_t) = -\hat{R}_t(\mathbf{y}_t, \mathbf{a}_t) \tag{26}$$

subject to the same constraints as in (24).

3.6 Dynamic Inverse Portfolio Optimization

When the model dynamics are known (or independently estimated from data), the dynamic portfolio optimization problem of Eq.(24) can be formulated as a problem of Stochastic Optimal Control (SOC), also known as a Dynamic Programming approach. This approach was pursued in Ref. [10] in a general setting of convex portfolio optimization, see also references there on previous work on this topic. In particular, one well-known example is a dynamic mean-variance model of Garleanu and Pedersen [21] with quadratic transaction costs.

We keep a convex multi-period portfolio formulation while adding to it modeling of market impact and external signals, and focusing on a *inverse* optimization problem, rather than a forward optimization problem as in [10]. We can refer to this problem as a Dynamic *Inverse* Portfolio Optimization (DIPO) problem. The word 'dynamic' here means that a learned optimal policy should be *adaptive* to predictors $\mathbf{z}_t$.

In DIPO learning, we assume that an optimal portfolio strategy has been already *found*, perhaps not quite optimally, in the past by an *expert* trader. We assume that we have a record of N different runs of such nearly optimal strategy, each of length T, performed by this expert trader. Following the common conventions of the RL/IRL literature, we can call this data samples expert demonstrations, or expert trajectories. The problem is then to find the optimal execution policy from these data.

We may differentiate between two possible settings for such data-driven DIPO learning that can be encountered in practice. First, in a setting of Reinforcement Learning we have access to historical data consisting of stock market prices, actions taken (i.e. portfolio trades), *and* risk-adjusted *rewards* received upon taking these actions (see below for details). In addition, the data consists of all predictive factors ("alpha-factors") that might be predictive of rewards. The

objective is to learn and improve a policy that was used in the data, so that the new improved policy can be used to generate higher rewards in the future.

The other setting is of Inverse Reinforcement Learning (IRL), where everything is the same as above, except we do *not* observe rewards anymore. The objective is to learn the reward function that leads to the observed behavior, and learn the policy as well.

This is the setting of this paper, where we use an IRL framework to represent *all* traders in the market as *one* market-wise 'expert trader' who is mathematically modeled as a bounded-rational RL agent. A reward function of this agent is learned from market data, plus whatever signals $\mathbf{z}_t$ that are used by the model. The learned parameters include market-implied risk aversion λ, market impact parameters μ_i, weights $\mathbf{W}$ of predictors $\mathbf{z}_t$, and market-implied 'rationality index' β.

Note that if *proprietary* trading data from a particular trader or broker are available, the same framework can be applied to learn a reward function of that particular trader. Such setting might be interesting given that the value of a 'true' risk aversion parameter is often unknown to investors themselves, as they may not base their decisions on a quadratic utility model. When applied to an individual investor, the model developed here may offer a probabilistic model of that *particular* investor, with parameters estimated on trading data of this investor, combined with the market data.

Regarding the policy optimization problem, as rewards are not observed in the IRL setting, this problem is in general both harder and less well-posed in comparison to the RL setting. In particular, unlike RL off-policy methods such as Q-learning that can learn, given enough data, even from data with purely random actions, IRL methods cannot proceed with data with entirely random actions. For IRL to work, data collected should correspond to some good, though not necessarily *optimal* policy. Probabilistic IRL methods are capable of learning when demonstrated data does not always correspond to optimal actions.

While our main focus in this paper is on the IRL setting, we will start below with RL approaches to the problem.

4 Reinforcement Learning of optimal trading

In this section, we will discuss a data-driven Reinforcement Learning approach to multi-period portfolio optimization of Eq.(24). We first introduce stochastic policies and a Bellman equation with stochastic policies, and then consider an entropy-regularized methods for MDP corresponding to Eq.(24).

4.1 Stochastic policy

Note that the multi-period portfolio optimization problem (24) assumes that an optimal policy that determines actions $\mathbf{a}_t$ is a deterministic policy that can also be described as a delta-like probability distribution

$$\pi(\mathbf{a}_t|\mathbf{y}_t) = \delta\left(\mathbf{a}_t - \mathbf{a}_t^\star(\mathbf{y}_t)\right) \tag{27}$$

where the optimal deterministic action $\mathbf{a}_t^\star(\mathbf{y}_t)$ is obtained by maximization of the objective (24) with respect to controls $\mathbf{a}_t$.

But the actual trading data may be sub-optimal, or noisy at times, because of model mis-specifications, market timing lags, human errors etc. Potential presence of such sub-optimal actions in data poses serious challenges, if we try to assume deterministic policy (27) that assumes the the action chosen is *always* an optimal action. This is because such events should

have zero probability under these model assumptions, and thus would produced vanishing path probabilities if observed in data.

Instead of assuming a deterministic policy (27), *stochastic* policies described by *smoothed* distributions $\pi(\mathbf{a}_t|\mathbf{y}_t)$, are more useful for inverse problems such as the problem of inverse portfolio optimization. In this approach, instead of maximization with respect to deterministic policy/action $\mathbf{a}_t$, we re-formulate the problem as maximization over *probability distributions* $\pi(\mathbf{a}_t|\mathbf{y}_t)$:

$$\begin{aligned}
&\text{maximize } \mathbb{E}_{q_\pi}\left[\sum_{t'=t}^{T-1}\gamma^{t'-t}\hat{R}_t(\mathbf{y}_{t'},\mathbf{a}_{t'})\right] \\
&\text{where } \hat{R}(\mathbf{y}_t,\mathbf{a}_t) = \mathbf{y}_t^T\mathbf{R}_{yy}\mathbf{y}_t + \mathbf{a}_t^T\mathbf{R}_{aa}\mathbf{a} + \mathbf{a}_t^T\mathbf{R}_{ay}\mathbf{y}_t + \mathbf{a}_t^T\mathbf{R}_a \\
&\text{w.r.t. } q_\pi(\bar{x},\bar{a}|\mathbf{y}_0) = \pi(\mathbf{a}_0|\mathbf{y}_0)\prod_{t=1}^{T-1}\pi(\mathbf{a}_t|\mathbf{y}_t)P\left(\mathbf{y}_{t+1}|\mathbf{y}_t,\mathbf{a}_t\right) \\
&\text{subject to } \int d\mathbf{a}_t\,\pi\left(\mathbf{a}_t|\mathbf{y}_t\right) = 1
\end{aligned} \tag{28}$$

Here $\mathbb{E}_{q_\pi}[\cdot]$ stands for expectations with respect to path probabilities defined according to the third line in Eqs.(28).

Note that due to inclusion of a quadratic risk penalty in the risk-adjusted return $\hat{R}(\mathbf{x}_t,\mathbf{a}_t)$ the original problem of risk-adjusted return optimization is re-stated in Eq.(28) as maximizing the expected cumulative reward in the standard MDP setting, thus making the problem amenable to a standard risk-neutral approach of MDP models. Such simple risk adjustment based on one-step variance penalties was suggested in a non-financial context by Gosavi [23], and used in a Reinforcement Learning based approach to option pricing in [26, 27].

Another comment that is due here is that a probabilistic approach to actions in portfolio trading appears, on many counts, a more natural way than a formalism based on deterministic policies. Indeed, even in a simplest one-period setting, because the Markowitz-optimal solution for portfolio weights is a function of *estimated* stock means and covariances, they are in fact *random* variables. Yet the probabilistic nature of portfolio optimization is not recognized as such in Markowitz-type single-period or multi-period optimization settings such as (24). A probabilistic portfolio optimization formulation was suggested in a one-period setting by Marshinski *et. al.* [33].

4.2 Reference policy

We assume that we are given a probabilistic *reference* (or *prior*) policy $\pi_0(\mathbf{a}_t|\mathbf{y}_t)$ which should be decided upon prior to attempting the portfolio optimization (28). Such policy can be chosen based on a parametric model, past historic data, etc. We will use a simple Gaussian reference policy

$$\pi_0(\mathbf{a}_t|\mathbf{y}_t) = \frac{1}{\sqrt{(2\pi)^N\left|\Sigma_p\right|}}\exp\left(-\frac{1}{2}\left(\mathbf{a}_t-\hat{\mathbf{a}}(\mathbf{y}_t)\right)^T\Sigma_p^{-1}\left(\mathbf{a}_t-\hat{\mathbf{a}}(\mathbf{y}_t)\right)\right) \tag{29}$$

where $\hat{\mathbf{a}}(\mathbf{y}_t)$ can be a deterministic policy chosen to be a linear function of a state vector $\mathbf{y}_t$:

$$\hat{\mathbf{a}}(\mathbf{y}_t) = \hat{\mathbf{A}}_0 + \hat{\mathbf{A}}_1\mathbf{y}_t \tag{30}$$

A simple choice of parameters in (29) could be to specify them in terms of only two scalars $\hat{a}_0,\ \hat{a}_1$ as follows: $\hat{\mathbf{A}}_0 = \hat{a}_0\mathbf{1}_{|A|}$ and $\hat{\mathbf{A}}_1 = \hat{a}_1\mathbf{1}_{|A|\times|A|}$ where $|A|$ is the the size of vector $\mathbf{a}_t$, $\mathbf{1}_A$ and $\mathbf{1}_{A\times A}$ are, respectively, a vector and matrix made of ones. The scalars $\hat{a}_0$ and $\hat{a}_1$ would then serve as hyper-parameters in our setting. Similarly, covariance matrix Σ_p for the prior policy can be taken to be a simple matrix with constant correlations ρ_p and constant variances σ_p.

As will be shown below, an *optimal* policy has the same Gaussian form as the prior policy (29), with updated parameters $\hat{\mathbf{A}}_0$, $\hat{\mathbf{A}}_1$ and Σ_p. These updates will be computed iteratively starting with their initial values defining the prior (29). Respectively, updates at iteration k will be denoted by upper subscripts, e.g. $\hat{\mathbf{A}}_0^{(k)}$, $\hat{\mathbf{A}}_1^{(k)}$.

Furthermore, it turns out that a linear dependence on $\mathbf{y}_t$ at iteration k, driven by the value of $\hat{\mathbf{A}}_1^{(k)}$ arises even if we set $\hat{\mathbf{A}}_1 = \hat{\mathbf{A}}_1^{(0)} = 0$ in the prior (29). Such choice of a state-independent prior $\pi_0(\mathbf{a}_t|\mathbf{y}_t) = \pi_0(\mathbf{a}_t)$, although not very critical, reduces the number of free parameters in the model by two, as well as simplifies some of the analyses below, and hence will be assumed going forward. It also makes it unnecessary to specify the value of $\bar{\mathbf{y}}_t$ in the prior (29) (equivalently, we can initialize it at zero). The final set of hyper-parameters defining the prior (29) therefore includes only three values of $\hat{a}_0$, ρ_a, Σ_p.

4.3 Bellman Optimality Equation

Let

$$V_t^\star(\mathbf{y}_t) = \max_{\pi(\cdot|y)} \mathbb{E}\left[\sum_{t'=t}^{T-1} \gamma^{t'-t} \hat{R}_{t'}(\mathbf{y}_{t'}, \mathbf{a}_{t'}) \middle| \mathbf{y}_t\right] \tag{31}$$

The optimal state value function $V_t^\star(\mathbf{x}_t)$ satisfies the Bellman optimality equation (see e.g. [49])

$$V_t^\star(\mathbf{y}_t) = \max_{\mathbf{a}_t} \hat{R}_t(\mathbf{y}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}_t}\left[V_{t+1}^\star(\mathbf{y}_{t+1})\right] \tag{32}$$

The optimal policy $\pi^\star$ can be obtained from $V^\star$ as follows:

$$\pi_t^\star(\mathbf{a}_t|\mathbf{y}_t) = \arg\max_{\mathbf{a}_t} \hat{R}_t(\mathbf{y}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}_t}\left[V_{t+1}^\star(\mathbf{y}_{t+1})\right] \tag{33}$$

The goal of Reinforcement Learning (RL) is to solve the Bellman optimality equation based on samples of data. Assuming that an optimal value function is found by means of RL, solving for the optimal policy $\pi^\star$ takes another optimization problem as formulated in Eq.(33).

4.4 Entropy-regularized Bellman optimality equation

Following [11], we start with reformulating the Bellman optimality equation using a Fenchel-type representation:

$$V_t^\star(\mathbf{y}_t) = \max_{\pi(\cdot|y)\in\mathcal{P}} \sum_{\mathbf{a}_t\in\mathcal{A}_t} \pi(\mathbf{a}_t|\mathbf{y}_t)\left(\hat{R}_t(\mathbf{y}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}_t}\left[V_{t+1}^\star(\mathbf{y}_{t+1})\right]\right) \tag{34}$$

Here $\mathcal{P} = \left\{\pi : \pi \geq 0, \mathbf{1}^T\pi = 1\right\}$ stands for a set of all valid distributions. Eq.(34) is equivalent to the original Bellman optimality equation (31), because for any $x \in \mathbb{R}^n$, we have $\max_{i\in\{1,\ldots,n\}} x_i = \max_{\pi\geq 0, ||\pi||\leq 1} \pi^T x$. Note that while we use discrete notations for simplicity of presentation, all formulas below can be equivalently expressed in continuous notations by replacing sums by integrals. For brevity, we will denote the expectation $\mathbb{E}_{\mathbf{y}_{t+1}|\mathbf{y}_t,\mathbf{a}_t}[\cdot]$ as $\mathbb{E}_{t,\mathbf{a}}[\cdot]$ in what follows.

The one-step *information cost* of a learned policy $\pi(\mathbf{a}_t|\mathbf{y}_t)$ relative to a reference policy $\pi_0(\mathbf{a}_t|\mathbf{y}_t)$ is defined as follows [18]:

$$g^\pi(\mathbf{y}, \mathbf{a}) = \log \frac{\pi(\mathbf{a}_t|\mathbf{y}_t)}{\pi_0(\mathbf{a}_t|\mathbf{y}_t)} \tag{35}$$

Its expectation with respect to policy π is the Kullback-Leibler (KL) divergence of $\pi(\cdot|\mathbf{y}_t)$ and $\pi_0(\cdot|\mathbf{y}_t)$:

$$\mathbb{E}_\pi\left[g^\pi(\mathbf{y},\mathbf{a})|\,\mathbf{y}_t\right] = KL[\pi||\pi_0](\mathbf{y}_t) \equiv \sum_{\mathbf{a}_t}\pi(\mathbf{a}_t|\mathbf{y}_t)\log\frac{\pi(\mathbf{a}_t|\mathbf{y}_t)}{\pi_0(\mathbf{a}_t|\mathbf{y}_t)} \tag{36}$$

The total discounted information cost for a trajectory is defined as follows:

$$I^\pi(\mathbf{y}) = \sum_{t'=t}^{T}\gamma^{t'-t}\mathbb{E}\left[g^\pi(\mathbf{y}_{t'},\mathbf{a}_{t'})|\,\mathbf{y}_t=\mathbf{y}\right] \tag{37}$$

The *free energy* function $F_t^\pi(\mathbf{y}_t)$ is defined as the value function (34) augmented by the information cost penalty (37):

$$\begin{aligned} F_t^\pi(\mathbf{y}_t) &= V_t^\pi(\mathbf{y}_t) - \frac{1}{\beta}I^\pi(\mathbf{y}_t) \\ &= \sum_{t'=t}^{T}\gamma^{t'-t}\mathbb{E}\left[\hat{R}_{t'}(\mathbf{y}_{t'},\mathbf{a}_{t'}) - \frac{1}{\beta}g^\pi(\mathbf{y}_{t'},\mathbf{a}_{t'})\right] \end{aligned} \tag{38}$$

Note that β in Eq.(38) serves as the "inverse temperature" parameter that controls a trade-off between reward optimization and proximity to the reference policy, see below. The free energy $F_t^\pi(\mathbf{y}_t)$ is the entropy-regularized value function, where the amount of regularization can be tuned to better cope with noise in data[11]. The reference policy π_0 provides a "guiding hand" in the stochastic policy optimization process that we describe next.

A Bellman equation for the free energy function $F_t^\pi(\mathbf{y}_t)$ is obtained from (38):

$$F_t^\pi(\mathbf{y}_t) = \mathbb{E}_{\mathbf{a}|y}\left[\hat{R}_t(\mathbf{y}_t,\mathbf{a}_t) - \frac{1}{\beta}g^\pi(\mathbf{y}_t,\mathbf{a}_t) + \gamma\mathbb{E}_{t,\mathbf{a}}\left[F_{t+1}^\pi(\mathbf{y}_{t+1})\right]\right] \tag{39}$$

For a finite-horizon setting, Eq.(39) should be supplemented by a terminal condition

$$F_T^\pi(\mathbf{y}_T) = \hat{R}_T(\mathbf{y}_T,\mathbf{a}_T)\Big|_{\mathbf{a}_T=-\mathbf{u}_{T-1}} \tag{40}$$

(see Eq.(10)). Eq.(39) can be viewed as a soft probabilistic relaxation of the Bellman optimality equation for the value function, with the KL information cost penalty (36) as a regularization controlled by the inverse temperature β. In addition to such regularized value function (free energy), we will next introduce an entropy regularized Q-function.

4.5 G-function: an entropy-regularized Q-function

Similarly to the action-value function, we define the state-action free energy function $G^\pi(\mathbf{x},\mathbf{a})$ as [18]

$$\begin{aligned} G_t^\pi(\mathbf{y}_t,\mathbf{a}_t) &= \hat{R}_t(\mathbf{y}_t,\mathbf{a}_t) + \gamma\mathbb{E}\left[F_{t+1}^\pi(\mathbf{y}_{t+1})\big|\,\mathbf{y}_t,\mathbf{a}_t\right] \\ &= \hat{R}_t(\mathbf{y}_t,\mathbf{a}_t) + \gamma\mathbb{E}_{t,\mathbf{a}}\left[\sum_{t'=t+1}^{T}\gamma^{t'-t-1}\left(\hat{R}_{t'}(\mathbf{y}_{t'},\mathbf{a}_{t'}) - \frac{1}{\beta}g^\pi(\mathbf{y}_{t'},\mathbf{a}_{t'})\right)\right] \\ &= \mathbb{E}_{t,\mathbf{a}}\left[\sum_{t'=t}^{T}\gamma^{t'-t}\left(\hat{R}_{t'}(\mathbf{y}_{t'},\mathbf{a}_{t'}) - \frac{1}{\beta}g^\pi(\mathbf{y}_{t'},\mathbf{a}_{t'})\right)\right] \end{aligned} \tag{41}$$

[11] Note that in physics, as well as in the free-energy principle literature [19, 39], free energy is defined with a negative sign relative to Eq.(38). This difference is purely a matter of a sign convention, as maximization of Eq.(38) can be re-stated as minimization of its negative. With our sign convention for the free energy function, we follow Reinforcement Learning and Information Theory literature [40, 52, 42, 30].

where in the last equation we used the fact that the first action $\mathbf{a}_t$ in the G-function is fixed, and hence $g^\pi(\mathbf{y}_t, \mathbf{a}_t) = 0$ when we condition on $\mathbf{a}_t = \mathbf{a}$.

If we now compare this expression with Eq.(38), we obtain the relation between the G-function and the free energy $F_t^\pi(\mathbf{y}_t)$:

$$F_t^\pi(\mathbf{y}_t) = \sum_{\mathbf{a}_t} \pi(\mathbf{a}_t|\mathbf{y}_t) \left[G_t^\pi(\mathbf{y}_t, \mathbf{a}_t) - \frac{1}{\beta} \log \frac{\pi(\mathbf{a}_t|\mathbf{y}_t)}{\pi_0(\mathbf{a}_t|\mathbf{y}_t)} \right] \tag{42}$$

This functional is maximized by the following distribution $\pi(\mathbf{a}_t|\mathbf{y}_t)$:

$$\begin{aligned} \pi(\mathbf{a}_t|\mathbf{y}_t) &= \frac{1}{Z_t} \pi_0(\mathbf{a}_t|\mathbf{y}_t) e^{\beta G_t^\pi(\mathbf{y}_t, \mathbf{a}_t)} \\ Z_t &= \sum_{\mathbf{a}_t} \pi_0(\mathbf{a}_t|\mathbf{y}_t) e^{\beta G_t^\pi(\mathbf{y}_t, \mathbf{a}_t)} \end{aligned} \tag{43}$$

The free energy (42) evaluated at the optimal solution (43) becomes

$$F_t^\pi(\mathbf{y}_t) = \frac{1}{\beta} \log Z_t = \frac{1}{\beta} \log \sum_{\mathbf{a}_t} \pi_0(\mathbf{a}_t|\mathbf{y}_t) e^{\beta G_t^\pi(\mathbf{y}_t, \mathbf{a}_t)} \tag{44}$$

Using Eq.(44), the optimal action policy (43) can be written as follows :

$$\pi(\mathbf{a}_t|\mathbf{y}_t) = \pi_0(\mathbf{a}_t|\mathbf{y}_t) e^{\beta (G_t^\pi(\mathbf{y}_t, \mathbf{a}_t) - F_t^\pi(\mathbf{y}_t))} \tag{45}$$

Eqs.(44), (45), along with the first form of Eq.(41) repeated here for convenience:

$$G_t^\pi(\mathbf{y}_t, \mathbf{a}_t) = \hat{R}_t(\mathbf{y}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}} \left[F_{t+1}^\pi(\mathbf{y}_{t+1}) \middle| \mathbf{y}_t, \mathbf{a}_t \right] \tag{46}$$

constitute a system of equations that should be solved self-consistently by backward recursion for $t = T-1, \ldots, 0$, with terminal conditions

$$\begin{aligned} G_T^\pi(\mathbf{y}_T, \mathbf{a}_T) &= \hat{R}_T(\mathbf{y}_T, \mathbf{a}_T) \\ F_T^\pi(\mathbf{y}_T) &= G_T^\pi(\mathbf{y}_T, \mathbf{a}_T) = \hat{R}_T(\mathbf{y}_T, \mathbf{a}_T) \end{aligned} \tag{47}$$

The self-consistent scheme of Eqs.(44, 45, 46) [18] can be used in both the RL setting, when rewards are *observed*, and in the IRL setting when they are *not*. Before proceeding with these methods, we want to digress on an alternative interpretation of entropy regularization in Eq.(38), that can be useful for clarifying the approach of this paper.

4.6 Adversarial interpretation of entropy regularization

A useful alternative interpretation of the entropy regularization term in Eq.(38) can be suggested using its representation as a Legendre-Fenchel transform of another function [41]:

$$-\frac{1}{\beta} \sum_{\mathbf{a}_t} \pi(\mathbf{a}_t|\mathbf{y}_t) \log \frac{\pi(\mathbf{a}_t|\mathbf{y}_t)}{\pi_0(\mathbf{a}_t|\mathbf{y}_t)} = \min_{C(\mathbf{a}_t, \mathbf{y}_t)} \sum_{\mathbf{a}_t} \left(-\pi(\mathbf{a}_t|\mathbf{y}_t) \left(1 + C(\mathbf{a}_t, \mathbf{y}_t)\right) + \pi_0(\mathbf{a}_t|\mathbf{y}_t) e^{\beta C(\mathbf{a}_t, \mathbf{y}_t)} \right) \tag{48}$$

where $C(\mathbf{a}_t, \mathbf{y}_t)$ is an arbitrary function. Eq.(48) can be verified by direct minimization of the right-hand side with respect to $C(\mathbf{a}_t, \mathbf{y}_t)$.

Using this representation of the KL term, the free energy maximization problem (42) can be re-stated as a max-min problem

$$F_t^\star(\mathbf{y}_t) = \max_\pi \min_C \sum_{\mathbf{a}_t} \pi(\mathbf{a}_t|\mathbf{y}_t) \left[G_t^\pi(\mathbf{y}_t, \mathbf{a}_t) - C(\mathbf{a}_t, \mathbf{y}_t) - 1\right] + \pi_0(\mathbf{a}_t|\mathbf{y}_t) e^{\beta C(\mathbf{a}_t, \mathbf{y}_t)} \tag{49}$$

The imaginary adversary's optimal cost obtained from (49) is

$$C^\star(\mathbf{a}_t, \mathbf{y}_t) = \frac{1}{\beta} \log \frac{\pi(\mathbf{a}_t|\mathbf{y}_t)}{\pi_0(\mathbf{a}_t|\mathbf{y}_t)} \tag{50}$$

Similarly to [41], one can check that this produces an *indifference* solution for the imaginary game between the agent and its adversarial environment where the total sum of the optimal G-function and the optimal adversarial cost (50) is constant: $G_t^\star(\mathbf{y}_t, \mathbf{a}_t) + C^\star(\mathbf{a}_t, \mathbf{y}_t) = \text{const}$, which means that the game of the original agent and its adversary is in a Nash equilibrium.

Therefore, portfolio optimization in a stochastic environment by a single agent that represents a bounded-rational component of the market as a whole, as is done in our approach using the entropy-regularized free energy, is mathematically equivalent to studying a Nash equilibrium in a two-party game of our our agent with an adversarial counter-party with an exponential budget given by the last term in Eq.(49).

4.7 G-learning and F-learning

In the RL setting when rewards are observed, the system Eqs.(44, 45, 46) can be reduced to one non-linear equation. Substituting the augmented free energy (44) into Eq.(41), we obtain

$$G_t^\pi(\mathbf{y}, \mathbf{a}) = \hat{R}(\mathbf{y}_t, \mathbf{a}_t) + \mathbb{E}_{t,\mathbf{a}} \left[\frac{\gamma}{\beta} \log \sum_{\mathbf{a}_{t+1}} \pi_0(\mathbf{a}_{t+1}|\mathbf{y}_{t+1}) e^{\beta G_{t+1}^\pi(\mathbf{y}_{t+1}, \mathbf{a}_{t+1})} \right] \tag{51}$$

This equation provides a soft relaxation of the Bellman optimality equation for the action-value Q-function, with the G-function defined in Eq.(41) being an entropy-regularized Q-function [18]. The "inverse-temperature" parameter β in Eq.(51) determines the strength of entropy regularization. In particular, if we take $\beta \to \infty$, we recover the original Bellman optimality equation for the Q-funciton. Because the last term in (51) approximates the $\max(\cdot)$ function when β is large but finite, Eq.(51) is known in the literature as soft Q-learning.

For finite values $\beta < \infty$, in a setting of Reinforcement Learning with observed rewards, Eq.(51) can be used to specify *G-learning* [18]: an off-policy time-difference (TD) algorithm that generalizes Q-learning to noisy environments where an entropy-based regularization can be needed. The G-learning algorithm of Ref. [18] was specified in a tabulated setting where both the state and action space are finite. In our case, we deal with high-dimensional state and action spaces, and in addition, we do not observe rewards, therefore we are in a setting of Inverse Reinforcement Learning.

Another possible approach is to bypass the G-function (i.e. the entropy-regulated Q-function) altogether, and proceed with the Bellman optimality equation for the free energy F-function (38). In this case, we have a pair of equations for $F_t^\pi(\mathbf{y}_t)$ and $\pi(\mathbf{a}_t|\mathbf{y}_t)$:

$$\begin{aligned} F_t^\pi(\mathbf{y}_t) &= \mathbb{E}_{\mathbf{a}|x} \left[\hat{R}(\mathbf{y}_t, \mathbf{a}_t) - \frac{1}{\beta} g^\pi(\mathbf{y}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}} \left[F_{t+1}^\pi(\mathbf{y}_{t+1}) \right] \right] \\ \pi(\mathbf{a}_t|\mathbf{y}_t) &= \frac{1}{Z_t} \pi_0(\mathbf{a}_t|\mathbf{y}_t) e^{\hat{R}(\mathbf{y}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}}\left[F_{t+1}^\pi(\mathbf{y}_{t+1})\right]} \end{aligned} \tag{52}$$

Here the first equation is the Bellman equation (39) for the F-function, and the second equation is obtained by substitution of Eq.(46) into Eq.(43). Also note that the normalization constant Z_t in Eq.(52) is in general different from the normalization constant in Eq.(43).

Eq.(52) shows that one-step rewards $\hat{R}(\mathbf{y}_t, \mathbf{a}_t)$ do *not* form on their own an alternative specification of single-step action probabilities $\pi(\mathbf{a}_t|\mathbf{y}_t)$. Rather, a specification of the sum $\hat{R}(\mathbf{y}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}} \left[F^{\pi}_{t+1}(\mathbf{y}_{t+1}) \right]$ is required [42]. However, in a special case when dynamics are *linear* and rewards $\hat{R}(\mathbf{y}_t, \mathbf{a}_t)$ are *quadratic*, the term $\mathbb{E}_{t,\mathbf{a}} \left[F^{\pi}_{t+1}(\mathbf{y}_{t+1}) \right]$ has the same parametric form as the time-t reward $\hat{R}(\mathbf{y}_t, \mathbf{a}_t)$, therefore addition of this term amounts to a 'renormalization' of parameters of the one-step reward function (see below). Therefore, if the only objective of IRL is to learn a policy from data via modeling a reward function, a model can directly learn these 'renormalized' parameters from data. Splitting these values into a current-reward and expected future-reward part would be unnecessary in this case, reducing the problem of finding an optimal policy in IRL to a standard Maximum Likelihood estimation. Such approach was considered e.g. in [28] in a different context.

5 Inverse Reinforcement Learning of optimal trading

In this section, we will simultaneously analyze two settings for our model: (i) a single investor IRL, and (ii) a market portfolio IRL. The main difference between these two cases is that while in the first case actions of an agent are observable, in the second case they are *not* directly observable, only their impact on market prices is observed.

A second difference has to do with a planning horizon in the model. For a single investor case, we have a finite-horizon MDP problem where a task starts a given initial time t_0 and ends in T steps at a specific time $t_0 + T$. On the contrary, for the market portfolio IRL we do not have a well defined notion of a starting time t_0 and an end time T. The only uncontroversial time-like parameter is the current time t.

A reasonable choice would be to get rid of an alleged time non-stationarity in a time homogeneous problem by setting $t_0 = t$ (which means we start our task now), and to set T to infinity. The latter means changing the problem to a problem of an *infinite-horizon* IRL.

On the other hand, as we will show below, computational algorithms for these two cases have many common or similar elements. In particular, an infinite-horizon setting can be numerically approximated by a fixed time horizon, while unobserved actions can be viewed as hidden variables that now become a part of inference of the model.

This implies that up to a certain point, inference of a market optimal portfolio and a single investor portfolio should involve many common elements. In our setting, as our bounded-rational market-agent is a *sum* of all individual investors, state variables in these two formulations are linked in a very explicit way: what was a dollar amount of a single investor's investment in a given stock becomes a total market capitalization for this stock in the market portfolio case.

Therefore, additivity of total individual investor's portfolios and actions into a single market-wise portfolio and a single action of a bounded-rational market-agent is built-in in our model by construction. This implies that the case of a market portfolio inference can be viewed as a generalization of a single investor case[12].

In this section, we will present a general solution to the problem of inference of optimal investment strategy from data made of observation of states, that works for both cases of a

[12]It also opens a way to build a model of influential 'market movers" in a top-down manner by a probabilistic dissection ("thinning") of a market-optimal portfolio into sub-portfolios of individual major investors. We leave this for a future research.

single investor and a market portfolio. This solution is based on a variational EM algorithm, and it can be used to find the original model parameters Θ. As will be shown in Sect. 6, for a specific case of a market portfolio, in addition to such general approach, our model can also be estimated in an alternative and simpler way, by re-formulating it as an *econometric* model of stock returns. Our presentation in the present section covers both cases of a single investor and market portfolio simultaneously when possible, and give separate analyses when it is not.

5.1 Likelihood functions

We first consider the case of observable actions. Data in this case includes a set of D trajectories ζ_i where $i = 1, \dots D$ of state-action pairs $(\mathbf{y}_t, \mathbf{a}_t)$ where trajectory i starts at some time t_{0i} and runs until time T_i.

We consider a single trajectory ζ where we set the start time $t = 0$ and the end time T. As individual trajectories are considered independent, they will enter additively in the final log-likelihood of the problem. We assume that dynamics are Markov in the pair $(\mathbf{y}_t, \mathbf{a}_t)$.

The probability of complete data for trajectory ζ is

$$P_c\left(\mathbf{y}, \mathbf{a}|\Theta\right) = p_\theta(\mathbf{y}_0) \prod_{t=0}^{T-1} \pi_\theta(\mathbf{a}_t|\mathbf{y}_t) p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{a}_t\right) \tag{53}$$

Here $p(\mathbf{y}_0)$ is a marginal probability of $\mathbf{y}_t$ at the start of the i-th demonstration, and $p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t\mathbf{a}_t\right)$ is a probability of a new state $\mathbf{y}_{t+1}$ conditional on the previous state $\mathbf{y}_t$ and action $\mathbf{a}_t$ taken at this step. Note that the first action $\mathbf{a}_0$ is fixed, therefore we have $\pi_\theta(\mathbf{a}_0|\mathbf{y}_0) = 1$. Also note that in our model-based IRL setting, both the action policy $\pi_\theta(\cdot|\mathbf{y}_t)$ and transition probability $p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t\mathbf{a}_t\right)$ depend on the same set of parameters. The joint distribution $p_\theta(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t) = \pi_\theta(\mathbf{a}_t|\mathbf{y}_t) p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t\mathbf{a}_t\right)$ is a generative model in our framework.

For a *complete data* (i.e. when both $\mathbf{y}_t$ and $\mathbf{a}_t$ are observable), we obtain the following log-likelihood

$$L_c(\theta) = \log P_c\left(\mathbf{y}, \mathbf{a}|\Theta\right) = \log p_\theta(\mathbf{y}_0) + \sum_{t\in\zeta} \left(\log \pi_\theta(\mathbf{a}_t|\mathbf{y}_t) + \log p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{a}_t\right)\right) \tag{54}$$

where $\mathbf{y}_t$ and $\mathbf{a}_t$ stand for values observed in data. Given some simple parametric forms for the policy and transition probability functions, maximization of such complete data log-likelihood is rather straightforward. Such inference problem with complete data corresponds to a single investor IRL in our model.

A different situation arises for IRL of the market portfolio. In this case, actions $\mathbf{a}_t$ of the agent are no longer observable. Respectively, we treat them as *hidden* variables and integrate over all values of $\mathbf{a}_t$ in the product over t in Eq.(53). This produces the *expected* complete log-likelihood of data

$$L_e(\theta) = \log p_\theta(\mathbf{y}_0) + \sum_{t=0}^{T-1} \log \int d\mathbf{a}_t \pi_\theta(\mathbf{a}_t|\mathbf{y}_t) p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{a}_t\right) \tag{55}$$

As the log-likelihood function involves an integral over $\mathbf{a}_t$, it is in general intractable in high dimensional action spaces. Therefore, we will next address an approximate approach to evaluation of log-likelihood (55). Furthermore, as Eq.(55) is additive in time steps, in what follows we focus on practical ways to compute the integral over $\mathbf{a}_t$ in a single term entering the sum over t in (55).

5.2 EM algorithm

Expectation Maximization (EM) algorithm is a powerful method for estimating parameters of models with incomplete observations and/or hidden variables. In our Eq.(55), the role of hidden variables is played by actions $\mathbf{a}_t$. In addition, we might introduce additional hidden variables for tractability of a resulting approximate likelihood.

Let $q(\mathbf{a}_t|\mathbf{y})$ be some distribution for actions $\mathbf{a}_t$ that can depend on the data $\mathbf{y} = (\mathbf{y}_t, \mathbf{y}_{t+1})$. We can use it to write the expected one-step log-likelihood L_t for time step $[t, t+1]$ as follows:

$$\begin{aligned} L_t(\theta) &\equiv \log \int d\mathbf{a}_t \, p_\theta \left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t\right) = \log \int d\mathbf{a}_t \, q(\mathbf{a}_t|\mathbf{y}) \frac{p_\theta \left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t\right)}{q(\mathbf{a}_t|\mathbf{y})} \\ &\geq \int d\mathbf{a}_t \, q(\mathbf{a}_t|\mathbf{y}) \log \frac{p_\theta \left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t\right)}{q(\mathbf{a}_t|\mathbf{y})} \end{aligned} \tag{56}$$

where in the second line we used Jensen's inequality. This produces the following low bound for expected log-likelihood of data:

$$\begin{aligned} \mathcal{F}(q, \theta) &\equiv \int d\mathbf{a}_t \, q(\mathbf{a}_t|\mathbf{y}) \log \frac{p_\theta \left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t\right)}{q(\mathbf{a}_t|\mathbf{y})} = \mathbb{E}_q \left[\log p_\theta \left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t\right)\right] + H\left[q\right] \\ &= -KL\left[q(\mathbf{a}_t|\mathbf{y}) || p_\theta \left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t\right)\right] \end{aligned} \tag{57}$$

where $H\left[q\right] = -\int d\mathbf{a}_t \, q(\mathbf{a}_t|\mathbf{y}) \log q(\mathbf{a}_t|\mathbf{y})$ is the entropy of distribution $q(\mathbf{a}_t|\mathbf{y})$. The low bound (57) can be interpreted as a free energy with the 'energy function' $\log p_\theta \left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t\right)$ [38].

The classical EM algorithm [12] amounts to iterative maximization of the free energy (57) with respect to the distribution q and model parameters θ:

$$\begin{aligned} &\textbf{E step: } q^{(k+1)} = \underset{\mathbf{q}}{\operatorname{argmax}} \, \mathcal{F}(q, \theta^{(k)}) \\ &\textbf{M step: } \theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}} \, \mathcal{F}(q^{(k+1)}, \theta) \end{aligned} \tag{58}$$

Note that the E-step can formally be done analytically by noting that the last form of the free energy $\mathcal{F}(q, \theta)$ in Eq.(57) indicates that its maximum as a function of q is attained when $q(\mathbf{a}_t|\mathbf{y}) = C p_\theta \left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t\right)$, where C is a normalization constant, which should be equal to $1/p_\theta \left(\mathbf{y}_{t+1}|\mathbf{y}_t\right)$ to have the right normalization of $q(\mathbf{a}_t|\mathbf{y})$. Together this produces the following analytical result for the E-step:

$$q^{(k+1)} = \frac{p_\theta \left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t, \theta^{(k-1)}\right)}{p_\theta \left(\mathbf{y}_{t+1}|\mathbf{y}_t, \theta^{(k-1)}\right)} = p_\theta \left(\mathbf{a}_t|\mathbf{y}_{t+1}, \mathbf{y}_t, \theta^{(k)}\right) \tag{59}$$

so that q for the k-th step is just the posterior distribution of $\mathbf{a}_t$ computed with the model parameters from the previous iteration. The M-step in Eq.(58) then amounts to maximization of the expectation of the 'energy' $\log p_\theta \left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t\right)$ in parameters θ. This procedure guarantees a monotonous convergence to a local maximum of the free energy (57) [12, 38].

5.3 Variational EM

As the M-step of the classical EM algorithm is intractable in our setting, we use the variational EM method where instead of a non-parametric specification of the approximating distribution q leading to a non-parametric optimal solution for the E-step, we use a model-based specification $q_w(\cdot)$ parametrized by a set of 'recognition model' parameters ω. The E-step then amounts to

maximization with respect to parameters ω, while the M-step is performed with an expectation defined by the distribution $q^{(k+1)}(\cdot)$.

A variational EM algorithm thus iteratively updates the recognition model parameters ω and the generative model parameters θ:

$$\begin{aligned} &\textbf{E step: } \omega^{(k+1)} = \underset{\omega}{\operatorname{argmax}}\, \mathcal{F}(\omega, \theta^{(k)}) \\ &\textbf{M step: } \theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{F}(\omega^{(k+1)}, \theta) \end{aligned} \tag{60}$$

While a variational version of the EM algorithm does not guarantees a monotonous increase of a log-likelihood at each step, it guarantees that the log-likelihood is non-decreasing (i.e. it either increases or stays constant) at each iteration.

To produce a practical computational scheme, we consider the following specification of a variational distribution $q_\omega(\cdot)$ as a joint distribution of *four* hidden variables $\mathbf{a}_t$, $\bar{\mathbf{a}}_t$, $\bar{\mathbf{y}}_t$, $\bar{\mathbf{y}}_{t+1}$:

$$q_\omega(\mathbf{a}_t|\mathbf{y}) = \int d\bar{\mathbf{a}}_t d\bar{\mathbf{y}}_t d\bar{\mathbf{y}}_{t+1} q_\omega(\mathbf{a}_t, \bar{\mathbf{a}}_t, \bar{\mathbf{y}}_t, \bar{\mathbf{y}}_{t+1}|\mathbf{y}) = \int d\bar{\mathbf{a}}_t d\bar{\mathbf{y}} q_{\bar{a}\bar{y}}(\bar{\mathbf{a}}_t, \bar{\mathbf{y}}|\mathbf{y}, \omega) q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t, \omega) \tag{61}$$

where $\mathbf{y} = (\mathbf{y}_t, \mathbf{y}_{t+1})$ and $\bar{\mathbf{y}} = (\bar{\mathbf{y}}_t, \bar{\mathbf{y}}_{t+1})$. The hidden variables $\bar{\mathbf{a}}_t, \bar{\mathbf{y}}$ will serve below for linearization of dynamics, similar to the Robust Controllable Embedding (RCE) method of [5].

Using this distribution in Eq.(57), we obtain the following variational EM bound on the log-likelihood of observed data:

$$\begin{aligned} \mathcal{F}(\omega, \theta) \;&= \int d\bar{\mathbf{a}}_t d\bar{\mathbf{y}}\, q_{\bar{a}\bar{y}}(\bar{\mathbf{a}}_t, \bar{\mathbf{y}}|\mathbf{y}, \omega) \int d\mathbf{a}_t\, q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t, \omega) \log \frac{p_\theta\left(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t\right)}{q_\omega(\mathbf{a}_t, \bar{\mathbf{a}}_t, \bar{\mathbf{y}}|\mathbf{y})} \\ &\equiv \int d\bar{\mathbf{a}}_t d\bar{\mathbf{y}}\, q_{\bar{a}\bar{y}}(\bar{\mathbf{a}}_t, \bar{\mathbf{y}}|\mathbf{y}, \omega) \mathcal{F}_a(\omega, \theta, \bar{\mathbf{a}}_t, \bar{\mathbf{y}}) \end{aligned} \tag{62}$$

where $\mathcal{F}_a(\omega, \theta, \bar{\mathbf{a}}_t, \bar{\mathbf{y}})$ is a conditional variational free energy :

$$\mathcal{F}_a(\omega, \theta, \bar{\mathbf{a}}_t, \bar{\mathbf{y}}) = \int d\mathbf{a}_t q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t, \omega) \log \frac{\pi_\theta\left(\mathbf{a}_t|\mathbf{y}_t, \bar{\mathbf{a}}_t, \bar{\mathbf{y}}_t\right) p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{a}_t, \bar{\mathbf{a}}_t, \bar{\mathbf{y}}\right)}{q_\omega(\mathbf{a}_t, \bar{\mathbf{a}}_t, \bar{\mathbf{y}}|\mathbf{y})} \tag{63}$$

where $q_\omega(\mathbf{a}_t|\mathbf{y})$ in the logarithm is computed as per Eq.(61). Eqs.(62) and (63) thus give a variational low bound on the likelihood of data for inference of a market portfolio, while for the case of an individual investor, we have to omit the inner integral over $\mathbf{a}_t$ in Eq.(62).

Note that in Eq.(63) we explicitly introduced the hidden variables into the generative model $p_\theta(\mathbf{y}_{t+1}, \mathbf{a}_t|\mathbf{y}_t)$. As will be shown below, these hidden variables are introduced to make two calculations involved in Eq.(63) tractable: computing the integral in (63), and computing the policy π_θ that this integral depends on.

These two tasks are clearly sequential. We will first use conditioning on hidden variables to find a tractable representation of the action policy π_θ, and then use this representation to compute the integral over $\mathbf{a}_t$. Eq.(63) suggests that if the distribution $q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t, \omega)$ is sharply peaked around $\mathbf{a}_t = \bar{\mathbf{a}}_t$, then the conditional free energy $\mathcal{F}_a(\omega, \theta, \bar{\mathbf{a}}_t, \bar{\mathbf{y}})$ can be computed using a saddle-point (Laplace) approximation. The remaining integral in Eq.(62) over the conditioning hidden variables $\bar{\mathbf{a}}_t$, $\bar{\mathbf{y}}_t$, $\bar{\mathbf{y}}_{t+1}$ can then be computed using another saddle point approximation. This scheme will be presented in details below after we specify the variational policy distribution q_w and the generative model p_θ.

5.4 Variational distribution q_w

Our variational model q_ω is defined as follows:

$$\begin{aligned} q_\omega\left(\mathbf{a}_t, \bar{\mathbf{a}}_t, \bar{\mathbf{y}}|\mathbf{y}\right) &= q_{\bar{a}\bar{y}}\left(\bar{\mathbf{a}}_t, \bar{\mathbf{y}}|\mathbf{y}\right) q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t, \omega) \\ &= q_\phi\left(\bar{\mathbf{y}}_{t+1}|\mathbf{y}_{t+1}\right) q_\varphi\left(\bar{\mathbf{y}}_t|\mathbf{y}_t, \bar{\mathbf{y}}_{t+1}\right) q_{\bar{a}}(\bar{\mathbf{a}}_t|\mathbf{y}_t, \omega) q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t, \omega) \end{aligned} \tag{64}$$

Here q_ϕ and q_φ are variational forward and backward encoders, respectively [5]. As we assume time homogeneity, a functional form of the encoder $q_\phi\left(\bar{\mathbf{y}}_{t+1}|\mathbf{y}_{t+1}\right)$ should be the same as of $q_\phi\left(\bar{\mathbf{y}}_t|\mathbf{y}_t\right)$.

We use Gaussian specifications for four marginals of the variational policy q_w:

$$\begin{aligned} &q_{\bar{a}}(\bar{\mathbf{a}}_t|\mathbf{y}_t, \omega) = \mathcal{N}\left(\bar{\mathbf{a}}_t|\mu_a(\mathbf{y}_t), \Sigma_a\right), \\ &q_\phi\left(\bar{\mathbf{y}}_t|\mathbf{y}_t\right) = \mathcal{N}\left(\bar{\mathbf{y}}_t|\mu_\phi(\mathbf{y}_t), \Sigma_\phi\right) \\ &q_\varphi\left(\bar{\mathbf{y}}_t|\mathbf{y}_t, \bar{\mathbf{y}}_{t+1}\right) = \mathcal{N}\left(\bar{\mathbf{y}}_{t+1}|\mu_\varphi(\mathbf{y}_t, \bar{\mathbf{y}}_{t+1}), \Sigma_\varphi\right) \\ &q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t, \omega) = \mathcal{N}\left(\mathbf{a}_t|\bar{\mathbf{a}}_t, \Sigma_\delta\right) \end{aligned} \tag{65}$$

with constant covariance matrices and linear mean functions:

$$\begin{aligned} &\mu_a(\mathbf{y}_t) = \mu_a + \Lambda_a \mathbf{y}_t \\ &\mu_\phi(\mathbf{y}_{t+1}) = \mu_\phi + \Lambda_\phi \mathbf{y}_{t+1} \\ &\mu_\varphi(\mathbf{y}_t, \bar{\mathbf{y}}_{t+1}) = \mu_\varphi + \Lambda_\varphi^{(1)} \mathbf{y}_t + \Lambda_\varphi^{(2)} \bar{\mathbf{y}}_{t+1} \end{aligned} \tag{66}$$

An alternative to these simple linear specifications could be non-linear means and covariances implemented by neural networks as in Ref.[5], or using some other universal function approximations such as Gaussian mixtures or trees. In this paper, we stick to simple linear Gaussian forms (65), (66).

The vector ω of parameters of the variational distribution q_ω thus includes three vectors μ_a, μ_ϕ, μ_φ, four 'slope' matrices Λ_a, Λ_ϕ, $\Lambda_\varphi^{(1)}$, $\Lambda_\varphi^{(2)}$, and four covariance matrices Σ_a, Σ_ϕ, Σ_φ, Σ_δ. For the marginalized distribution $q_\omega(\mathbf{a}_t|\mathbf{y}_t)$ in Eq.(61), we obtain

$$q_\omega(\mathbf{a}_t|\mathbf{y}_t) = \int d\bar{\mathbf{a}}\, q_{\bar{a}}(\bar{\mathbf{a}}|\mathbf{y}_t) q_a(\mathbf{a}_t|\bar{\mathbf{a}}) = \mathcal{N}\left(\mathbf{a}_t|\mu_a(\mathbf{y}_t), \Sigma_w\right), \;\; \Sigma_w = \Sigma_a + \Sigma_\delta \tag{67}$$

We can also marginalize over $\bar{\mathbf{y}}_{t+1}$:

$$q_{\bar{y}}\left(\bar{\mathbf{y}}_t|\mathbf{y}_t, \mathbf{y}_{t+1}\right) = \int d\bar{\mathbf{y}}_{t+1}\, q_\phi\left(\bar{\mathbf{y}}_{t+1}|\mathbf{y}_{t+1}\right) q_\varphi\left(\bar{\mathbf{y}}_t|\mathbf{y}_t, \bar{\mathbf{y}}_{t+1}\right) = \mathcal{N}\left(\bar{\mathbf{y}}_t|\mu_h(\mathbf{y}_t, \mathbf{y}_{t+1}), \Sigma_h\right) \tag{68}$$

where

$$\begin{aligned} &\mu_h(\mathbf{y}_t, \mathbf{y}_{t+1}) = \Lambda_\varphi^{(2)}\left(\mu_\phi + \Lambda_\phi \mathbf{y}_{t+1}\right) + \Lambda_\varphi^{(1)} \mathbf{y}_t + \mu_\varphi \\ &\Sigma_h = \Sigma_\varphi + \Lambda_\varphi^{(2)} \Sigma_\phi \left(\Lambda_\varphi^{(2)}\right)^T \end{aligned} \tag{69}$$

Finally, the joint distribution $q_h(\bar{\mathbf{y}}_t, \bar{\mathbf{y}}_{t+1}|\mathbf{y})$ is a Gaussian with the following inverse covariance matrix:

$$\Sigma_j^{-1} = \begin{bmatrix} \Sigma_\phi^{-1} + \Lambda_\varphi^{(2)} \Sigma_\varphi^{-1} \left(\Lambda_\varphi^{(2)}\right)^T & -\Lambda_\varphi^{(2)} \Sigma_\varphi^{-1} \\ -\Sigma_\varphi^{-1} \Lambda_\varphi^{(2)} & \Sigma_\varphi^{-1} \end{bmatrix} \tag{70}$$

5.5 Calculation of conditional free energy $\mathcal{F}_a$

Let us write the conditional free energy (63) as follows:

$$\begin{aligned}\mathcal{F}_a(\omega,\theta,\bar{\mathbf{a}}_t) &= \mathbb{E}_{q_a}\left[\log \pi_\theta\left(\mathbf{a}_t|\mathbf{y}_t\right) p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t,\mathbf{a}_t\right)\right] - \mathbb{E}_{q_a}\left[\log q_\omega(\mathbf{a}_t,\bar{\mathbf{a}}_t,\bar{\mathbf{y}}|\mathbf{y})\right] \\ &\equiv \mathcal{E}_a(\omega,\theta,\bar{\mathbf{a}}_t) + \mathcal{H}_a \end{aligned} \tag{71}$$

The second term in this expression is given by the following expression:

$$\mathcal{H}_a \equiv -\log q_\phi(\bar{\mathbf{y}}_{t+1}|\mathbf{y}_{t+1}) - \log q_\varphi(\bar{\mathbf{y}}_t|\mathbf{y}_t,\bar{\mathbf{y}}_{t+1}) - \log q_{\bar{a}}(\bar{\mathbf{a}}_t|\mathbf{y}_t) + H\left[q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t)\right] \tag{72}$$

where $H\left[q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t)\right]$ is the entropy of the marginal $q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t)$:

$$H\left[q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t)\right] = -\int d\mathbf{a}_t q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t)\log q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t) = \frac{1}{2}\log\left\{(2\pi e)^N\left|\Sigma_\delta\right|\right\} \tag{73}$$

Using specifications of marginals in Eq.(65), we obtain a closed-form expression for $\mathcal{H}_a$:

$$\begin{aligned}\mathcal{H}_a &= -\frac{1}{2}\left(\bar{\mathbf{y}}_{t+1}-\mu_\phi\right)^T\Sigma_\phi^{-1}\left(\bar{\mathbf{y}}_{t+1}-\mu_\phi\right) - \frac{1}{2}\left(\bar{\mathbf{y}}_t-\mu_\varphi\right)^T\Sigma_\varphi^{-1}\left(\bar{\mathbf{y}}_t-\mu_\varphi\right) \\ &-\frac{1}{2}\left(\bar{\mathbf{a}}_t-\mu_a\right)^T\Sigma_a^{-1}\left(\bar{\mathbf{a}}_t-\mu_a\right) + \frac{1}{2}\log\left\{(2\pi e)^N\left|\Sigma_\delta\right|\right\} \\ &-\frac{1}{2}\log\left|\Sigma_\phi\right| - \frac{1}{2}\log\left|\Sigma_\varphi\right| - \frac{1}{2}\log\left|\Sigma_a\right| - \frac{1}{2}\left(2N+N_a\right)\log 2\pi \end{aligned} \tag{74}$$

where N and N_a stand for dimensions of vectors $\bar{\mathbf{y}}_t$ and $\bar{\mathbf{a}}_t$, respectively.

On the other hand, the first 'energy' term $\mathcal{E}_a(\omega,\theta,\bar{\mathbf{a}}_t)$ in the conditional free energy (71) cannot be computed in closed form. Changing the integration variable $\mathbf{a}_t \to \delta\mathbf{a}_t = \mathbf{a}_t - \bar{\mathbf{a}}_t$, we write this term as follows:

$$\mathcal{E}_a(\omega,\theta,\bar{\mathbf{a}}_t) = \int d\delta\mathbf{a}_t\, q_a(\bar{\mathbf{a}}_t+\delta\mathbf{a}_t|\bar{\mathbf{a}}_t,\omega)\log\left[\pi_\theta\left(\bar{\mathbf{a}}_t+\delta\mathbf{a}_t|\mathbf{y}_t\right) p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t,\bar{\mathbf{a}}_t+\delta\mathbf{a}_t\right)\right] \tag{75}$$

As the distribution $q_a(\mathbf{a}_t|\bar{\mathbf{a}}_t,\omega)$ is sharply peaked around $\mathbf{a}_t = \bar{\mathbf{a}}_t$ (as long as Σ_δ is small enough), we can calculate this integral using a saddle point approximation. To this end, we need to compute $\pi_\theta\left(\bar{\mathbf{a}}_t+\delta\mathbf{a}_t|\mathbf{y}_t\right)$ and $p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t,\bar{\mathbf{a}}_t+\delta\mathbf{a}_t\right)$ for small values of $\delta\mathbf{a}_t$.

Let us start with a calculation of $p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t,\bar{\mathbf{a}}_t+\delta\mathbf{a}_t\right)$. A full transition probability for the state vector $\mathbf{y}_t = [\mathbf{x}_t,\mathbf{z}_t]^T$ is given by the following expression:

$$p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t,\bar{\mathbf{a}}_t+\delta\mathbf{a}_t\right) = p_z(\mathbf{z}_{t+1}|\mathbf{z}_t)\, p_\theta\left(\mathbf{x}_{t+1}|\mathbf{x}_t,\bar{\mathbf{a}}_t+\delta\mathbf{a}_t\right) \tag{76}$$

where

$$p_z(\mathbf{z}_{t+1}|\mathbf{z}_t) = \frac{1}{\sqrt{(2\pi)^K\left|\Sigma_z\right|}} e^{-\frac{1}{2}(\mathbf{z}_{t+1}-(\mathbf{I}-\Phi)\circ\mathbf{z}_t)^T\Sigma_z^{-1}(\mathbf{z}_{t+1}-(\mathbf{I}-\Phi)\circ\mathbf{z}_t)} \tag{77}$$

(see Eq.(13)), where K is the number of component in the vector of predictors $\mathbf{z}_t$. This term is independent of $\delta\mathbf{a}_t$ and serves as a constant term in Eq.(75).

The second conditional transition probability $p_\theta\left(\mathbf{x}_{t+1}|\mathbf{x}_t,\bar{\mathbf{a}}_t+\delta\mathbf{a}_t\right)$ in (76) can be computed as follows. First, we obtain the dynamics of the portfolio vector $\mathbf{x}_t$ using Eqs. (7) and (11):

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t + \mathbf{u}_t + \mathbf{r}_t\circ(\mathbf{x}_t+\mathbf{u}_t) \\ &= \mathbf{x}_t + \mathbf{u}_t + \left(r_f\mathbf{1} + \mathbf{W}\mathbf{z}_t - \mathbf{M}^T\mathbf{u}_t + \varepsilon_t\right)\circ(\mathbf{x}_t+\mathbf{u}_t) \\ &= (1+r_f)(\mathbf{x}_t+\mathbf{u}_t) + \mathrm{diag}\left(\mathbf{W}\mathbf{z}_t - \mathbf{M}\mathbf{u}_t\right)(\mathbf{x}_t+\mathbf{u}_t) + \varepsilon(\mathbf{x}_t,\mathbf{u}_t)\end{aligned} \tag{78}$$

Here we assumed that the matrix M of market impacts is diagonal with elements μ_i, and set

$$\mathbf{M} = \mathrm{diag}\,(\mu_i)\,,\ \varepsilon(\mathbf{x}_t, \mathbf{u}_t) \equiv \varepsilon_t \circ (\mathbf{x}_t + \mathbf{u}_t) \tag{79}$$

Eq.(78) shows that the dynamics are non-linear in controls $\mathbf{u}_t$ due to the market impact $\sim \mathbf{M}$. Expanding the action $\mathbf{u}_t$ as follows:

$$\mathbf{u}_t = [\mathbf{1}, -\mathbf{1}]\mathbf{a}_t = [\mathbf{1}, -\mathbf{1}]\bar{\mathbf{a}}_t + [\mathbf{1}, -\mathbf{1}]\delta\mathbf{a}_t \equiv \bar{\mathbf{u}}_t + \delta\mathbf{u}_t$$

so that $\delta\mathbf{u}_t = [\mathbf{1}, -\mathbf{1}]\delta\mathbf{a}_t = \mathbf{1}_{-1}^T\delta\mathbf{a}_t$ where $\mathbf{1}_{-1} \equiv [\mathbf{1}, -\mathbf{1}]^T$, a one-step conditional transition probability for $\mathbf{x}_t$ reads

$$p_\theta\left(\mathbf{x}_{t+1}|\mathbf{x}_t, \bar{\mathbf{a}}_t + \delta\mathbf{a}_t\right) = \frac{1}{\sqrt{(2\pi)^N\,|\Sigma_r|}} e^{-\frac{1}{2}\Delta_t^T\Sigma_r^{-1}\Delta_t} \tag{80}$$

where

$$\begin{aligned}\Delta_t \ &\equiv \frac{\mathbf{x}_{t+1}}{\mathbf{x}_t + \bar{\mathbf{u}}_t + \delta\mathbf{u}_t} - 1 - r_f - \mathbf{W}\mathbf{z}_t + \mathbf{M}^T\left(\bar{\mathbf{u}}_t + \delta\mathbf{u}_t\right) \\ &= \mathbf{d}_0(\bar{\mathbf{a}}_t) + \mathbf{d}_1(\bar{\mathbf{a}}_t)\delta\mathbf{a}_t + \mathbf{d}_2(\bar{\mathbf{a}}_t)\left(\delta\mathbf{a}_t\right)^2 + \ldots\end{aligned} \tag{81}$$

Here

$$\begin{aligned}\mathbf{d}_0(\bar{\mathbf{a}}_t) &= \frac{\mathbf{x}_{t+1}}{\mathbf{x}_t + \mathbf{1}_{-1}^T\bar{\mathbf{a}}_t} - 1 - r_f - \mathbf{W}\mathbf{z}_t + \mathbf{M}^T\mathbf{1}_{-1}^T\bar{\mathbf{a}}_t \\ \mathbf{d}_1(\bar{\mathbf{a}}_t) &= -\mathrm{diag}\left(\frac{\mathbf{x}_{t+1}}{\left(\mathbf{x}_t + \mathbf{1}_{-1}^T\bar{\mathbf{a}}_t\right)^2}\right)\mathbf{1}_{-1}^T + \mathbf{M}^T\mathbf{1}_{-1}^T \\ \mathbf{d}_2(\bar{\mathbf{a}}_t) &= \mathrm{diag}\left(\frac{\mathbf{x}_{t+1}}{\left(\mathbf{x}_t + \mathbf{1}_{-1}^T\bar{\mathbf{a}}_t\right)^3}\right)[\mathbf{1}, \mathbf{1}]\end{aligned} \tag{82}$$

These expressions depend non-linearly on $\bar{\mathbf{a}}_t$, and within a saddle point approximation, values $\bar{\mathbf{a}}_t$ in these expressions will be replaced by their mean values according to the distribution $q_{\bar{a}}$ defined in Eq.(65). On the other hand, other arguments of these expressions, namely $\mathbf{x}_t$ and $\mathbf{x_{t+1}}$ (and $\mathbf{z}_t$) are values directly observed in the variational likelihood (62), as well as in the full likelihood (54).

Next we have to compute the action policy $\pi_\theta\left(\bar{\mathbf{a}}_t + \delta\mathbf{a}_t|\mathbf{y}_t\right)$ for small values of $\delta\mathbf{a}_t$. To this end, we write the state vector as $\mathbf{y}_t = \bar{\mathbf{y}}_t + \delta\mathbf{y}_t$ (the meaning of this decomposition will be explained below), and introduce a locally-quadratic parametrization for the G-function:

$$G_t^\pi(\mathbf{y}_t, \bar{\mathbf{a}}_t + \delta\mathbf{a}_t) = \delta\mathbf{a}_t^T\mathbf{G}_{aa}\delta\mathbf{a}_t + \delta\mathbf{y}_t^T\mathbf{G}_{yy}\delta\mathbf{y}_t + \delta\mathbf{a}_t^T\mathbf{G}_{ay}\delta\mathbf{y}_t + \delta\mathbf{a}_t^T\mathbf{G}_a + \delta\mathbf{y}_t^T\mathbf{G}_y + g(\bar{\mathbf{y}}_t, \bar{\mathbf{a}}_t) \tag{83}$$

As the optimal action policy is given by Eq.(45), we have (where now $\mathbf{y}_t = \bar{\mathbf{y}}_t + \delta\mathbf{y}_t$)

$$\pi(\bar{\mathbf{a}}_t + \delta\mathbf{a}_t|\mathbf{y}_t) = \pi_0(\bar{\mathbf{a}}_t + \delta\mathbf{a}_t|\mathbf{y}_t)e^{\beta(G_t^\pi(\mathbf{y}_t, \bar{\mathbf{a}}_t + \delta\mathbf{a}_t) - F_t^\pi(\mathbf{y}_t))} \tag{84}$$

Substituting these expressions in Eq.(75) and retaining only quadratic terms in $\delta\mathbf{a}_t$ in in the $\log p_\theta\left(\mathbf{x}_{t+1}|\mathbf{x}_t, \bar{\mathbf{a}}_t + \delta\mathbf{a}_t\right)$ term (see Eq.(81)), we obtain

$$\mathcal{E}_a(\omega, \theta, \bar{\mathbf{a}}_t) = \mathcal{E}_a^{(0)}(\omega, \theta) + \mathcal{E}_a^{(1)}(\omega, \theta, \bar{\mathbf{a}}_t) \tag{85}$$

where

$$\begin{aligned}
\mathcal{E}_a^{(0)}(\omega,\theta) &= -\frac{1}{2}\left(\bar{\mathbf{a}}_t - \hat{A}_0 - \hat{A}_1\mathbf{y}_t\right)^T \Sigma_p^{-1}\left(\bar{\mathbf{a}}_t - \hat{A}_0 - \hat{A}_1\mathbf{y}_t\right) - \frac{1}{2}\mathbf{d}_0^T\Sigma_r^{-1}\mathbf{d}_0 \\
&+ \log p_z(\mathbf{z}_{t+1}|\mathbf{z}_t) - \frac{1}{2}\mathrm{Tr}\left[\Sigma_\delta \mathbf{d}_1^T\Sigma_r^{-1}\mathbf{d}_1\right] - \mathrm{Tr}\left[\mathrm{diag}\left(\Sigma_\delta\right)\mathbf{d}_0^T\Sigma_r^{-1}\mathbf{d}_2\right] \\
&-\frac{1}{2}\mathrm{Tr}\left[\Sigma_\delta\Sigma_p^{-1}\right] - \frac{1}{2}\log|\Sigma_p| - \frac{1}{2}\log|\Sigma_r| - \frac{N}{2}\log(2\pi) \\
\mathcal{E}_a^{(1)}(\omega,\theta,\bar{\mathbf{a}}_t) &= \beta\left(g(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t) - F_t^\pi(\mathbf{y}_t) + \delta\mathbf{y}_t^T\mathbf{G}_{yy}\delta\mathbf{y}_t + \delta\mathbf{y}_t^T\mathbf{G}_y + \mathrm{Tr}\left[\Sigma_\delta\mathbf{G}_{aa}\right]\right) \qquad (86)
\end{aligned}$$

where we omitted for compactness the dependence of $\mathbf{d}_0$, $\mathbf{d}_1$ and $\mathbf{d}_2$ on $\bar{\mathbf{a}}_t$, see Eq.(82), and $\mathbf{y}_t$ in $\mathcal{E}_a^{(0)}(\omega,\theta)$ stands for the observed state vector at time t. The second expression $\mathcal{E}_a^{(1)}(\omega,\theta,\bar{\mathbf{a}}_t)$ in Eq.(85) thus collects all terms that depend on the G- and F-functions, while terms independent of these functions are combined in $\mathcal{E}_a^{(0)}(\omega,\theta)$.

To summarize so far, Eqs.(85), (86), (72) jointly specify the conditional variational free energy (71), provided model parameters as well as the G-function (83) and the F-function are known. Once the conditional free energy $\mathcal{E}_a(\omega,\theta,\bar{\mathbf{a}}_t)$ is computed, the unconditional variational free energy (62) can be calculated using another saddle point approximation for the integral over $\bar{\mathbf{a}}_t$. This calculation will be presented next, while the following sections will describe the method of finding the policy π_θ and the G-function (83) and a corresponding F-function by linearization around $\bar{\mathbf{a}}_t, \bar{\mathbf{y}}$.

5.6 Calculation of variational free energy $\mathcal{F}$

Recall that in Eq.(83) we used the representation of the state vector $\mathbf{y}_t = \bar{\mathbf{y}}_t + \delta\mathbf{y}_t$. This decomposes the *observable* vector $\mathbf{y}_t$ into a sum of two *unobservable* quantities $\bar{\mathbf{y}}_t$ and $\delta\mathbf{y}_t$. When we condition on the linearization variable $\bar{\mathbf{y}}_t$, we can write $\delta\mathbf{y}_t = \mathbf{y}_t - \bar{\mathbf{y}}_t$ when performing integration over the outer hidden variables $\bar{\mathbf{a}}_t, \bar{\mathbf{y}}$.

The advantage of such decomposition of the observable $\mathbf{y}_t$ into two unobservables $\bar{\mathbf{y}}_t$, $\delta\mathbf{y}_t$ is that now we can assume that the F-function is locally quadratic around a random hidden conditioning (linearization) value $\bar{\mathbf{a}}_t, \bar{\mathbf{y}}$, and parametrize it as follows:

$$F_t^\pi(\mathbf{y}_t) = \delta\mathbf{y}_t^T\mathbf{F}_{yy}\delta\mathbf{y}_t + \delta\mathbf{y}_t^T\mathbf{F}_y + F_0(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t) \qquad (87)$$

Here

$$\mathbf{F}_{yy} = \begin{bmatrix} \mathbf{F}_{xx} & \mathbf{F}_{xz} \\ \mathbf{F}_{zx} & \mathbf{F}_{zz} \end{bmatrix}, \quad \mathbf{F}_y = \begin{bmatrix} \mathbf{F}_x \\ \mathbf{F}_z \end{bmatrix}, \qquad (88)$$

In a finite-horizon setting, parameters $\mathbf{F}_{yy}$, $\mathbf{F}_y$, F_0 become time-dependent, while in an infinite-horizon setting they do not explicitly depend on time. As will be shown below, the last term $F_0(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t)$ in (87) is a quadratic functional of $(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t)$.

Using Eq.(87) in (86), we have the following decomposition of the unconditional free energy (62):

$$\begin{aligned}
\mathcal{F}(\omega,\theta) &= \int d\bar{\mathbf{a}}_t d\bar{\mathbf{y}}\, q_{\bar{a}\bar{y}}(\bar{\mathbf{a}}_t,\bar{\mathbf{y}}|\mathbf{y},\omega)\left(\mathcal{H}_a + \mathcal{E}_a^{(0)}(\omega,\theta) + \mathcal{E}_a^{(1)}(\omega,\theta,\bar{\mathbf{a}}_t)\right) \\
&\equiv \mathcal{H} + \mathcal{F}^{(0)}(\omega,\theta) + \mathcal{F}^{(1)}(\omega,\theta) \qquad (89)
\end{aligned}$$

Here the first term can be computed analytically:

$$\begin{aligned}
\mathcal{H} &= -\int d\bar{\mathbf{y}}_t d\bar{\mathbf{y}}_{t+1} q_h(\bar{\mathbf{y}}_t,\bar{\mathbf{y}}_{t+1}|\mathbf{y})\log q_h(\bar{\mathbf{y}}_t,\bar{\mathbf{y}}_{t+1}|\mathbf{y}) + H\left[q_{\bar{a}}(\bar{\mathbf{a}}_t|\mathbf{y}_t)\right] \\
&= \frac{1}{2}\log\left\{(2\pi e)^{2N}\left|\Sigma_j\right|\right\} + \frac{1}{2}\log\left\{(2\pi e)^{N_a}\left|\Sigma_a\right|\right\} \qquad (90)
\end{aligned}$$

where the joint covariance matrix Σ_j is defined in Eq.(70).

The second term $\mathcal{F}^{(0)}(\omega,\theta)$ in Eq.(89) involves the integral of $\mathcal{E}_a^{(0)}(\omega,\theta)$ that collects all terms that are independent of the G- and F-functions. With a saddle point approximation, we replace $\bar{\mathbf{a}}_t$ in coefficients (82) by its mean value $\langle\bar{\mathbf{a}}_t\rangle = \mu_a(\mathbf{y}_t)$. Therefore, with this approximation, the remaining dependence of $\mathcal{E}_a^{(0)}(\omega,\theta)$ on $\bar{\mathbf{a}}_t$ is quadratic due to the first term. Integrating this expression with the Gaussian distribution q_a given by Eq.(65), we obtain

$$\begin{aligned}
\mathcal{F}^{(0)}(\omega,\theta) &= \int d\bar{\mathbf{a}}_t d\bar{\mathbf{y}}\, q_{\bar{a}\bar{y}}(\bar{\mathbf{a}}_t,\bar{\mathbf{y}}|\mathbf{y},\omega)\mathcal{E}_a^{(0)}(\omega,\theta) = \int d\bar{\mathbf{a}}_t\, q_{\bar{a}}(\bar{\mathbf{a}}_t|\mathbf{y}_t,\omega)\mathcal{E}_a^{(0)}(\omega,\theta) \\
&= -\frac{1}{2}\left(\mu_a - \hat{A}_0 + \left(\Lambda_a - \hat{A}_1\right)\mathbf{y}_t\right)^T \Sigma_p^{-1}\left(\mu_a - \hat{A}_0 + \left(\Lambda_a - \hat{A}_1\right)\mathbf{y}_t\right) - \frac{1}{2}\mathbf{d}_0^T\Sigma_r^{-1}\mathbf{d}_0 \\
&+ \log p_z(\mathbf{z}_{t+1}|\mathbf{z}_t) - \frac{1}{2}\mathrm{Tr}\left[\Sigma_\delta \mathbf{d}_1^T\Sigma_r^{-1}\mathbf{d}_1\right] - \mathrm{Tr}\left[\mathrm{diag}\left(\Sigma_\delta\right)\mathbf{d}_0^T\Sigma_r^{-1}\mathbf{d}_2\right] \\
&- \frac{1}{2}\mathrm{Tr}\left[\Sigma_\delta\Sigma_p^{-1}\right] - \frac{1}{2}\mathrm{Tr}\left[\Sigma_a\Sigma_p^{-1}\right] - \frac{1}{2}\log|\Sigma_p| - \frac{1}{2}\log|\Sigma_r| - \frac{N}{2}\log(2\pi)
\end{aligned} \tag{91}$$

Lastly, we consider the third term in Eq.(89) that depends on the G-function (83) and the F-function (87). Using these expressions, we can write the integrand $\mathcal{E}_a^{(1)}$ of this term defined in the second of Eqs.(86) as follows

$$\begin{aligned}
\mathcal{E}_a^{(1)}(\omega,\theta,\bar{\mathbf{a}}_t) &= \beta\left(g(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t) - F_t^\pi(\mathbf{y}_t) + \delta\mathbf{y}_t^T\mathbf{G}_{yy}\delta\mathbf{y}_t + \delta\mathbf{y}_t^T\mathbf{G}_y + \mathrm{Tr}\left[\Sigma_\delta\mathbf{G}_{aa}\right]\right) \\
&= \beta\left[\delta\mathbf{y}_t^T\left(\mathbf{G}_{yy} - \mathbf{F}_{yy}\right)\delta\mathbf{y}_t + \delta\mathbf{y}_t^T\left(\mathbf{G}_y - \mathbf{F}_y\right) + g(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t) - F_0(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t) + \mathrm{Tr}\left[\Sigma_\delta\mathbf{G}_{aa}\right]\right]
\end{aligned} \tag{92}$$

Relations between parameters of the G-function and F-function are derived in Appendix A in Sect. A.3, see Eqs.(A.23). Using the following auxiliary quantities (as defined below in Eq.(A.21) and repeated here for convenience)

$$\begin{aligned}
&\mathbf{b}_t = \bar{\mathbf{a}}_t - \hat{\mathbf{A}}_0 - \hat{\mathbf{A}}_1\bar{\mathbf{y}}_t,\ \tilde{\Sigma}_p = \Sigma_p^{-1} - 2\beta\mathbf{G}_{aa}, \\
&\Gamma_\beta = \frac{1}{\beta}\left(\mathbf{I} - \left(\Sigma_p^{-1}\right)^T\tilde{\Sigma}_p^{-1}\right)\Sigma_p^{-1},\ \Upsilon_\beta = \tilde{\Sigma}_p^{-1}\Sigma_p^{-1} \\
&\mathbf{E}_{ay} = \Upsilon_\beta\hat{\mathbf{A}}_1 + \frac{1}{2}\beta\tilde{\Sigma}_p^{-1}\mathbf{G}_{ay},\ \mathbf{D}_{ay} = \mathbf{G}_{ay}^T\Upsilon_\beta - \hat{\mathbf{A}}_1^T\Gamma_\beta \\
&\mathbf{E}_a = \hat{\mathbf{A}}_1^T\Upsilon_\beta\mathbf{G}_a + \beta\mathbf{G}_{ay}^T\tilde{\Sigma}_p^{-1}\mathbf{G}_a,\ \mathcal{L}_\beta = \frac{1}{2\beta}\left(\log|\Sigma_p| + \log\left|\tilde{\Sigma}_p\right|\right)
\end{aligned}$$

we obtain

$$\begin{aligned}
&\mathbf{F}_{yy} = \mathbf{G}_{yy} + \mathbf{G}_{ay}^T\mathbf{E}_{ay} - \frac{1}{2}\hat{\mathbf{A}}_1^T\Gamma_\beta\hat{\mathbf{A}}_1 \\
&\mathbf{F}_y = \mathbf{G}_y - \mathbf{D}_{ay}\mathbf{b}_t + \hat{\mathbf{A}}_1^T\Upsilon_\beta\mathbf{G}_a + \beta\mathbf{G}_{ay}^T\tilde{\Sigma}_p^{-1}\mathbf{G}_a \\
&F_0(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t) = g(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t) - \frac{1}{2}\mathbf{b}_t^T\Gamma_\beta\mathbf{b}_t - \mathbf{G}_a^T\Upsilon_\beta\mathbf{b}_t + \frac{\beta}{2}\mathbf{G}_a^T\tilde{\Sigma}_p^{-1}\mathbf{G}_a - \mathcal{L}_\beta
\end{aligned} \tag{93}$$

These relations suggest the following dependencies of different terms in free energy (87) on hidden variables $\bar{\mathbf{a}}_t$ and $\bar{\mathbf{y}}_t$. First, the quadratic term $\delta\mathbf{y}_t^T\mathbf{F}_{yy}\delta\mathbf{y}_t$ is quadratic in $\bar{\mathbf{y}}_t$ (as $\delta\mathbf{y}_t = \mathbf{y}_t - \bar{\mathbf{y}}_t$), and independent of $\bar{\mathbf{a}}_t$. The second term $\delta\mathbf{y}_t^T\mathbf{F}_y$ is quadratic in $\bar{\mathbf{y}}_t$ and linear in $\bar{\mathbf{a}}_t$. The free term $F_0(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t)$ is given by a sum of the term $g(\bar{\mathbf{x}}_t,\bar{\mathbf{a}}_t)$ that cancels out in Eq.(92), and a quadratic form in both $\bar{\mathbf{y}}_t$ and $\bar{\mathbf{a}}_t$, as indicated by the last of Eqs.(93).

The integral of this expression can therefore be computed in closed form with Gaussian hidden variable distributions (65). Using Eqs.(92) and (93), we obtain the following results for

expectations $\mathbb{E}_{\bar{\mathbf{a}}_t,\bar{\mathbf{y}}}[\cdot]$ of three terms in Eq.(87) under the variational distribution $q_{\bar{a}\bar{y}}\left(\bar{\mathbf{a}}_t,\bar{\mathbf{y}}|\mathbf{y}\right)$:

$$
\begin{aligned}
\mathcal{E}^{(1)}_{yy}(\omega,\theta,\bar{\mathbf{a}}_t) &\equiv \mathbb{E}_{\bar{\mathbf{a}}_t,\bar{\mathbf{y}}}\left[\beta\delta\mathbf{y}_t^T\left(\mathbf{G}_{yy}-\mathbf{F}_{yy}\right)\delta\mathbf{y}_t\right] = \beta\mathrm{Tr}\left[\Sigma_h^{-1}\left(\frac{1}{2}\hat{\mathbf{A}}_1^T\Gamma_\beta\hat{\mathbf{A}}_1-\mathbf{G}_{ay}^T\mathbf{E}_{ay}\right)\right]\\
&+\beta\left(\mathbf{y}_t-\mu_h(\mathbf{y})\right)^T\left(\frac{1}{2}\hat{\mathbf{A}}_1^T\Gamma_\beta\hat{\mathbf{A}}_1-\mathbf{G}_{ay}^T\mathbf{E}_{ay}\right)\left(\mathbf{y}_t-\mu_h(\mathbf{y})\right)\\
\mathcal{E}^{(1)}_{y}(\omega,\theta,\bar{\mathbf{a}}_t) &\equiv \mathbb{E}_{\bar{\mathbf{a}}_t,\bar{\mathbf{y}}}\left[\beta\delta\mathbf{y}_t^T\left(\mathbf{G}_{y}-\mathbf{F}_{y}\right)\right] = \beta\mathrm{Tr}\left(\Sigma_h^{-1}\mathbf{D}_{ay}\hat{\mathbf{A}}_1\right)-\beta\mu_h(\mathbf{y})\hat{\mathbf{A}}_1^T\mathbf{D}_{ay}^T\\
&+\beta\left(\mathbf{y}_t-\mu_h(\mathbf{y})\right)^T\left(\mathbf{E}_a+\mathbf{D}_{ay}\left(\mu_a(\mathbf{y}_t)-\hat{\mathbf{A}}_0\right)\right) \qquad (94)\\
\mathcal{E}^{(1)}_{0}(\omega,\theta,\bar{\mathbf{a}}_t) &\equiv \mathbb{E}_{\bar{\mathbf{a}}_t,\bar{\mathbf{y}}}\left[\beta\left(g(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t)-F_0(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t)+\mathrm{Tr}\left[\Sigma_\delta\mathbf{G}_{aa}\right]\right)\right] = \frac{\beta}{2}\mathrm{Tr}\left[\Sigma_a\Gamma_\beta+\Sigma_h\hat{\mathbf{A}}_1^T\Gamma_\beta\hat{\mathbf{A}}_1\right]\\
&+\frac{\beta}{2}\hat{\mathbf{A}}_0^T\Gamma_\beta\hat{\mathbf{A}}_0-\beta\hat{\mathbf{A}}_0^T\Gamma_\beta\mu_a(\mathbf{y}_t)-\beta\left(\mu_a(\mathbf{y}_t)-\hat{\mathbf{A}}_0\right)^T\Gamma_\beta\hat{\mathbf{A}}_1\mu_h(\mathbf{y})\\
&+\beta\mathbf{G}_a^T\Upsilon_\beta\left(\mu_a(\mathbf{y}_t)-\hat{\mathbf{A}}_0-\hat{\mathbf{A}}_1\mu_h(\mathbf{y})\right)-\frac{\beta^2}{2}\mathbf{G}_a^T\tilde{\Sigma}_p^{-1}\mathbf{G}_a+\beta\mathrm{Tr}\left[\Sigma_\delta\mathbf{G}_{aa}\right]+\beta\mathcal{L}_\beta
\end{aligned}
$$

where linear Gaussian mean functions $\mu_a(\mathbf{y}_t)$ and $\mu_h(\mathbf{y})$ are defined in Eqs.(66) and (69), respectively.

The final *closed form* result for the variational free energy (89) is therefore given by the sum of equations (90), (91) and (94):

$$\mathcal{F}(\omega,\theta,\pi_\theta) = \mathcal{H}+\mathcal{F}^{(0)}(\omega,\theta)+\mathcal{F}^{(1)}(\omega,\theta,\pi_\theta) \qquad (95)$$

Here we added the policy π_θ as an argument to $\mathcal{F}(\omega,\theta,\pi_\theta)$ to emphasize that the latter depends on three sets of inputs: variational parameters ω, generative model parameters Θ, and the optimal policy π_θ. The variational free energy (95) depends on the policy π_θ via its dependence on the parameter $\mathbf{G}_{aa},\mathbf{G}_{\mathbf{ay}}$ etc. that determine the locally-quadratic representation (83) of the optimal G-function (i.e. the optimal entropy-regularized Q-function).

The variational EM algorithm amounts to iterative maximization of Eq.(95). As the whole expression for the variational free energy (95) is analytical, both the E-step and the M-step of the algorithm are computationally light. In the E-step, we maximize it with respect to variational parameters ω while keeping parameters Θ and the G-function from the previous iteration. In the M-step, we maximize it with respect to generative model parameters Θ and policy π_θ. The outputs of the M-step are updated values of parameters Θ and updated values of parameters of G-function (83). We will now consider the M-step in more details.

5.7 M-step: policy optimization

In the M-step, updates of G-functions are done using Eqs.(A.9), (A.15), (A.17) derived in Appendix A. These equations provide a practical implementation of the general self-consistent system of equations (44), (45), (46) in our setting of locally-quadratic expansion for the G-function. In this setting, all integrations in these equations are performed analytically, thus providing a tractable version of this approach in our highly dimensional continuous state-action setting. Note that the original version of G-learning was only explored in [18] in a low-dimensional discrete state setting.

As discussed in Appendix A, Eqs.(A.9), (A.15), (A.17) can be used for either a single investor or a market portfolio. In the former case, the update is performed backward in time, starting with a terminal time T and a specific terminal condition on the F-function or/and G-function.

In the latter case of a market portfolio, these equations can be used in a time-stationary setting as update rules for time-independent coefficients of the G-function.

When coefficients of the Q-functions are computed in this way for time step t, the optimal action distribution for $\delta\mathbf{a}_t$ is computed using Eq.(84) which we repeat here for convenience:

$$\pi_\theta(\bar{\mathbf{a}}_t + \delta\mathbf{a}_t|\mathbf{y}_t) = \pi_0(\bar{\mathbf{a}}_t + \delta\mathbf{a}_t|\mathbf{y}_t)e^{\beta(G_t^\pi(\mathbf{y}_t,\bar{\mathbf{a}}_t+\delta\mathbf{a}_t)-F_t^\pi(\mathbf{y}_t))} \tag{96}$$

When $\bar{\mathbf{a}}_t$ is fixed by conditioning, we view the distribution as a Gaussian distribution for $\delta\mathbf{a}_t$ with the mean $\widehat{\delta\mathbf{a}}_t = \hat{\mathbf{A}}_0 + \hat{\mathbf{A}}_1\mathbf{y}_t - \bar{\mathbf{a}}_t$. As the reference distribution π_0 is Gaussian and the Q-function is quadratic, the optimal action policy π is again Gaussian with a new mean and covariance:

$$\pi_\theta(\delta\mathbf{a}_t|\mathbf{y}_t) = \pi_0(\delta\mathbf{a}_t|\mathbf{y}_t)e^{\beta(G_t^\pi(\mathbf{y}_t,\mathbf{a}_t)-F_t^\pi(\mathbf{y}_t))} = \mathcal{N}\left(\delta\mathbf{a}_t|\widehat{\delta\mathbf{a}}_t', \Sigma_p'\right) \tag{97}$$

where $\mathcal{N}(\cdot)$ is a multivariate Gaussian distribution with the following mean and covariance matrix:

$$\begin{aligned}
\widehat{\delta\mathbf{a}}_t' &= \Sigma_p'\left(\Sigma_p^{-1}\widehat{\delta\mathbf{a}}_t + \beta\mathbf{G}_{ay}\delta\mathbf{y}_t + \beta\mathbf{G}_a\right)\\
\Sigma_p' &= \left[\Sigma_p^{-1} - 2\beta\mathbf{G}_{aa}\right]^{-1}
\end{aligned} \tag{98}$$

These relations can be viewed as Bayesian updates for the current iteration mean $\widehat{\delta\mathbf{a}}_t$ (see Eq.(A.19)) and variance Σ_p of the optimal action policy relative to their values for the "prior" reference policy (A.18). Note that in the limit $\beta \to 0$, Eq.(98) produces no update, $\widehat{\delta\mathbf{a}}_t' = \widehat{\delta\mathbf{a}}_t$. This is as expected, as in this 'high-temperature' limit the agent only maximizes the negative of the KL entropy but not rewards.

They can be also expressed as updates for the action policy (29) in terms of original policy variables. As $\widehat{\delta\mathbf{a}}_t = \hat{\mathbf{A}}_0 + \hat{\mathbf{A}}_1\mathbf{y}_t - \bar{\mathbf{a}}_t$, the update (98) of the mean $\widehat{\delta\mathbf{a}}_t$ implies an update of parameters $\hat{\mathbf{A}}_0$ and $\hat{\mathbf{A}}_1$. Substituting this expression into Eq.(98) and comparing an intercept and linear terms in this equation produces an update for the mean of the policy (29):

$$\begin{aligned}
\Sigma_p^{(k+1)} &= \left[\left(\Sigma_p^{(k)}\right)^{-1} - 2\beta\mathbf{G}_{aa}^{(k)}\right]^{-1}\\
\hat{\mathbf{A}}_0^{(k+1)} &= \bar{\mathbf{a}}_t + \Sigma_p^{(k+1)}\left(\Sigma_p^{(k)}\right)^{-1}\left(\hat{\mathbf{A}}_0^{(k)} - \bar{\mathbf{a}}_t\right) + \beta\Sigma_p^{(k+1)}\left(\mathbf{G}_a^{(k)} - \mathbf{G}_{ay}^{(k)}\bar{\mathbf{y}}_t\right)\\
\hat{\mathbf{A}}_1^{(k+1)} &= \Sigma_p^{(k+1)}\left(\left(\Sigma_p^{(k)}\right)^{-1}\hat{\mathbf{A}}_1^{(k)} + \beta\mathbf{G}_{ay}^{(k)}\right)
\end{aligned} \tag{99}$$

where we use values of parameters $\mathbf{G}_{aa}$ etc. corresponding to the current iteration of the algorithm. Again, these updates degenerate and become identities in the high temperature limit $\beta \to 0$. On the other hand, in the opposite limit $\beta \to \infty$ we obtain finite and non-trivial updates.

Note that in a finite-horizon setting of a single investor, parameters $\mathbf{G}_{aa}, \mathbf{G}_{ay}$ etc. are time-dependent, therefore coefficients $\hat{\mathbf{A}}_1$ will be also be time-dependent. On the other hand, for a market portfolio inference, parameters of the G-function are time-independent, thus parameters $\hat{\mathbf{A}}_0$ and $\hat{\mathbf{A}}_1$ would also be time-independent[13].

The updated policy for step $k+1$ now takes the form

$$\pi^{(k+1)}(\mathbf{a}_t|\mathbf{y}_t) = \mathcal{N}\left(\mathbf{a}_t|\hat{\mathbf{A}}_0^{(k+1)} + \hat{\mathbf{A}}_1^{(k+1)}\mathbf{y}_t, \Sigma_p^{(k+1)}\right) \tag{100}$$

[13] An apparent dependence of $\hat{\mathbf{A}}_0$ on $\bar{\mathbf{a}}_t, \bar{\mathbf{y}}_t$ is a result of our conditioning on these values in the outside integral in Eq.(62). While *updates* of $\hat{\mathbf{A}}_0$ may depend on the conditioning/linearization variables $\bar{\mathbf{a}}_t, \bar{\mathbf{y}}_t$ as in Eq.(99), a final fixed-point value of $\hat{\mathbf{A}}_0$ obtained with this method is a constant parameter that is independent of $\bar{\mathbf{a}}_t, \bar{\mathbf{y}}_t$.

Equations (99) and (100) represent one of our main results. The point is that the last of Eqs.(99) shows that a non-zero coefficients $\hat{\mathbf{A}}_1^{(k+1)}$ is obtained even if its value at the previous iteration was zero. Applying this for $k = 0$, it means that this coefficient (which induces the dependence of the optimal policy on the state $\mathbf{y}_t$) becomes non-zero even if we start with $\hat{\mathbf{A}}_1^{(0)}$ in the policy prior (29).

Furthermore, it implies that at convergence, updates (100) produce some fixed values $\hat{\mathbf{A}}_0$, $\hat{\mathbf{A}}_1$ of policy parameters. Our model therefore predicts that the optimal investment policy is Gaussian whose mean is is *linear* in the state variable $\mathbf{y}_t = [\mathbf{x}_t, \mathbf{z}_t]$, as in the Iterative Linear-Quadratic Gaussian (iLQG) regulator of Todorov and Li [53].

When $\mathbf{x}_t$ is identified with a market portfolio and an agent is our bounded-rational market-agent, Eq.(100) (used with such fixed-point values $\hat{\mathbf{A}}_0$, $\hat{\mathbf{A}}_1$) defines an optimal "market-implied" action policy. This provides a probabilistic and multi-period extension of a market-optimal static portfolio in a one-period setting of the Black-Litterman model [8] and inverse portfolio optimization approach of Bertsimas *et. al.* [7].

On the other hand, as we mentioned above, the same framework can be applied to an individual investor provided we have access to proprietary trading data of that particular investor. In this case, actions $\mathbf{a}_t$ will be actions of that investor. If these actions are observable, Eq.(100) can be directly used within a Maximum Likelihood estimation. We discuss this as a special case of our model in Appendix B, while here we proceed with the case when actions (of either a market-agent or an individual investor) are unobservable.

While the main focus of this paper is on inference of a market-wide bounded-rational agent, the algorithm can also be used for a single large investor whose trades impact the market but cannot be directly observed. Such setting may be of interest for intraday trading when the market moves have stronger causality relations with impacts of individual large trades. While for this case variables $\mathbf{x}_t$ correspond to the dollar values of positions in different stocks, they become total capitalizations of all firms in a market portfolio for inference of a market.

5.8 IRL for a market portfolio vs IRL for a single investor

Up to this point in the paper, our mathematical formulation for a single-investor and market portfolio was nearly uniform. In both cases, the optimal investment policy is given by Eq.(100), and in both cases, inference can be made using variational EM algorithm with a single-step variational free energy given by Eq.(95). Now we come to differences between these two cases.

The first difference is in computational procedures for computing parameters entering these equations. For a single investor case, if actions are unobserved, coefficients in Eqs.(100) and (.(95) are time-dependent, and should be computed by a backward recursion starting from a terminal date $t = T$, as described in Appendix A[14]. For a market portfolio case, the problem is stationary, as there is no single unique horizon T for planning in the market.

This means that coefficients are now time-independent. The self-consistent set of Eqs.(44), (45), (46) for the stationary case reads

$$\begin{aligned} F^\pi(\mathbf{y}_t) &= \frac{1}{\beta}\log\sum_{\mathbf{a}_t}\pi_0(\mathbf{a}_t|\mathbf{y}_t)e^{\beta G^\pi(\mathbf{y}_t,\mathbf{a}_t)} \\ G^\pi(\mathbf{y}_t,\mathbf{a}_t) &= \hat{R}(\mathbf{y}_t,\mathbf{a}_t) + \gamma\mathbb{E}_{t,\mathbf{a}}\left[F^\pi(\mathbf{y}_{t+1})|\,\mathbf{y}_t,\mathbf{a}_t\right] \\ \pi(\mathbf{a}_t|\mathbf{y}_t) &= \pi_0(\mathbf{a}_t|\mathbf{y}_t)e^{\beta(G^\pi(\mathbf{y}_t,\mathbf{a}_t)-F^\pi(\mathbf{y}_t))} \end{aligned} \tag{101}$$

[14] A single investor case with unobserved actions may probably be less common than a scenario with observable actions, but the latter is a straightforward case as it does not need hidden variables at all, see Appendix B.

Computationally, this formulation amounts to solving the self-consistent system Eqs.(101) as fixed point equations for time stationary G-function, F-function, and policy π_θ. In this setting, equations (A.23) become fixed point matrix equations, because now they relate matrix coefficients of the F-function (A.22) with themselves, rather than with their next-period values, as was the case in a finite-horizon specification. In the stationary setting, these equations can be used as update rules for parameters of the F-function by reading them from the right to the left, the same way as they are used in each step of a time-depending case.

A second major difference of IRL for the market portfolio from the single investor case is that while states $\mathbf{y}_t$ are directly observable in this settings, actions $\mathbf{a}_t$ are *not*. They *might* be made observable in a multi-agent version of the model, where the objective would be to model market-beating strategies, rather than just market-fitting strategies. However, in the inverse optimization IRL setting of this paper, we have only one agent representing a bounded-rational component of the market itself, thus it cannot trade stocks with other agents.

Therefore, its actions $\mathbf{a}_t$ cannot be observed or interpreted as changes in *numbers* of stocks in the portfolio. Our agent does only a fictitious self-play of its trading decisions, but does not trade directly with any other counter-party. The only observable effects of actions of the agent are price changes resulting from heating the market via the trading impact mechanism.

We are now ready to formulate our final variational EM algoritm for inference of either an individual investor or a market optimal portfolio. A different and simpler algorithm for a special (and the most interesting) case of a market optimal portfolio will be presented in Sect. 6.

5.9 Invisible Hand Inference with Free energy (IH-IF) algorithm

The complete IRL algorithm for learning the optimal policy of a bounded-rational agent (either a market-agent or a single investor) whose actions are unobservable that we call the Invisible Hand Inference with Free energy (IH-IF) is given by Algorithm 1.

Our algorithm is a variational EM algorithm that amounts to iterative maximization of Eq.(95). In the E-step, we maximize it with respect to variational parameters ω while keeping parameters $\theta = (\lambda, \mu_i, \beta, \mathbf{W}, \Gamma, \Upsilon), \hat{\mathbf{A}}_0$, $\hat{\mathbf{A}}_1$, Σ_p and the G-function from the previous iteration. In the M-step, we maximize it with respect to generative model parameters Θ and policy π_θ. The outputs of the M-step are updated values of parameters θ and updated values of parameters of G-function (83).

In more details, at each iteration, we sample a new random mini-batch of N_b T-step trajectories $(\mathbf{y}_1, \ldots, \mathbf{y}_{t+T})$. For the case of a market portfolio, we can take $T = 1$, so that a mini-batch has N_b one-step transitions $(\mathbf{y}_1, \mathbf{y}_{t+1})$. For inference of a single large investor, T should be set to be a finite planning horizon of the investor.

All subsequent calculations in a given iteration of the algorithm are done for this mini-batch. We define the free energy of a mini-batch as

$$\mathcal{F}_b(\omega, \theta) = \sum_{b=1}^{N_b} \sum_{t=0}^{T} \mathcal{F}(\omega, \theta, t) \tag{102}$$

where $\mathcal{F}(\omega, \theta, t)$ is defined in Eq.(95), while here we add a third argument to emphasize the time dependence in observations.

In the E-step, we maximize $\mathcal{F}_b(\omega, \theta)$ with respect to variational parameters ω. In the M-step, we compute updates of parameters of the G-function, policy π_θ as functions of θ, and then use these expressions to compute $\mathcal{F}_b(\omega, \theta)$ as a function of θ.

This is done as follows. In step 1, the expectation of the next-time F-function is computed with Eq.(A.14) used as an update equation for parameters of the model, or within a backward

recursion that starts with a fixed terminal condition at time $t = T$, for IRL of an individual investor. In step 2, we compute the reward using Eq.(A.8). In step 3, an update of the Q-function is performed using Eq.(A.17). The time-t F-function is computed in step 4 using Eq.(A.23). Finally, in step 5, the optimal policy as a function of θ is recomputed using Eq.(100). Computing these quantities for all transitions in the mini-batch, we obtain the free energy (102) for the mini-batch. This is used to produce an update of the current estimation of θ using a learning rate α_θ. The new updated values of θ are then used to update parameters $\hat{\mathbf{A}}_1^{(k)}$, $\hat{\mathbf{A}}_1^{(k)}$, $\Sigma_p^{(k)}$ of the policy π_θ. Then the algorithm proceeds to the next iteration.

Data: a sequence of states and signals
Result: the reward function, optimal policy, and value function
Set the learning rates α_θ, α_ω, batch size N_b, initial parameters $\theta^{(0)}$, $\omega^{(0)}$, $\hat{\mathbf{A}}_0^{(0)}$, $\hat{\mathbf{A}}_1^{(0)}$, $\Sigma_p^{(0)}$
Set $k = 1$
while *not converged* **do**
  Draw a new mini-batch of N_b T-step trajectories $(\mathbf{y}_t, \ldots, \mathbf{y}_{t+T})$ (can set $T = 1$ for a market portfolio)
  E-step:
  Compute the free energy $\mathcal{F}_b(\omega, \theta^{(k-1)})$ of the mini-batch using Eq.(102)
  Update recognition model parameters $\omega^{(k)} = (1 - \alpha_\omega)\omega^{(k-1)} + \alpha_\omega \frac{\partial}{\partial \omega} \mathcal{F}_b(\omega, \theta^{(k-1)})$
  M-step: Maximize $\mathcal{F}_b(\omega^{(k)}, \theta)$ as a function of θ:
  for *each transition* $(\mathbf{y}_t, \mathbf{y}_{t+1})$ *(for a single investor, take* $t = T - 1, \ldots, 0$*)* **do**
    1. Compute the expected value at time t of the F-function at time $t + 1$.
    2. Compute the reward as a function of θ.
    3. Use steps 1 and 2 to update the Q-function at time t
    4. Compute the value of the F-function at time t.
    5. Recompute the policy distribution $\pi_\theta(\mathbf{a}_t|t, \mathbf{y}_t)$ as a function of θ by updating its mean and variance.
  end
  Compute the free energy $\mathcal{F}_b(\omega^{(k)}, \theta)$ of the mini-batch using Eq.(102)
  Update the parameter vector $\theta^{(k)} = (1 - \alpha_\theta)\theta^{(k-1)} + \alpha_\theta \frac{\partial}{\partial \theta} \mathcal{F}_B(\omega^{(k)}, \theta)$
  Use the new value $\theta^{(k)}$ to compute $\hat{\mathbf{A}}_1^{(k)}$, $\hat{\mathbf{A}}_1^{(k)}$, $\Sigma_p^{(k)}$
  Increment $k = k + 1$
end

Algorithm 1: The Invisible Hand Inference with the Free energy (IH-IF) variational EM IRL algorithm that learns the reward function, optimal policy and value function from a history of prices and signals, for either a market portfolio or a single investor.

6 IRL for the market portfolio

When actions are unobserved or unobservable, the variational EM formulation (95) provides a general and tractable algorithm to estimate the original model parameters Θ from observed trajectories of stock capitalizations. The price one has to pay to solve the problem in this way is a need to specify a variational distribution with its own parameters ω, and estimate these parameters jointly with Θ in a way specified by a variational EM algorithm.

As we will show next, an alternative and simpler method of estimation model can be obtained simply by plugging Eq.(100) into the market return model (11). To this end, we note that that once we obtained Eq.(100), we can 'forget' how it was derived using RL, IRL, neuroscience

etc., and simply treat it as a model with free tunable parameters $\hat{\mathbf{A}}_0$, $\hat{\mathbf{A}}_1$ and Σ_p. Substituting Eq.(100) into Eq.(11) gives rise to a purely *econometric* model of market returns, which can be viewed (and estimated) as a model on its own. As will be shown below, this produces a model that predicts *mean reversion* in stock returns.

6.1 Market dynamics: dynamically generated mean reversion

Recall that for a vector of N stocks, we introduced a size $2N$-action vector $\mathbf{a}_t = [\mathbf{u}_t^{(+)}, \mathbf{u}_t^{(-)}]$, so that an action $\mathbf{u}_t$ was defined as a difference of two non-negative numbers $\mathbf{u}_t = \mathbf{u}_t^{(+)} - \mathbf{u}_t^{(-)} = [\mathbf{1}, -\mathbf{1}]\mathbf{a}_t \equiv \mathbf{1}_{-1}^T \mathbf{a}_t$.

Therefore, the joint distribution of $\mathbf{a}_t = [\mathbf{u}_t^{(+)}, \mathbf{u}_t^{(-)}]$ is given by our Gaussian policy $\pi_\theta(\mathbf{a}_t|\mathbf{y}_t)$. This means that the distribution of $\mathbf{u}_t = \mathbf{u}_t^{(+)} - \mathbf{u}_t^{(-)}$ is also Gaussian. Let us write it therefore as follows:

$$\pi_\theta(\mathbf{u}_t|\mathbf{y}_t) = \mathcal{N}\left(\mathbf{u}_t|\mathbf{U_0} + \mathbf{U_1}\mathbf{y_t}, \mathbf{\Sigma_u}\right) \tag{103}$$

Here $\mathbf{U_0} = \mathbf{1}_{\mathbf{-1}}^{\mathbf{T}}\mathbf{A_0}$ and $\mathbf{U_1} = \mathbf{1}_{\mathbf{-1}}^{\mathbf{T}}\mathbf{A_1}$.

Eq.(103) means that $\mathbf{u}_t$ is a Gaussian random variable that we can write as follows:

$$\mathbf{u}_t = \mathbf{U_0} + \mathbf{U_1}\mathbf{y_t} + \varepsilon_{\mathbf{t}}^{(\mathbf{u})} = \mathbf{U_0} + \mathbf{U}_{\mathbf{1}}^{(\mathbf{x})}\mathbf{x_t} + \mathbf{U}_{\mathbf{1}}^{(\mathbf{z})}\mathbf{z_t} + \varepsilon_{\mathbf{t}}^{(\mathbf{u})} \tag{104}$$

where $\varepsilon_t^{(u)} \sim \mathcal{N}(0, \Sigma_u)$ is a Gaussian random noise.

The most important feature of this expression that we need going forward is is linear dependence on the state $\mathbf{x}_t$. As can be seen in Eqs.(99) and (100), the variational EM algorithm developed above suggests that a coefficient of such dependence should be non-vanishing.

This is the only result from the model developed in this paper that we will use in this section in order to construct a simple dynamic market model resulting from our approach. In order to end up with non-negative market prices in the model, we use a deterministic limit of Eq.(104), where in addition we set $\mathbf{U_0} = \mathbf{U}_{\mathbf{1}}^{(\mathbf{z})} = \mathbf{0}$, and replace $\mathbf{U}_{\mathbf{1}}^{(\mathbf{x})} \to \phi$ to simplify the notation. We thus obtain a simple deterministic policy

$$\mathbf{u}_t = \phi \mathbf{x}_t \tag{105}$$

Next, let us recall Eqs.(7) and (11), which we repeat were with a substitution $\mathbf{W} \to \mathbf{w}$ and $\mathbf{M} \to \mu$:

$$\begin{aligned} \mathbf{x}_{t+1} &= (1 + r_t) \circ (\mathbf{x}_t + \mathbf{u}_t) \\ \mathbf{r}_t - r_f \mathbf{1} &= \mathbf{w}\mathbf{z}_t - \mu \mathbf{u}_t + \varepsilon_t^{(r)} \end{aligned} \tag{106}$$

where r_f is a risk-free rate, $\mathbf{z}_t$ is a vector of predictors with factor loading matrix $\mathbf{w}$, μ is a matrix of permanent market impacts with a linear impact specification, and $\varepsilon_t^{(r)}$ is a vector of residuals with $\mathbb{E}\left[\varepsilon_t^{(r)}\right] = 0$ and $\mathrm{Var}_t\left[\varepsilon_t^{(r)}\right] = \Sigma_r$.

In general case, the second equation in (106) assumes a single vector of predictor $\mathbf{z}_t$ for all stocks in a market portfolio. If we have K individual predictors $\mathbf{z}_t^{(i)} = [z_{t1}^{(i)}, \ldots, z_{tK}^{(i)}]$ for each stock i, we can stack them together as $\mathbf{z}_t = [\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(N)}]^T$, so that $\mathbf{z}_t$ has length KN. Respectively, matrix $\mathbf{w}$ will have the size $N \times KN$. Each row i in this matrix will only have K non-zero elements in positions $i, \ldots, i + K$ (so that to only include i's name predictors). This results in KN free parameters in matrix $\mathbf{w}$. If desired or needed, the number of free parameters

can be reduced if we enforce some symmetries, e.g. enforce a requirement that factor loadings for all names in a given sector should have the same value.

Substituting Eq.(105) into Eqs.(106) and simplifying, we obtain

$$\Delta \mathbf{x}_t = \mu \circ \phi \circ (1+\phi) \circ \mathbf{x}_t \circ \left(\frac{\phi + (1+\phi)(r_f + \mathbf{w}\mathbf{z}_t)}{\mu\phi(1+\phi)} - \mathbf{x}_t \right) + (1+\phi) \circ \mathbf{x}_t \circ \varepsilon_t^{(r)} \tag{107}$$

Introducing parameters

$$\kappa \Delta t = \mu \circ \phi \circ (1+\phi),\ \theta(\mathbf{z}_t) = \frac{\phi + (1+\phi)(r_f + \mathbf{w}\mathbf{z}_t)}{\mu\phi(1+\phi)},\ \sigma(\mathbf{x}_t)\sqrt{\Delta t} = (1+\phi) \circ \mathbf{x}_t \tag{108}$$

(here Δt is a time step) and replacing $\varepsilon_t^{(r)} \to \varepsilon_t$, we can write Eq.(107) more suggestively as

$$\Delta \mathbf{x}_t = \kappa \circ \mathbf{x}_t \circ (\theta(\mathbf{z}_t) - \mathbf{x}_t)\, \Delta t + \sigma(\mathbf{x}_t)\sqrt{\Delta t} \circ \varepsilon_t \tag{109}$$

In this equation, $\circ$ stands for an element-wise (Hadamard) product. Note that this equation has a *quadratic* mean reversion. It is quite different from models with *linear* mean reversion such as the Ornstein-Uhlenbeck (OU) process. Eq.(109) is the second main result of this paper.

Equation (109) describes mean reverting dynamics with a signal-driven mean reversion level $\theta(\mathbf{z}_t)$, and a mean reversion speed κ proportional to market impact parameter vector μ. It is easy to see that in the limit of vanishing market impact $\mu \to 0$, $\phi \to 0$, Eq.(109) reduces to the log-normal return model given by Eq.(11) without the action term $\mathbf{u}_t$:

$$\frac{\Delta \mathbf{x}_t}{\mathbf{x}_t} = r_f + \mathbf{w}\mathbf{z}_t + \varepsilon_t \tag{110}$$

Therefore, the conventional log-normal return dynamics (with signals) is reproduced in our framework in the limit $\mu \to 0$, $\phi \to 0$. However, when parameters μ, ϕ are small but non-zero, Eqs. (110) and (109) describe *qualitatively different* dynamics. While Eq.(110) is scale-invariant with respect to scale transformations $\mathbf{x}_t \to \alpha \mathbf{x}_t$ with α being a scaling parameter, the non-linear mean reverting dynamics (109) are *not* scale invariant.

This is of course due to the fact that our market-wide agent aggregates all agents in the market. As their individual trade impacts induce a dependence of dynamics on a dimensional market impact parameter μ, scale invariance is broken in the resulting market dynamics (109).

Therefore, even if parameters κ, ϕ are small but non-vanishing, Eq.(109) produces a potentially highly complex non-linear dynamics with broken scale invariance and ensuing multi-period auto-correlations.

These non-linear dynamics with a *dynamically* generated mean reversion level $\theta(\mathbf{z}_t)$ are produced from simple linear dynamics (11) with a Linear-Quadratic-Gaussian (LQG) control $\mathbf{u}_t$. A peculiar feature of our model is that it has very clear origins for both the *level* and the *speed* of mean reversion. As can be seen from Eqs.(109), the level $\theta(\mathbf{z}_t)$ is driven by external signals $\mathbf{z}_t$, which makes an intuitive sense. On the other hand, the *speed* of reverting to such 'target' price values is proportional to the market impact parameter vector μ, that also intuitively makes sense.

It is important to note here that our model demonstrates some features that are typical for self-organizing systems, such as non-linear mean reversion effects, long-term correlations resulting from such mean reversion, and a dynamic adaptivity to external signals $\mathbf{z}_t$. Therefore, our construction of self-learning by a fictitious self-play by an agent, that imitates simultaneously all traders in the market, provides a specific illustration of equivalence between self-organization and decision-making that was suggested in [57].

Another important comment has to do with time scales in the problem. There are a few of them in our model. First, we have a vector of external signals $\mathbf{z}_t$. Each one of them has its own relaxation time τ_{zk} where $k = 1, \ldots, K$ is a number of signals.

Assume for simplicity that we have only one scalar signal z_t with a characteristic relaxation time $\tau_z \sim 1/\kappa_z$ where κ_z is the mean reversion speed of the signal. This can be compared with the characteristic relaxation time of the *system* $\tau_x \sim 1/\kappa$. The setting of this paper implicitly assumes that $\tau_x \leq \tau_z$, that is, $\kappa \geq \kappa_z$, so that the market is close to a non-equilibrium steady state, and it manages to digest a new information in signals $\mathbf{z}_t$ at each step, and fully adjust market prices (at the price of the information cost g_t, see Eq.(35)).

On the other hand, we might have a very different dynamics if $\kappa \leq \kappa_z$. In this case, the market would be in non-equilibrium transient state without a steady state. Yet a different scenario may occur when a large jump in $\mathbf{z}_t$ occurs at time t relative to its previous value (following e.g. a major financial, economic or political event), and then continues to fluctuate only mildly around a new level. In this case, the mean stock price level $\theta(\mathbf{z}_t)$ that adjusted at time t to the *previous* value of the signals, becomes not the true dynamic optimum, but only a *metastable* state. Further comments on such scenario will be given in Sect. 8.

In a one-dimensional (1D) case with a constant mean reversion level $\theta(\mathbf{z}_t) = \theta$, Eq.(109) produces the following dynamics for a re-scaled variable $s_t = x_t/\theta$:

$$\Delta s_t = \mu s_t(1 - s_t) + \sigma\sqrt{\Delta t} s_t \varepsilon_t, \quad \mu \equiv \kappa\theta\Delta t \tag{111}$$

Dynamics described by Eq.(111) or its noiseless limit $\sigma \to 0$ are widely encountered or used in physics and biology. In particular, the limit $\sigma \to 0$ of Eq.(111) describes the logistic map dynamics, that arises e.g. in the Malthus-Verhulst model of population growth (see e.g. [55]), or in Feigenbaum bifurcations in the logistic map chaos, that arise when $3 \leq \mu < 4$ in Eq.(111), see e.g. [47]. When $\sigma > 0$, Eq.(111) describes a logistic map with a multiplicative thermal noise, which may produce highly complex dynamics [4].

We can also consider a continuous-time limit of 1D dynamics implied by Eq.(109):

$$dx_t = \kappa x_t \left(\theta - x_t\right) dt + \sigma x_t \, dW_t \tag{112}$$

where W_t is a standard Brownian motion. This 1D process is known in the economics and finance literature as a Geometric Mean Reversion (GMR) process. Equivalently, we can introduce a scaled variable $s_t = \kappa x_t$, for which we obtain

$$ds_t = \left(\lambda_t s_t - s_t^2\right) dt + \sigma s_t \, dW_t, \quad \lambda_t \equiv \kappa\theta_t \tag{113}$$

which is a form mostly used in physics literature [24]. As discussed in [24], if we keep parameter $\lambda_t \equiv \kappa\theta_t$ constant in time, i.e. $\lambda_t \to \lambda$ and look at the behavior of the system in the limit $\sigma \to 0$, the system exhibits a second-order phase transition at $\lambda = 0$.

When $\sigma > 0$ while $\theta_t = \theta$ is kept fixed, Eq.(113) has one or two transition points corresponding to two extrema of its stationary distribution:

$$s_1 = 0, \quad s_2 = \kappa\theta - \nu\frac{\sigma^2}{2} \tag{114}$$

where $\nu = 2$ and $\nu = 1$ for the Ito and Stratonovich interpretation of SDE (113), respectively. The second transition point exists only if $\kappa\theta > \nu\frac{\sigma^2}{2}$. When this constraint is satisfied, the system (113) undergoes a noise-induced transition [24].

We can produce a few equivalent descriptions of the dynamics described by Eq.(112) by using changes of variable in this equation. In particular, if we define $s_t = 1/x_t$, then the stochastic differential equation for s_t using Ito's prescription reads

$$ds_t = \left(\kappa - (\kappa\theta - \sigma^2)s_t\right) dt - \sigma s_t dW_t \tag{115}$$

where now the drift becomes linear in the transformed variable $s_t = 1/x_t$.

Another useful form is obtained if instead we define $s_t = \log x_t/c$ where $c > 0$ is a fixed number having dimension of the currency of the market portfolio (e.g. the USD) that we need to introduce on the grounds of dimensionality analysis. For example, we can choose $c = \langle x\rangle$ to be a time-average value of x_t within an observation period. Using Ito's prescription with this choice of c, the SDE for $s_t = \log x_t/\langle x\rangle$ reads

$$ds_t = \kappa\left(\theta - \frac{\sigma^2}{2\kappa} - \langle x\rangle e^{s_t}\right) dt + \sigma dW_t \tag{116}$$

Note that with this form, the noise becomes additive rather than multiplicative as in Eqs.(112) or (115). On the other hand, the drift becomes exponential. It is easy to see that Eq.(116) requires the condition $2\kappa\theta > \sigma^2$ in order for Eq.(116) to have a stationary distribution.

Note that because x_t is a total market capitalization of a firm (or all firms in the index, depending on how we use the 1D setting here), $\log x_t$ will be given by a log-stock price plus a log of total number of shares outstanding. When the latter is constant, $s_t = \log x_t/c$ is equal to the log-price of the stock plus a constant term.

The GMR model (112) was used by Dixit and Pindyck [13], and its properties were further studied by Ewald and Yang [17] who have shown that this process is bounded, non-negative, and has a stationary distribution under the constraint $2\kappa\theta > \sigma^2$. Rather than introducing such mean-reverting dynamics phenomenologically, our model *derives* them (in a multi-variate setting) from an underlying dynamic optimization problem of a bounded-rational agent.

The non-stationary multivariate Geometric Mean reverting process (109) can be interpreted as either an equilibrium or quasi-equilibrium statistical process (which is the case usually assumed in econometric and financial models), or as an non-equilibrium Langevin process [55]. In the rest of this section, we assume the former setting, while some further comments on the latter case will be provided in Sect. 8.2.

6.2 IRL by Maximum Likelihood: market portfolio

Here we assume a quasi-equilibrium setting when the market manages to attain an equilibrium distribution (100) in each period, following changes in signals $\mathbf{z}_t$. In this case, standard statistical methods, such as Maximum Likelihood, can be applied to estimate the model. The negative log-likelihood function for observable data with this model reads

$$LL_M(\Theta) = -\log \prod_{t=0}^{T-1} \frac{1}{\sqrt{(2\pi)^N |\Sigma_x|}} e^{-\frac{1}{2}(\mathbf{v}_t)^T \Sigma_x^{-1}(\mathbf{v}_t)}, \;\; \mathbf{v}_t \equiv \frac{\mathbf{x}_{t+1} - \mathbf{x}_t}{\mathbf{x}_t} - \kappa \circ \left(\theta(\mathbf{z}_t) - \mathbf{x}_t\right) \Delta t \tag{117}$$

where $\mathbf{x}_t$ now stands for observed stock market prices, and $\Sigma_x = \sqrt{\Delta t}\Sigma_r$. Note that because the model is Markov, the product over $t = 0, \ldots, T-1$ does not necessarily mean a product of transitions along the same trajectory, but can be viewed as a product of T one-step transitions that do not correspond to consecutive time moments.

Parameters that can be estimated from data are therefore the vector of mean reversion speed parameters κ, factor loading matrix $\mathbf{w}$, and covariance matrix Σ_x.

Note that instead of defining the likelihood in terms of the original variables $\mathbf{x}_t$, we could defined it instead in terms of a transformed variable $\mathbf{s}_t = \log \mathbf{x}_t / \langle x \rangle$. The negative log-likelihood, when re-expressed in terms of the original observables $\mathbf{x}_t$ would then be of the same Gaussian form as in Eq.(117) where the variable $\mathbf{v}_t$ would be defined as

$$\mathbf{v}_t = \log \frac{\mathbf{x}_{t+1}}{\mathbf{x}_t} - \kappa \circ \left(\theta(\mathbf{z}_t) - \frac{\sigma^2}{2\kappa} - \mathbf{x}_t \right) \Delta t \tag{118}$$

7 Experiments

In this section we describe our experiments with the market model Eq.(109). Further details for calibrated model parameters are provided in Appendix C.

To show detailed results, we use the DJI index instead of the S&P500 index that is more commonly used as a market portfolio. We analyze the daily data on market caps of all firms in the DJI index from 2010 to the end of 2017. We use the current composition of DJI that includes Apple that was added in 2016. We re-scale all data points by dividing by the average total market cap of the index for the whole period, which is approximately equal to \$160Bn for our dataset.

Similar to [10], our approach takes signals $\mathbf{z}_t$ as given, and assumes that they are obtained through a search for 'alpha' that is beyond the scope of our framework. Calibrated model parameters will necessarily depend on the choice of predictors $\mathbf{z}_t$. One of our objectives here is to illustrate such dependence on the choice of signals.

To this end, we test our model using two sets of experiments with two different sets of predictors $\mathbf{z}_t$. We build both sets as predictors of market caps (or equivalently prices) rather than predictors of returns.

The first set of predictors includes two predictors for each stock: a perfect signal and a random signal. The perfect (oracle) signal is obtained as a (demeaned) realized next-day return. This test can serve as a sanity/implementation test for the model. It is expected to provide a stable calibration of parameters, nearly zero volatility, and an order of magnitude of difference between estimated weights of the perfect signal and the random signal. The results are as expected, see tables 1 and 2 in Appendix C where we show calibrated parameters for separate annual runs (we do not report weights to save space).

The second set of predictors are given by a pair of demeaned exponential moving averages of the (re-scaled) market caps. The two signals use parameters $\gamma = 0.9$ and $\gamma = 0.96$ of the exponential moving averaging, corresponding to the lookback windows of 7 days and 15 days, respectively.

In both sets of experiment, we estimate the resulting model parameters by minimizing the negative log-likelihood (117) subject to constraints of non-negativity of weights w_1, w_2 of two predictors, and adding a regularization term $\lambda(w_1 + w_2 - 1)^2$. While the results are only weakly dependent on the value of regularization parameter in the range $\lambda \sim 10^{-3} - 10^{-2}$, we report the results for the value of $\lambda = 10^{-2}$. The covariance matrix Σ_x is taken to be diagonal $\Sigma_x = \mathrm{diag}(\sigma_i^2)$. We set $\Delta t = 1$, thus we report daily rather than annualized values of κ and σ^2.

Calibrated parameters κ and σ^2 for the exponential moving average signals are shown in tables 3 and 4 in Appendix C. As could be expected, the resulting parameters are substantially different from those obtained with the first set of signals. Calibrated parameters are less stable, which is unsurprising given that moving averages are not very good predictors of future prices. In particular, we observe occasional negative values of κ that suggest a local divergence from a predicted value, rather than a convergence to this value. In Figs. 1, 2, 3 we show the market

cap vs a fitted mean level for the IBM, JPM and XOM stocks for a two-month period in 2017. Results obtained for other stocks and other periods are similar.

Note that it would be wrong to try to estimate parameter κ by simply running a regression of Δx_t on x_t and treating the signals $\mathbf{z}_t$ as a part of a noise term in such regression. As $\mathbf{z}_t$ is a random process itself, such procedure would violate the *i.i.d.* assumption for a noise term in such regression.

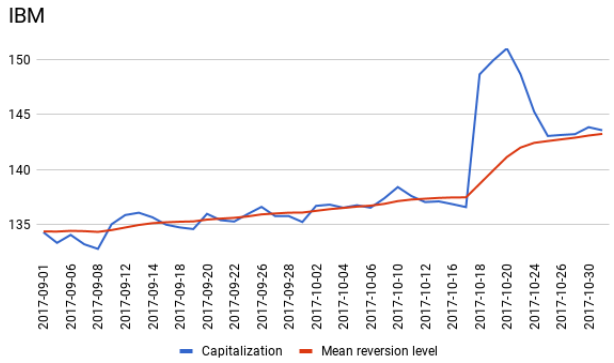

Figure 1: Market cap vs estimated mean level: IBM

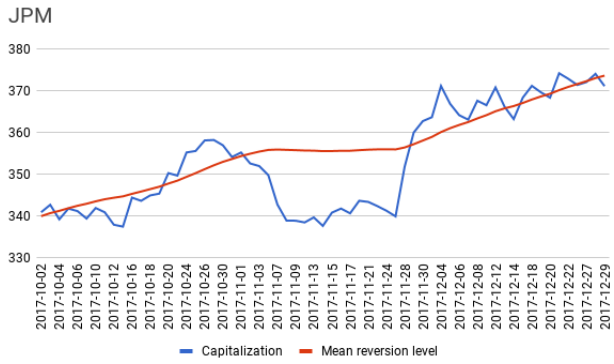

Figure 2: Market cap vs estimated mean level: JPM

8 Discussion and future directions

8.1 Mean reversion in asset returns

One of the most interesting implications of our model is its prediction of a non-linear mean-reverting behavior of asset returns. While mean reversion in intraday data for stock markets is a well established fact, its presence in longer-horizon returns is a topic of a long discussion in the literature. The latter started with Poterba and Summers who argued for mean reversion in stock returns as resulting from actions of 'noise traders' that do not have any objectives in trading, i.e. have zero intelligence [44]. Implications of mean reversion for a long-term optimal asset management were discussed in [50].

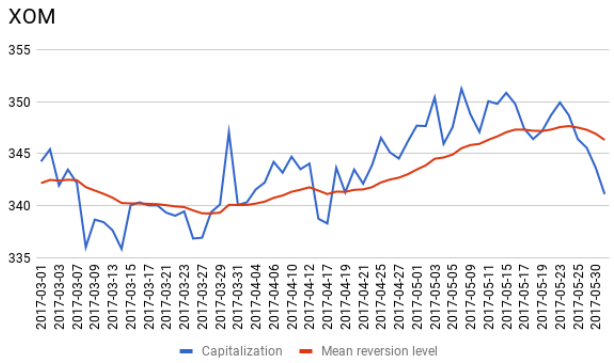

Figure 3: Market cap vs estimated mean level: XOM

In our model, mean reversion in asset returns has a very transparent origin. It results from a total market impact from traders that optimize their investment portfolios following mean-variance Markowitz-type optimization strategies by adapting to changing signals and changes in the market. The resulting stock price dynamics are of *non-linear mean reversion* type (a multi-variate Geometric Mean Reversion process with external factors), even though we start with a simple Gaussian policy π for the agent in our model. Non-linearity of the dynamics in our model is both a manifestation and a result of a feedback loop via the price impact mechanism.

Interestingly, the *dynamically* generated mean reversion in our model produces time-decaying auto-correlations in the system, i.e. multi-period effects that were absent in the original formulation[15]. Note that presence of slowly time-decaying auto-correlations and adaptivity to external signals are typical for self-organizing systems, see e.g. [57]. Therefore, our model demonstrates some features of a self-organizing behavior by the dynamic generation of a mean reversion level for stocks.

8.2 Non-equilibrium behavior and market crashes

As was discussed in Sect. 6.1, our setting above assumes that changes of external signals are slow enough, so that market has sufficient time to adjust to new information in signals $\mathbf{z}_t$ from one period to another. If external signals were just constant in time, the system would eventually settle in a stationary equilibrium state.

A different situation can occur if signals $\mathbf{z}_t$ exhibits a large jump at time t relative to their value at time $t-1$. In this case, the system can find itself trapped in a meta-stable state - a previously globally optimal state that becomes a local minimum following a jump of $\mathbf{z}_t$ to a new value. A meta-stability, rather than stability of this state is ensured by the presence of a potential barrier separating the global and local minima. A transition from a metastable state to a new dynamically optimal stable state would be activated by noise $\varepsilon_t^{(x)}$, see e.g. [55] on how such transitions are modeled in physics. In the financial setting, decays of such meta-stable states via a thermally-activated diffusion can describe market crashes. Such transitions can be studied either numerically using simulations, or theoretically using methods of [55]. Non-equilibrium phase transitions induced by multiplicative noise as in Eq.(111) were studied in [54]. Statistical

[15]In particular, our model did not originally include any "permanent impact" effect which may not *a priori* be an well-defined notion in an MDP setting.

physics of driven non-equilibrium dynamics of system with thermal fluctuations is studied in [43].

8.3 Multi-agent formulations: market-fitting or market-beating strategies?

As the objective of this paper was to make inference of a Bounded-Rational 'Invisible Hand' that drives the market as a whole using Inverse Reinforcement Learning, we used a single agent setting. In our formulation, this single agent self-learns by self-playing. As we showed in this paper, this formulation, though may appear somewhat abstract or even 'theological', gives rise to quite specific observable and computable consequences such as the prediction of mean reversion in asset returns, implied rationality and risks aversion parameters, and a market-implied optimal strategy.

On the other hand, it would be interesting to extend the setting of this model to a multi-agent formulation. On-line multi-agent Reinforcement Learning, where two or more bounded-rational agents implement Markowitz-like, or possibly more advanced investment strategies in a noisy market environment with external signals, can create potentially rich market-*beating* strategies.

8.4 Implied rationality of the market

Recall that the inverse temperature parameter β controls the degree of rationality of the RL agent that dynamically replicates the market portfolio by minimizing its trading cost. We have showed the result of calibration of the market model (109) implied by our framework. In this setting, the original model parameters are embedded in parameters defining Eq.(109), see Eqs.(108). The latter parameters are calibrated to market data.

To infer the original parameters of the model including β, one can instead use the IH-IF algorithm from Sect. 5.9. Inference of market-implied rationality parameter β and risk aversion λ will be addressed elsewhere.

8.5 The market as an information perception-action system

Analysis of the RL agent representing a coherent bounded-rational component of the market that we developed above included analysis of information costs of actions. This analysis can be extended by including information costs of information *extraction* [52, 42, 22, 48].

The value of this extension is in its focus on the external signals $\mathbf{z}_t$. In our model, we took them as given, effectively leaving the information costs of their *extraction* outside of the scope of the model. Analysis along the lines of [52, 42, 22, 48] allows one to assess the value of these signals for the *full* perception-action cycle. Note that traditionally, signals are accessed based on their ability to predict the future, e.g. their own future.

However, this is not the same as the ultimate goal of these signals, which is to improve rewards. A perception-action cycle analysis in [52, 42, 22] specifies *useful* information in signals, as opposed to *useless* information that should be discarded as its use amounts to a dissipated energy (heat) instead of an increase of the free energy. Extensions of the model developed in this paper along these lines of analysis of the perception-action cycle of financial markets will be presented elsewhere.

9 Summary

As was discussed e.g. by Sornette in [46], economic models differ from models in the physical sciences in that economic agents anticipate the future and act accordingly, thus impacting the

present. A value in finance depends on views of market participants on the future. This is very different from physics where quantities such as e.g. the mass of a proton are clearly independent of public views on the future. Such observations led many researchers to suggest that ideas from biology and genetics can be useful for financial modeling [46].

As we discussed in Sect. 2, our model shares a number of similarities with models in biology, e.g. [19], [39]. Our bounded-rational market-wide agent aggregates all traders in the market who anticipate the future in their trading decisions. Optimal actions of the agent are those that maximize its free energy, similar to models of [19], [39].

Our model provides a computational scheme based on Inverse Reinforcement Learning and the variational EM algorithm to infer parameters of the model. As in our model the market-wide agent that implements the 'Invisible Hand' is a *sum* of all agents, it provides a unifying framework for inference of either a market portfolio or a single investor. Furthermore, for the most interesting case of a dynamic inference of the market portfolio, our model provides a multi-period extension of the Black-Litterman model [8]. Finally, our approach suggests a non-stationary multivariate Geometric Mean Reversion (GMR) process (109) as a model for market dynamics.

Appendix A: Optimal action and optimal G-function with locally-quadratic expansion

A.1 Linearization of dynamics

Here we develop a tractable computational scheme based on conditioning on the linearization variables $\bar{\mathbf{a}}_t,\ \bar{\mathbf{y}}_t$, and expanding the dynamics and functions of interest (the G-function and the action policy π_θ) in Taylor series in small deviations from these values.

In this section we use the symbols $\bar{\mathbf{a}}_t,\ \bar{\mathbf{y}}_t$ as fixed conditioning values in calculation of conditional variational free energy (71), or equivalently as *realizations* of random hidden variables $\bar{\mathbf{a}}_t,\ \bar{\mathbf{y}}_t$. Note that when these values are fixed, we also have fixed values of a related pair $(\bar{\mathbf{u}}_t, \bar{\mathbf{x}}_t) \equiv (\mathbf{1}_{-1}^T \bar{\mathbf{a}}_t, \mathbf{1}_0^T \bar{\mathbf{y}}_t)$, where $\mathbf{1}_0 = [\mathbf{1}, \mathbf{0}]^T$ and $\mathbf{1}_{-1} = [\mathbf{1}, -\mathbf{1}]^T$.

Let us come back to Eq.(78) that shows that the dynamics are non-linear in controls $\mathbf{u}_t$ and the state vector $\mathbf{y}_t$. Define deviations $\delta\mathbf{x}_t$ and $\delta\mathbf{u}_t$ from linearization points in the $(\mathbf{x}, \mathbf{u})$ space:

$$\mathbf{x}_t = \bar{\mathbf{x}}_t + \delta\mathbf{x}_t,\ \mathbf{u}_t = \bar{\mathbf{u}}_t + \delta\mathbf{u}_t \tag{A.1}$$

We linearize the dynamics equation (78) by keeping linear terms in deviations $\delta\mathbf{x}_t,\ \delta\mathbf{u}_t$. This yields

$$\delta\mathbf{x}_{t+1} = \Omega_0 + \Omega_x \delta\mathbf{x}_t + \Omega_u \delta\mathbf{u}_t + \Omega_z \delta\mathbf{z}_t + \varepsilon_t \circ (\mathbf{x}_t + \mathbf{u}_t) \tag{A.2}$$

where

$$\begin{aligned}
\Omega_0 &= (1 + r_f + \mathrm{diag}\,(\mathbf{W}\bar{\mathbf{z}}_t - \mathbf{M}\bar{\mathbf{u}}_t))\,(\bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t) - \bar{\mathbf{x}}_{t+1} \\
\Omega_x &= 1 + r_f + \mathrm{diag}\,(\mathbf{W}\bar{\mathbf{z}}_t - \mathbf{M}\bar{\mathbf{u}}_t) \\
\Omega_u &= 1 + r_f + \mathrm{diag}\,(\mathbf{W}\bar{\mathbf{z}}_t - \mathbf{M}\bar{\mathbf{u}}_t) - (\bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t) \circ \mathbf{M} \\
\Omega_z &= (\bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t) \circ \mathbf{W}
\end{aligned} \tag{A.3}$$

Here $(\bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t) \circ \mathbf{M}$ stands for an element-wise multiplication of a k-th component of vector $(\bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t)$ with a k-th row of matrix $\mathbf{M}$, and a similar convention is used in the last relation.

Deviations can also be defined for the extended space (14). In this case, we expand around conditioning values $\bar{\mathbf{a}}_t, \bar{\mathbf{y}}_t$ in a similar way to Eq.(A.1):

$$\mathbf{y}_t = \bar{\mathbf{y}}_t + \delta\mathbf{y}_t, \quad \mathbf{a}_t = \bar{\mathbf{a}}_t + \delta\mathbf{a}_t \tag{A.4}$$

so that linearization points in Eqs.(A.1) and (A.4) are related as follows: $(\bar{\mathbf{u}}_t, \bar{\mathbf{x}}_t) \equiv (\mathbf{1}_{-1}^T \bar{\mathbf{a}}_t, \mathbf{1}_0^T \bar{\mathbf{y}}_t)$.

Stacking Eq.(A.2) and Eq.(13) written in terms of the increment $\delta\mathbf{z}_t$ together, we can write a linearized equation for $\delta\mathbf{y}_t$ as follows:

$$\delta\mathbf{y}_{t+1} = \Psi_0 + \Psi_y \delta\mathbf{y}_t + \Psi_a \delta\mathbf{a}_t + \varepsilon_t^y \left(\delta\mathbf{y}_t, \delta\mathbf{a}_t\right) \tag{A.5}$$

where

$$\begin{aligned}
&\Psi_0 = \left[\begin{array}{c} \Omega_0 \\ (\mathbf{I} - \Phi) \circ \bar{\mathbf{z}}_t - \bar{\mathbf{z}}_{t+1} \end{array}\right], \; \Psi_y = \left[\begin{array}{cc} \Omega_x & \Omega_z \\ 0 & \mathbf{I} - \Phi \end{array}\right], \; \Psi_a = \left[\begin{array}{c} \Omega_u \mathbf{1}_{-1}^T \\ 0 \end{array}\right] \\
&\varepsilon_t^y \left(\delta\mathbf{y}_t, \delta\mathbf{a}_t\right) = \left[\begin{array}{c} \varepsilon_t \circ (\mathbf{x}_t + \mathbf{u}_t) \\ \varepsilon_t^z \end{array}\right] = \left[\begin{array}{c} \varepsilon_t \circ \left(\mathbf{1}_0^T \mathbf{y}_t + \mathbf{1}_{-1}^T \mathbf{a}_t\right) \\ \varepsilon_t^z \end{array}\right]
\end{aligned} \tag{A.6}$$

Note that matrices Ψ_0, Ψ_y, Ψ_a implicitly depend on time via their dependence of $\bar{\mathbf{y}}_t$ and $\bar{\mathbf{a}}_t$. Also note that Eq.(A.5) implies that

$$\begin{aligned}
&\widehat{\delta\mathbf{y}}_{t+1} \equiv \mathbb{E}_{t,a}\left[\delta\mathbf{y}_{t+1}\right] = \Psi_0 + \Psi_y \delta\mathbf{y}_t + \Psi_a \delta\mathbf{a}_t \\
&\Sigma_y \equiv \mathrm{Cov}\left[\delta\mathbf{y}_{t+1}\right] = \left[\begin{array}{cc} \Sigma_{xx} & 0 \\ 0 & \Sigma_z \end{array}\right] \\
&\Sigma_{xx} = \Sigma_r \circ \left[\left(\mathbf{1}_0^T \mathbf{y}_t + \mathbf{1}_{-1}^T \mathbf{a}_t\right)\left(\mathbf{1}_0^T \mathbf{y}_t + \mathbf{1}_{-1}^T \mathbf{a}_t\right)^T\right]
\end{aligned} \tag{A.7}$$

We can also express the reward Eq.(22) in terms of $\delta\mathbf{y}_t$ and $\delta\mathbf{a}_t$:

$$\hat{R}_t(\mathbf{y}_t, \mathbf{a}_t) = \delta\mathbf{a}_t^T \hat{\mathbf{R}}_{aa} \delta\mathbf{a}_t + \delta\mathbf{y}_t^T \hat{\mathbf{R}}_{yy} \delta\mathbf{y}_t + \delta\mathbf{a}_t^T \hat{\mathbf{R}}_{ay} \delta\mathbf{y}_t + \delta\mathbf{a}_t^T \hat{\mathbf{R}}_a + \delta\mathbf{y}_t^T \hat{\mathbf{R}}_y + r(\bar{\mathbf{y}}_t, \bar{\mathbf{a}}_t) \tag{A.8}$$

Here we defined

$$\begin{aligned}
&\hat{\mathbf{R}}_{aa} = \mathbf{R}_{aa}, \; \hat{\mathbf{R}}_{yy} = \mathbf{R}_{yy}, \; \hat{\mathbf{R}}_{ay} = \mathbf{R}_{ay}, \\
&\hat{\mathbf{R}}_a = \mathbf{R}_a + 2\mathbf{R}_{aa}\bar{\mathbf{a}}_t + \mathbf{R}_{ay}\bar{\mathbf{y}}_t, \; \hat{\mathbf{R}}_y = 2\mathbf{R}_{yy}\bar{\mathbf{y}}_t + \mathbf{R}_{ay}^T \bar{\mathbf{a}}_t \\
&r(\bar{\mathbf{y}}_t, \bar{\mathbf{a}}_t) = \bar{\mathbf{a}}_t^T \mathbf{R}_{aa} \bar{\mathbf{a}}_t + \bar{\mathbf{y}}_t^T \mathbf{R}_{yy} \bar{\mathbf{y}}_t + \bar{\mathbf{a}}_t^T \mathbf{R}_{ay} \bar{\mathbf{y}}_t + \bar{\mathbf{a}}_t^T \mathbf{R}_a
\end{aligned} \tag{A.9}$$

Recall that as the original parameters of the reward function coefficients $\mathbf{R}_{aa}$ etc. were defined in terms of the original model parameters, the new 'hat' coefficients $\hat{\mathbf{R}}_{aa}$ etc. are now functions of the original model parameters and conditioning variables $\bar{\mathbf{y}}_t$, $\bar{\mathbf{a}}_t$.

A.2 Recursion for the G-function

In this section, we consider a finite-horizon setting. In this case, a time dependence of coefficients will be implicit in equations to follow, and will be supplemented by an additional upper script, e.g. $\mathbf{F}_{yy}^{(t)}$, where needed for clarity.

For a finite-horizon setting with a planning horizon T, as positions $\mathbf{x}_T$ are fixed by (10), we can use Eqs.(47) and (A.8) to get

$$F_T^\pi(\mathbf{y}_T) = \hat{R}_T(\bar{\mathbf{y}}_T + \delta\mathbf{y}_T, \bar{\mathbf{a}}_T + \delta\mathbf{a}_T) \tag{A.10}$$

We use this to fix $\mathbf{F}_{yy}$, $\mathbf{F}_y$, $F_0(\bar{\mathbf{y}}_t)$ in Eq.(87) in terms of coefficient of reward function (A.8):

$$\begin{aligned}
\mathbf{F}_{yy}^{(T)} &= \hat{\mathbf{R}}_{yy} = \mathbf{R}_{yy} \\
\mathbf{F}_{y}^{(T)} &= \hat{\mathbf{R}}_{ay}^T \delta\mathbf{a}_T + \hat{\mathbf{R}}_y = \mathbf{R}_{ay}^T \left(\bar{\mathbf{a}}_T + \delta\mathbf{a}_T\right) + 2\mathbf{R}_{yy}\bar{\mathbf{y}}_T \\
F_0(\bar{\mathbf{y}}_T, \bar{\mathbf{a}}_T) &= \delta\mathbf{a}_T \hat{\mathbf{R}}_{aa} \delta\mathbf{a}_T + \delta\mathbf{a}_T^T \hat{\mathbf{R}}_a + r(\bar{\mathbf{y}}_T, \bar{\mathbf{a}}_T)
\end{aligned} \tag{A.11}$$

For values $t = T-1, \ldots, 0$, we use Eqs.(A.5) and (A.7) to compute the conditional expectation of the next-period F-function as follows:

$$\mathbb{E}_{t,\mathbf{a}}\left[F_{t+1}^{\pi}(\mathbf{y}_{t+1})\right] = F_0(\bar{\mathbf{y}}_{t+1}, \bar{\mathbf{a}}_{t+1}) + \widehat{\delta\mathbf{y}}_{t+1}^T \mathbf{F}_y^{(t+1)} + \widehat{\delta\mathbf{y}}_{t+1}^T \mathbf{F}_{yy}^{(t+1)} \widehat{\delta\mathbf{y}}_{t+1} + \mathrm{Tr}\left[\mathbf{F}_{yy}^{(t+1)} \Sigma_y\right] \tag{A.12}$$

The last term can be expressed in a more convenient form using Eq.(88):

$$\begin{aligned}
\mathrm{Tr}\left[\mathbf{F}_{yy}^{(t+1)} \Sigma_y\right] &= \mathrm{Tr}\left[\left(\left(\mathbf{1}_0^T \mathbf{y}_t + \mathbf{1}_{-1}^T \mathbf{a}_t\right)\left(\mathbf{1}_0^T \mathbf{y}_t + \mathbf{1}_{-1}^T \mathbf{a}_t\right)^T\right)\left(\mathbf{F}_{xx} \circ \Sigma_r\right)\right] + \mathrm{Tr}\left[\mathbf{F}_{zz}\Sigma_z\right] \\
&= \left(\mathbf{1}_0^T \mathbf{y}_t + \mathbf{1}_{-1}^T \mathbf{a}_t\right)^T \left(\mathbf{F}_{xx} \circ \Sigma_r\right)\left(\mathbf{1}_0^T \mathbf{y}_t + \mathbf{1}_{-1}^T \mathbf{a}_t\right) + \mathrm{Tr}\left[\mathbf{F}_{zz}\Sigma_z\right]
\end{aligned} \tag{A.13}$$

After some algebra, we put Eq.(A.12) in a form similar to Eq.(A.8):

$$\mathbb{E}_{t,\mathbf{a}}\left[F_{t+1}^{\pi}(\mathbf{y}_{t+1})\right] = \delta\mathbf{a}_t^T \mathbf{H}_{aa} \delta\mathbf{a}_t + \delta\mathbf{y}_t^T \mathbf{H}_{yy} \delta\mathbf{y}_t + \delta\mathbf{a}_t^T \mathbf{H}_{ay} \delta\mathbf{y}_t + \delta\mathbf{a}_t^T \mathbf{H}_a + \delta\mathbf{y}_t^T \mathbf{H}_y + \widehat{f}(\bar{\mathbf{y}}_t, \bar{\mathbf{a}}_t) \tag{A.14}$$

where

$$\begin{aligned}
\mathbf{H}_{aa} &= \Psi_a^T \mathbf{F}_{yy} \Psi_a + \mathbf{1}_{-1}\left(\mathbf{F}_{xx} \circ \Sigma_r\right)\mathbf{1}_{-1}^T \\
\mathbf{H}_{yy} &= \Psi_y^T \mathbf{F}_{yy} \Psi_y + \mathbf{1}_0\left(\mathbf{F}_{xx} \circ \Sigma_r\right)\mathbf{1}_0^T \\
\mathbf{H}_{ay} &= 2\Psi_a^T \mathbf{F}_{yy} \Psi_y + 2 \cdot \mathbf{1}_{-1}\left(\mathbf{F}_{xx} \circ \Sigma_r\right)\mathbf{1}_0^T \\
\mathbf{H}_a &= \Psi_a^T \mathbf{F}_y + 2\Psi_a^T \mathbf{F}_{yy} \Psi_0 + 2 \cdot \mathbf{1}_{-1}\left(\mathbf{F}_{xx} \circ \Sigma_r\right)\left(\mathbf{1}_0^T \bar{\mathbf{y}}_t + \mathbf{1}_{-1}^T \bar{\mathbf{a}}_t\right) \\
\mathbf{H}_y &= \Psi_y^T \mathbf{F}_y + 2\Psi_y^T \mathbf{F}_{yy} \Psi_0 + 2 \cdot \mathbf{1}_0\left(\mathbf{F}_{xx} \circ \Sigma_r\right)\left(\mathbf{1}_0^T \bar{\mathbf{y}}_t + \mathbf{1}_{-1}^T \bar{\mathbf{a}}_t\right) \\
\widehat{f}(\bar{\mathbf{y}}_t, \bar{\mathbf{a}}_t) &= F_0(\bar{\mathbf{y}}_{t+1}, \bar{\mathbf{a}}_{t+1}) + \Psi_0^T \mathbf{F}_y + \Psi_0^T \mathbf{F}_{yy} \Psi_0 \\
&+ \left(\mathbf{1}_0^T \bar{\mathbf{y}}_t + \mathbf{1}_{-1}^T \bar{\mathbf{a}}_t\right)^T \left(\mathbf{F}_{xx} \circ \Sigma_r\right)\left(\mathbf{1}_0^T \bar{\mathbf{y}}_t + \mathbf{1}_{-1}^T \bar{\mathbf{a}}_t\right) + \mathrm{Tr}\left[\mathbf{F}_{zz}\Sigma_z\right]
\end{aligned} \tag{A.15}$$

These equations can be used for both the finite-horizon and infinite-horizon settings. For the former case, all parameters in the right-hand sides of Eqs.(A.15) refer to the future time moment $t+1$, so that Eqs.(A.15) serve as a part of a backward recursion scheme to be completed below. On the other hand, for an infinite-horizon case, they can be used as updates equations for time-independent parameters of the free energy function (87).

Next we take the Bellman equation for the G-function

$$G_t^{\pi}(\mathbf{y}_t, \mathbf{a}_t) = \hat{R}_t(\mathbf{y}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}}\left[F_{t+1}^{\pi}(\mathbf{y}_{t+1})\right] \tag{A.16}$$

where we substitute Eqs.(83), (A.8) and (A.14). Equating coefficients in front of like powers of $\delta\mathbf{x}_t$ and $\delta\mathbf{a}_t$ in the left-hand side and the right-hand side of the resulting equation, we get a set of recursive relations for matrix coefficients defining the G-function in Eq.(83):

$$\begin{aligned}
\mathbf{G}_{aa} &= \hat{\mathbf{R}}_{aa} + \mathbf{H}_{aa},\ \mathbf{G}_{yy} = \hat{\mathbf{R}}_{yy} + \mathbf{H}_{yy},\ \mathbf{G}_{ay} = \hat{\mathbf{R}}_{ay} + \mathbf{H}_{ay} \\
\mathbf{G}_a &= \hat{\mathbf{R}}_a + \mathbf{H}_a,\ \mathbf{G}_y = \hat{\mathbf{R}}_y + \mathbf{F}_y,\ g(\bar{\mathbf{y}}_t, \bar{\mathbf{a}}_t) = r(\bar{\mathbf{y}}_t, \bar{\mathbf{a}}_t) + \widehat{f}(\bar{\mathbf{y}}_t, \bar{\mathbf{a}}_t)
\end{aligned} \tag{A.17}$$

In these equations, coefficients in the left-hand side and the right-hand side refer to the same time t, therefore they can be used in the same way for both the finite- and infinite horizon cases.

A.3 Backward recursion with observable rewards

We first consider a complete backward recursion scheme for a finite-horizon case, that would apply if rewards were observed. Below, we will modify this scheme to replace observed rewards by their estimated values. In both cases, Eqs.(A.17) should be solved by backward recursion, starting at the planning horizon T with a terminal condition.

For an arbitrary time step $t < T$, we proceed as follows. First, we use Eqs.(A.17) to obtain parameters of the Q-function at time t . Note that parameters entering the right-hand of Eqs.(A.17) are known at time t, as they are computed using the values defined at time step $t + 1$.

Second, we use the computed Q-function as parametrized by Eq.(83) to compute the F-function at time t according to Eq.(44). To this end, we express the prior π_0 in Eq.(29) in terms of increments $\delta\mathbf{a}_t$ with the mean $\widehat{\delta\mathbf{a}}_t = \hat{\mathbf{a}}_t - \bar{\mathbf{a}}_t$ (recall that we condition on the value of $\bar{\mathbf{a}}_t$):

$$\pi_0(\delta\mathbf{a}_t|\mathbf{y}_t) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_p|}} \exp\left(-\frac{1}{2}\left(\delta\mathbf{a}_t - \widehat{\delta\mathbf{a}}_t\right)^T \Sigma_p^{-1}\left(\mathbf{a}_t - \widehat{\delta\mathbf{a}}_t\right)\right) \tag{A.18}$$

where

$$\widehat{\delta\mathbf{a}}_t = \hat{\mathbf{a}}_t - \bar{\mathbf{a}}_t = \hat{\mathbf{A}}_0 + \hat{\mathbf{A}}_1\left(\bar{\mathbf{y}}_t + \delta\mathbf{y}_t\right) - \bar{\mathbf{a}}_t \tag{A.19}$$

Using this in Eq.(44) along with we Eqs.(A.17) and replacing a discrete sum by an integral[16], we obtain

$$\begin{aligned} F_t^\pi(\mathbf{y}_t) &= \frac{1}{\beta}\log Z_t = \frac{1}{\beta}\log\sum_{\delta\mathbf{a}_t}\pi_0(\bar{\mathbf{a}}_t + \delta\mathbf{a}_t|\mathbf{y}_t)e^{\beta G_t^\pi(\mathbf{y}_t,\mathbf{a}_t)} \\ &= \frac{1}{\beta}\left[-\frac{N_a}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_p| + \beta\delta\mathbf{y}_t^T\mathbf{G}_{yy}\delta\mathbf{y}_t + \beta\delta\mathbf{y}_t^T\mathbf{G}_y + \beta g(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t)\right. \\ &\left. -\frac{1}{2}\widehat{\delta\mathbf{a}}_t^T\Sigma_p^{-1}\widehat{\delta\mathbf{a}}_t + \log\int d\,\mathbf{a}e^{-\frac{1}{2}\mathbf{a}^T\left(\Sigma_p^{-1}-2\beta\mathbf{G}_{aa}\right)\mathbf{a}+\mathbf{a}^T\left(\Sigma_p^{-1}\widehat{\delta\mathbf{a}}_t+\beta\mathbf{G}_{ay}\delta\mathbf{y}_t+\beta\mathbf{G}_a\right)}\right] \end{aligned} \tag{A.20}$$

To simplify formulae below, we introduce auxiliary quantities

$$\begin{aligned} &\mathbf{b}_t = \bar{\mathbf{a}}_t - \hat{\mathbf{A}}_0 - \hat{\mathbf{A}}_1\bar{\mathbf{y}}_t,\ \tilde{\Sigma}_p = \Sigma_p^{-1} - 2\beta\mathbf{G}_{aa}, \\ &\Gamma_\beta = \frac{1}{\beta}\left(\mathbf{I} - \left(\Sigma_p^{-1}\right)^T\tilde{\Sigma}_p^{-1}\right)\Sigma_p^{-1},\ \Upsilon_\beta = \tilde{\Sigma}_p^{-1}\Sigma_p^{-1} \\ &\mathbf{E}_{ay} = \Upsilon_\beta\hat{\mathbf{A}}_1 + \frac{1}{2}\beta\tilde{\Sigma}_p^{-1}\mathbf{G}_{ay},\ \mathbf{D}_{ay} = \mathbf{G}_{ay}^T\Upsilon_\beta - \hat{\mathbf{A}}_1^T\Gamma_\beta \\ &\mathbf{E}_a = \hat{\mathbf{A}}_1^T\Upsilon_\beta\mathbf{G}_a + \beta\mathbf{G}_{ay}^T\tilde{\Sigma}_p^{-1}\mathbf{G}_a,\ \mathcal{L}_\beta = \frac{1}{2\beta}\left(\log|\Sigma_p| + \log\left|\tilde{\Sigma}_p\right|\right) \end{aligned} \tag{A.21}$$

Note that $\lim_{\beta\to 0}\Gamma_\beta = 0$ and $\lim_{\beta\to 0}\Upsilon_\beta = 1$. Using Eqs.(A.21) for the Gaussian integral (A.20), we can express it as in the same form as in Eq.(87):

$$F_t^\pi(\mathbf{y}_t) = \delta\mathbf{y}_t^T\mathbf{F}_{yy}\delta\mathbf{y}_t + \delta\mathbf{y}_t^T\mathbf{F}_y + F_0(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t) \tag{A.22}$$

where the coefficients are now *computed* as follows:

$$\begin{aligned} &\mathbf{F}_{yy} = \mathbf{G}_{yy} + \mathbf{G}_{ay}^T\mathbf{E}_{ay} - \frac{1}{2}\hat{\mathbf{A}}_1^T\Gamma_\beta\hat{\mathbf{A}}_1 \\ &\mathbf{F}_y = \mathbf{G}_y - \mathbf{D}_{ay}\mathbf{b}_t + \hat{\mathbf{A}}_1^T\Upsilon_\beta\mathbf{G}_a + \beta\mathbf{G}_{ay}^T\tilde{\Sigma}_p^{-1}\mathbf{G}_a \\ &F_0(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t) = g(\bar{\mathbf{y}}_t,\bar{\mathbf{a}}_t) - \frac{1}{2}\mathbf{b}_t^T\Gamma_\beta\mathbf{b}_t - \mathbf{G}_a^T\Upsilon_\beta\mathbf{b}_t + \frac{\beta}{2}\mathbf{G}_a^T\tilde{\Sigma}_p^{-1}\mathbf{G}_a - \mathcal{L}_\beta \end{aligned} \tag{A.23}$$

[16]Recall that we used a discrete notation for convenience only, while working in fact in a continuous-action formulation.

Appendix B: IRL for a single investor case

In this appendix, we consider the case of a single investor with observable actions as a special case of our model. To recall, in this case, we build a probabilistic model of a specific trader, assuming that we have access to trader's trading record. This model is given by the Gaussian policy of Eq.(100) where the mean and variance in Eq.(98) are computed using trader's trading data, interpreted as trader's observed actions $\mathbf{a}_t$.

A major simplification of the single investor inference in our model is that when actions are observed, we do not need an inner integral over $\mathbf{a}_t$ in Eq.(62). The only integration that we need in this case is the outer integration over $\bar{\mathbf{a}}_t$.

For such setting with investor-specific actions and rewards, estimation of parameters of Eq.(A.8) amounts to the EM algorithm with the free energy for a set of N_b trajectories of length T of the following form (compare with Eq.(62))

$$\mathcal{F}_s(\mathbf{w},\theta) = \sum_{b=1}^{N_b}\sum_{t=0}^{T}\int d\bar{\mathbf{a}}_t\, q_{\bar{a}}(\bar{\mathbf{a}}_t|\mathbf{y},\mathbf{w}) \log \frac{\pi_\theta\left(\mathbf{a}_t|\mathbf{y}_t\right) p_\theta\left(\mathbf{y}_{t+1},|\mathbf{y}_t,\mathbf{a}_t\right)}{q_{\bar{a}}(\bar{\mathbf{a}}_t|\mathbf{y},\mathbf{w})}$$

where $\mathbf{y}_t$ and $\mathbf{a}_t$ stands for observed values of investments, signals and trades in the investor portfolio, stored as a historical dataset, and conditional transition probability $p_\theta\left(\mathbf{y}_{t+1}|\mathbf{y}_t,\mathbf{a}_t\right)$ is defined in Eq.(76).

The complete variational EM IRL algorithm for a single investor is given by Algorithm **??**. In step 1, the expectation of the next-time F-function is computed using Eq.(A.14) within a backward recursion that starts with a fixed terminal condition at time $t = T$. In step 2, we compute the reward using Eq.(A.8). In step 3, an update of the Q-function is performed using Eq.(A.17). The time-t F-function is computed in step 4 using Eq.(A.23). Finally, in step 5, the optimal policy as a function of θ is recomputed using Eq.(100). Computing these quantities for all transitions in the mini-batch, we obtain the free energy (102) for the mini-batch. This is used to produce an update of the current estimation of θ using a learning rate α_θ. The new updated values of θ are then used to update parameters $\hat{\mathbf{A}}_1^{(k)}$, $\hat{\mathbf{A}}_1^{(k)}$, $\Sigma_p^{(k)}$ of the policy π_θ. Then the algorithm proceeds to the next iteration.

Data: a sequence of states and signals
Result: the reward function, optimal policy, and value function
Set the learning rates α_θ, α_ω, batch size N_b, initial parameters $\theta^{(0)}$, $\omega^{(0)}$, $\hat{\mathbf{A}}_0^{(0)}$, $\hat{\mathbf{A}}_1^{(0)}$, $\Sigma_p^{(0)}$
Set $k = 1$
while *not converged* **do**
  Draw a new mini-batch of N_b T-step trajectories $(\mathbf{y}_t, \ldots, \mathbf{y}_{t+T})$
  E-step:
  Compute the free energy $\mathcal{F}_s(\omega, \theta^{(k-1)})$ of the mini-batch using Eq.(B.1)
  Update recognition model parameters $\omega^{(k)} = (1 - \alpha_\omega)\omega^{(k-1)} + \alpha_\omega \frac{\partial}{\partial \omega} \mathcal{F}_s(\omega, \theta^{(k-1)})$
  M-step: Maximize $\mathcal{F}_s(\omega^{(k)}, \theta)$ as a function of θ:
  for *each transition* $(\mathbf{y}_t, \mathbf{y}_{t+1})$ *for* $t = T-1, \ldots, 0$ **do**
    1. Compute the expected value at time t of the F-function at time $t+1$.
    2. Compute the reward as a function of θ.
    3. Use steps 1 and 2 to update the Q-function at time t
    4. Compute the value of the F-function at time t.
    5. Recompute the policy distribution $\pi_\theta(\mathbf{a}_t | t, \mathbf{y}_t)$ as a function of θ by updating its mean and variance.
  end
  Compute the free energy $\mathcal{F}_s(\omega^{(k)}, \theta)$ of the mini-batch using Eq.(B.1)
  Update the parameter vector $\theta^{(k)} = (1 - \alpha_\theta)\theta^{(k-1)} + \alpha_\theta \frac{\partial}{\partial \theta} \mathcal{F}_s(\omega^{(k)}, \theta)$
  Use the new value $\theta^{(k)}$ to compute $\hat{\mathbf{A}}_1^{(k)}$, $\hat{\mathbf{A}}_1^{(k)}$, $\Sigma_p^{(k)}$
  Increment $k = k + 1$
end

Algorithm 2: IRL algorithm that learns the optimal policy, reward, and value function for a single investor.

Appendix C: Calibration results for the DJI portfolio

Here we report results of Maximum Likelihood estimation of the market model (109) for two sets of signals described in Sect.(7). We show the results for the calibrated daily mean reversion parameter κ and variance $\Sigma = \sigma^2$ in Eq.(109). Fitted weights of the signals are not shown to save space.

	2010	2011	2012	2013	2014	2015	2016	2017
AAPL	0.7006	0.4707	0.3024	0.3621	0.2846	0.2403	0.2875	0.2036
AXP	3.2127	-0.0447	2.5010	2.0031	1.7278	2.0213	2.6951	2.1649
BA	3.3122	3.2815	2.9620	2.0125	1.7078	1.6523	1.9196	1.2189
CAT	3.9048	2.6793	2.6194	2.8312	2.6242	3.4815	3.6305	2.4664
CSCO	1.1863	1.6795	1.6632	1.3183	1.3022	1.1573	1.1861	0.9684
CVX	1.0389	0.8057	0.7649	0.6913	0.7266	0.9405	0.9175	0.7601
DIS	2.4381	2.3823	1.9216	1.4093	1.1324	0.8764	1.0116	0.9794
DWDP	5.0047	-0.0785	4.2726	3.6760	2.7862	2.9058	2.8816	1.9451
GE	0.9095	0.8770	0.7563	0.6556	0.6221	0.6057	0.5860	0.7540
GS	1.9353	2.7332	2.9583	2.2502	2.0742	1.9744	2.3197	1.7672
HD	3.0667	2.9361	1.9388	1.4898	1.3529	1.0774	1.0159	0.8540
IBM	0.9639	0.7677	0.7085	0.7414	0.8652	1.0601	1.1543	1.0965
INTC	1.3863	1.3800	1.2918	1.3920	1.0265	1.0671	1.0228	0.8517
JNJ	0.9459	0.8814	0.8717	0.6484	0.5680	0.5876	0.5311	0.4577
JPM	1.0068	1.1603	1.0930	0.8261	0.7321	0.6890	0.7049	0.4972
KO	1.2571	1.0468	0.9509	0.9079	0.8886	0.8958	0.8481	0.8447
MCD	-0.0194	1.8295	1.7101	1.6392	1.7204	1.7170	1.5119	1.3050
MMM	2.7017	2.7410	2.6042	2.0222	1.7473	1.6580	1.6336	1.2609
MRK	1.4462	1.5575	1.2520	1.1635	0.9826	1.0299	1.0075	0.9625
MSFT	0.6883	0.7343	0.6487	0.5723	0.4595	0.4326	0.3776	0.2787
NKE	5.4456	4.9397	-0.0221	3.6118	2.8779	2.0821	2.0890	2.1891
PFE	1.1989	1.0767	0.9167	0.7890	0.8235	0.7841	0.8232	0.7923
PG	0.9066	0.9232	0.8940	0.7440	0.7239	0.7564	0.7194	0.7010
TRV	6.3443	-0.0210	6.4233	5.1396	5.0635	4.8245	4.9080	4.6598
UNH	4.2802	3.2558	2.8095	2.3977	1.9617	1.4535	1.2777	0.8714
UTX	2.4634	2.3068	2.2702	1.7594	1.5741	1.6942	1.9644	1.7134
V	4.1773	3.7475	2.4586	1.6999	1.4692	1.1614	1.1130	0.8799
VZ	1.8906	1.5762	1.3284	1.1525	1.1527	0.8486	0.7779	0.8242
WMT	0.8198	0.8641	0.7082	0.6513	0.6365	0.6957	0.7640	0.6456
XOM	0.5292	0.4262	0.4018	0.3992	0.3930	0.4779	0.4671	0.4656

Table 1: Calibrated κ for a combination of a "perfect signal" and a "noise signal"

	2010	2011	2012	2013	2014	2015	2016	2017
AAPL	0.0001	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
AXP	0.0001	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
BA	0.0001	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
CAT	0.0001	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0001
CSCO	0.0001	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
CVX	0.0000	0.0002	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
DIS	0.0000	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
DWDP	0.0001	0.0020	0.0001	0.0000	0.0000	0.0000	0.0000	0.0001
GE	0.0001	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
GS	0.0001	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
HD	0.0001	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
IBM	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
INTC	0.0001	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
JNJ	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
JPM	0.0001	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
KO	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MCD	0.0002	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MMM	0.0000	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MRK	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MSFT	0.0001	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
NKE	0.0001	0.0002	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000
PFE	0.0001	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PG	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
TRV	0.0000	0.0015	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
UNH	0.0001	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
UTX	0.0000	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
V	0.0001	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
VZ	0.0000	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000
WMT	0.0000	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
XOM	0.0001	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 2: Calibrated values of σ^2 for a combination of a "perfect signal" and a "noise signal"

	2010	2011	2012	2013	2014	2015	2016	2017
AAPL	-0.0122	0.0288	-0.0041	0.0047	-0.0038	0.0080	0.0045	-0.0038
AXP	0.1524	0.1748	0.1317	-0.0354	0.0452	0.2109	-0.0387	-0.0299
BA	0.0640	0.1488	0.2515	-0.0406	0.0542	0.0116	0.0550	-0.0307
CAT	-0.0730	0.0430	-0.0278	0.0852	-0.0132	0.2088	0.1437	-0.0401
CSCO	0.0309	0.1211	-0.0416	0.0326	-0.0234	0.0530	0.0277	-0.0177
CVX	-0.0171	0.0376	0.0119	0.0276	-0.0092	0.0144	0.0884	-0.0070
DIS	0.0971	0.0549	0.1424	-0.0284	0.0811	-0.0134	0.0196	0.0308
DWDP	0.3140	0.2110	0.1000	0.1250	0.0446	0.1225	0.0668	-0.0277
GE	-0.0152	0.0148	0.0287	0.0276	0.0194	0.0227	0.0141	-0.0182
GS	0.0189	0.2639	-0.0738	0.0749	0.0348	0.0938	-0.0396	0.0526
HD	-0.0557	0.0521	0.1519	0.0496	0.0260	0.1190	0.0225	-0.0193
IBM	0.0938	0.0532	0.0128	0.0261	-0.0152	0.0801	0.0313	-0.0258
INTC	0.0223	0.0315	-0.0071	0.0685	-0.0152	0.0055	0.0260	-0.0266
JNJ	0.0081	0.0514	0.0191	-0.0163	0.0172	0.0665	-0.0052	-0.0090
JPM	0.0326	0.1172	-0.0333	0.0286	0.0345	0.0470	-0.0071	0.0122
KO	0.0519	0.0920	0.0188	0.0580	0.0246	0.0288	0.0788	0.0512
MCD	0.1993	0.2217	0.0468	0.0669	0.0451	0.1489	0.0370	-0.0211
MMM	0.1316	0.2049	0.1524	-0.0348	-0.0231	0.1088	0.0558	-0.0243
MRK	0.0691	0.0122	0.0144	0.0530	0.0498	0.0572	0.1009	-0.0206
MSFT	-0.0133	0.0349	0.0183	0.0273	0.0353	0.0084	0.0217	-0.0043
NKE	0.2952	0.3174	0.2199	-0.0478	0.1105	0.1755	0.1088	-0.0258
PFE	0.0197	0.0554	0.0293	0.0149	0.0264	0.0454	0.0208	-0.0141
PG	0.0709	0.0314	0.0061	0.0340	0.0351	0.0116	0.0683	0.0246
TRV	0.5598	0.3278	0.1963	0.3414	-0.1580	0.2416	0.1205	0.1552
UNH	0.2908	0.1356	0.1251	-0.0413	0.0682	0.1874	-0.0204	-0.0139
UTX	0.0452	0.0463	0.0799	-0.0270	0.0181	-0.0163	0.0469	-0.0223
V	0.1820	0.4427	0.2446	0.1132	-0.0145	0.0954	0.1141	-0.0130
VZ	-0.0785	0.1175	0.0159	0.0113	0.0170	0.0722	-0.0274	-0.0328
WMT	0.0607	0.0184	-0.0087	0.0095	-0.0074	0.0456	0.1101	-0.0103
XOM	0.0144	0.0103	0.0163	0.0040	0.0115	0.0209	0.0380	0.0068

Table 3: Calibrated values of κ for exponential moving averages signals ($\gamma = 0.9$ and 0.96)

	2010	2011	2012	2013	2014	2015	2016	2017
AAPL	0.0073	0.0088	0.0060	0.0053	0.0045	0.0065	0.0053	0.0041
AXP	0.0082	0.0097	0.0047	0.0039	0.0041	0.0056	0.0055	0.0033
BA	0.0080	0.0098	0.0044	0.0044	0.0044	0.0057	0.0054	0.0040
CAT	0.0082	0.0112	0.0055	0.0039	0.0045	0.0062	0.0058	0.0045
CSCO	0.0080	0.0108	0.0055	0.0049	0.0040	0.0059	0.0050	0.0038
CVX	0.0061	0.0092	0.0042	0.0032	0.0042	0.0065	0.0052	0.0036
DIS	0.0067	0.0101	0.0043	0.0039	0.0041	0.0059	0.0043	0.0039
DWDP	0.0091	0.0123	0.0055	0.0047	0.0051	0.0070	0.0047	0.0126
GE	0.0074	0.0098	0.0044	0.0038	0.0036	0.0059	0.0044	0.0045
GS	0.0080	0.0114	0.0058	0.0044	0.0041	0.0057	0.0058	0.0044
HD	0.0067	0.0089	0.0044	0.0039	0.0040	0.0053	0.0045	0.0034
IBM	0.0057	0.0086	0.0040	0.0040	0.0041	0.0055	0.0047	0.0039
INTC	0.0070	0.0091	0.0048	0.0043	0.0048	0.0061	0.0052	0.0041
JNJ	0.0045	0.0084	0.0028	0.0031	0.0036	0.0047	0.0036	0.0031
JPM	0.0080	0.0117	0.0058	0.0041	0.0041	0.0057	0.0055	0.0039
KO	0.0051	0.0067	0.0035	0.0035	0.0036	0.0043	0.0038	0.0027
MCD	0.0050	0.0066	0.0036	0.0030	0.0033	0.0052	0.0040	0.0034
MMM	0.0061	0.0092	0.0037	0.0033	0.0037	0.0051	0.0038	0.0033
MRK	0.0061	0.0081	0.0039	0.0037	0.0043	0.0056	0.0049	0.0038
MSFT	0.0064	0.0083	0.0047	0.0048	0.0042	0.0067	0.0051	0.0037
NKE	0.0065	0.0098	0.0051	0.0043	0.0046	0.0058	0.0051	0.0047
PFE	0.0062	0.0084	0.0035	0.0039	0.0039	0.0053	0.0046	0.0030
PG	0.0046	0.0062	0.0035	0.0036	0.0031	0.0046	0.0038	0.0032
TRV	0.0058	0.0091	0.0041	0.0035	0.0034	0.0049	0.0044	0.0034
UNH	0.0070	0.0099	0.0048	0.0042	0.0042	0.0062	0.0046	0.0036
UTX	0.0061	0.0092	0.0045	0.0036	0.0037	0.0053	0.0046	0.0034
V	0.0082	0.0101	0.0047	0.0043	0.0047	0.0056	0.0049	0.0033
VZ	0.0054	0.0073	0.0037	0.0038	0.0078	0.0046	0.0041	0.0041
WMT	0.0048	0.0067	0.0040	0.0031	0.0034	0.0055	0.0046	0.0042
XOM	0.0059	0.0088	0.0038	0.0032	0.0039	0.0058	0.0046	0.0031

Table 4: Calibrated values of σ^2 for exponential moving averages signals ($\gamma = 0.9$ and 0.96)

References

[1] F. Abergel, H. Aoyama, B.K. Chakrabarti, A. Chakraborti, and A. Ghosh, *Econophysics of Agent-Based Models*, Springer (2014).

[2] R. Almrgen and N. Chriss, "Optimal Execution of Portfolio Transactions", *Journal of Risk*, **3**, 5-29 (2000).

[3] Y. Amihud, H. Mendelson, and L.H. Pedersen, "Liquidity and Asset Prices", *Foundations and Trends in Finance*, 2005, vol. 1, no. 4, pp. 269-364.

[4] F. Baldovin and A. Robledo, "Parallels between the dynamics at the noise-perturbed onset of chaos in logistic maps and the dynamics of glass formation", *Phys. Rev. E* **72**, 066213 (2005). https://arxiv.org/pdf/cond-mat/0504033.pdf (2005).

[5] E. Banijamali, R. Shu, M. Ghavamzadeh, H. Bui, and A. Ghodsi, "Robust Locally-Linear Controllable Embedding", http://lanl.arxiv.org/pdf/1710.05373 (2018).

[6] D. Bertsimas and A.W. Lo, "Optimal Control of Execution Costs", *Journal of Financial Markets*, **1** (1), 1-50, (1998).

[7] D. Bertsimas, V. Gupta, and I.Ch. Paschalidis, "Inverse Optimization: A New Perspective on the Black-Litterman Model", *Operations Research*, Vol.60, No.6, pp. 1389-1403 (2012).

[8] F. Black and R. Litterman, "Global Portfolio Optimization", *Financial Analyst Journal*, Sept-Oct. 1992, 28-43.

[9] J.P. Bouchaud and M. Potters, *Theory of Financial Risk and Derivative Pricing*, second edition, Cambridge University Press (2004).

[10] S. Boyd, E, Busetti, S. Diamond, R.N. Kahn, K. Koh, P. Nystrup, and J. Speth, "Multi-Period Trading via Convex Optimization", *Foundations and Trends in Optimization*. Vol. XX, no. XX, 1-74 (2017).

[11] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, J. Chen, and L. Song, "Smoothed Dual Embedding Control", https://arxiv.org/pdf/1712.10285.pdf (2018).

[12] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm", *Journal of the Royal Statistical Society* B, **39**, 1-38.

[13] A. Dixit and R. Pindyck, *Investment Under Uncertainty*, Princeton University Press, Princeton NJ (1994).

[14] D. Duffie, "Black, Scholes and Merton - Their Central Contributions to Economics" (1997).

[15] D. Ernst, P. Geurts, and L. Wehenkel, "Tree-Based Batch Model Reinforcement Learning", Journal of Machine Learning Research, 6, 405-556, 2005.

[16] S. Esipov, 'Weakly Inefficient Markets: Stability of High-Frequency Trading Strategies", *Lithuanian Journal of Physics*, **52**(2), 102-114 (2012).

[17] C.O. Ewald and Z. Yang, "Geometric Mean Reversion: Formulas for the Equilibrium Density and Analytic Moment Matching", *University of St. Andrews Economics Preprints* (2007).

[18] R. Fox, A. Pakman, and N. Tishby, "Taming the Noise in Reinforcement Learning via Soft Updates", *32nd Conference on Uncertainty in Artificial Intelligence (UAI)* (2016). https://arxiv.org/pdf/1512.08562.pdf (2015).

[19] K. Friston, "The Free-Energy Principle: a Unified Brain Theory?", *Nat. Rev. Neurosci.* **11**, 127-138 (2010). https://doi.org/10.1038.nrn2787.

[20] N. Garleanu, LH.. Pedersen, and A.M. Poteshman, "Demand-Based Option Pricing", *The Review of Financial Studies*, Volume 22, Issue 10 (2009).

[21] N. Garleanu and L.H. Pedersen, "Dynamic Trading with Predictable Returns and Transaction Costs", *Journal of Finance*, vol. 68, issue 6, 2309-2340 (2013).

[22] T. Genewein, F. Liebfried, J. Grau-Mori, and D.A. Braun, "Bounded Rationality, Abstraction, and Hierarchical Decision-Making: An Information-Theoretic Optimality Principle", *Frontiers in Robotics and AI*, **11**. https://doi.org/10.3389/frobt.2015.00027 (2015).

[23] A. Gosavi, "Finite Horizon Markov Control with One-Step Variance Penalties", Conference Proceedings of the Allerton Conferences, Allerton, IL, 2010.

[24] W. Horsthemke and R. Lefever, *Noise-Induced Transitions: Theory and Applications in Physics, Chemistry and Biology*, Springer (1984).

[25] G. Ritter, "Machine Learning for Trading", Risk, October 2017.

[26] I. Halperin, "QLBS: Q-Learner in the Black-Scholes (-Merton) Worlds", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3087076 (2017).

[27] I. Halperin, "The QLBS Q-Learner Goes NuQLear: Fitted Q Iteration, Inverse RL, and Option Portfolios", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3102707 (2018).

[28] I. Halperin, "Inverse Reinforcement Learning for Marketing" , https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3087057 (2017).

[29] L.D. Landau and E.M. Lifschitz, *Statistical Physics*, Elsevier (1980).

[30] D.T. Larsson, D. Braun, and P. Tsiotras, "Hierarchical State Abstractions for Decision-Making Problems with Computational Constraints", https://arxiv.org/pdf/1710.07990.pdf (2017).

[31] S. Levine and V. Koltun, "Guided Policy Search", in *Proceeding of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA (2013).

[32] H. Markowitz, *Portfolio Selection: Efficient Diversification of Investment*, John Wiley, 1959.

[33] R. Marschinski, P. Rossi, M. Tavoni, and F. Cocco, "Portfolio selection with probabilistic utility", *Annals of Operations Research*, vol. 151, issue 1, 223-239 (2007).

[34] R.A. Marsland, "The Edge of Thermodynamics: Driven Steady States in Physics and Biology", Ph.D. Thesis, MIT (2017).

[35] R. Merton, "Theory of Rational Option Pricing", Bell Journal of Economics and Management Science, Vol.4(1), 141-183, 1974.

[36] M. Monfort, A. Liu, and B.D. Ziebart, "Intent prediction and trajectory forecasting via Predictive inverse linear-quadratic regulation", in *Proceedings of 29th AAAI Conference on Artificial Intelligence*, AAAI (2015).

[37] S.A. Murphy, "A Generalization Error for Q-Learning", Journal of Machine Learning Research, 6, 1073-1097, 2005.

[38] R.M. Neal and G.E. Hinton, "A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants", in *Learning in Graphical Models*, Springer 1998, 355-368.

[39] M.J.D. Ramstead, P.B. Badcock, and K.J. Friston, "Answering Schrödinger Question: A Free-Energy Formulation", *Physics of Life Reviews* **24**, 1-16 (2018).

[40] P.A. Ortega and D.A. Braun, "Thermodynamics as a Theory of Decision-Making with Information Processing Costs", *Proceedings of the Royal Society A*, March 2013. https://doi.org/10.1098/rspa.2012.0683 (2013). https://arxiv.org/pdf/1204.6481.pdf (2012).

[41] P.A. Ortega and D.D. Lee, "An Adversarial Interpretation of Information-Theoretic Bounded Rationality", *Proceedings of the Twenty-Eighth AAAI conference on AI* (2014). https://arxiv.org/abs/1404.5668.

[42] P.A. Ortega, D.A. Braun, J.Dyer, K.E. Kim, and N.Tishby, "Information-Theoretic Bounded Rationality", https://arxiv.org/pdf/1512.06789.pdf (2015).

[43] N. Perunov, R.A. Marsland, and J.L. England, "Statistical Physics of Adaptation", *Physical Review X*, **6**, 021036 (2016).

[44] J.M. Poterba and L.H. Summers, "Mean Reversion in Stock Prices: Evidence and Implications", *Journal of Financial Economics*, **22**, 27-59 (1988).

[45] H.A. Simon, "Rational Choice and the Structure of the Environment", *Psychological Review*, **63** (2), 129-138 (1956).

[46] D. Sornette, *Why Stock Markets Crash*, Princeton University Press (2003).

[47] S. Sternberg, *Dynamic Systems*, Dover Publications (2010).

[48] S. Still, "Thermodynamic cost and benefit of data representations", https://arxiv.org/abs/1705.00612 (2017).

[49] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, A Bradford Book (1998).

[50] L. Spierdijk and J.A. Bikker, "Mean Reversion in Stock Prices: Implications for Long-Term Investors", *De Nederlandsche Bank Working Paper No. 343*. Available at SSRN: https://ssrn.com/abstract=2046093 or http://dx.doi.org/10.2139/ssrn.2046093.

[51] E.Schrödinger, *What is Life: The Physical Aspect of the Living Cell*, Cambridge University Press, Cambridge, 1948.

[52] N. Tishby and D. Polani, "Information Theory of Decisions and Actions", in Cutsuridis V., Hussain A., Taylor J. (eds), *Perception-Action Cycle. Springer Series in Cognitive and Neural Systems*. Springer, New York, NY (2011). https://doi.org/10.1007/978-1-4419-1452-1_19.

[53] E. Todorov and W. Li, "A Generalized Iterative LQG Method for Locally-Optimal Feedback Control of Constrained Nonlinear Stochastic Systems", in *Proceeding of the 2005, American COntrol Conference*, Portland OR, USA, pp. 300-306 (2005).

[54] C. Van den Broeck, J.M.R. Parrondo, R. Toral, and R. Kawai, "Nonequilibrium Phase Transitions Induced by Multiplicative Noise", *Phys. Rev. E*, **55** (4), 4084-4094 (1997).

[55] N.G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, North-Holland (1981).

[56] C.J. Watkins and P. Dayan, "Q-Learning", Machine Learning, 8(3-4), 179-192, 1992.

[57] B.I.Yukalov and D. Sornette, "Self-organization in Complex Systems as Decision Making", *Advances in Complex Systems*, **17** (3-4), 1450016 (2014).

[58] B.D. Ziebart, A. Maas, J.A. Bagnell, and A.K. Dey, "Maximum Entropy Inverse Reinforcement Learning" (2008), *AAAI*, p. 1433-1438 (2008).