Dirichlet policies for reinforced factor portfolios

Eric André*   Guillaume Coqueret†

November 12, 2020

Abstract

This article aims to combine factor investing and reinforcement learning (RL). The agent learns through sequential random allocations which rely on firms' characteristics. Using Dirichlet distributions as the driving policy, we derive closed forms for the policy gradients and analytical properties of the performance measure. This enables the implementation of REINFORCE methods, which we perform on a large dataset of US equities. Across a large range of implementation choices, our result indicates that RL-based portfolios are very close to the equally-weighted ($1/N$) allocation. This implies that the agent learns to be *agnostic* with regard to factors. This is partly consistent with cross-sectional regressions showing a strong time variation in the relationship between returns and firm characteristics.

Keywords: Reinforcement learning; Factor Investing; Equally-weighted Portfolio; Asset Pricing.
JEL classifications: C38; G11; G12

1. Introduction

The traditional econometric approaches to asset pricing have recently seen a surge in competition from machine learning tools. A flow of recent studies[1] have shown the benefits that can be reaped when switching from the conventional linear models to more complex structures such as tree methods or neural networks. Supervised learning algorithms help the econometrician link financial performance (asset returns) to key indicators such as firm characteristics (Gu et al. (2020b)) or latent factors (Kelly et al. (2019), Lettau and Pelger (2020a,b)). Based on large datasets, these black boxes reveal intricate correlations between variables that are not captured by standard linear models, thereby (often) improving cross-sectional fit. Depending on the quality of the sample and the algorithm's architecture, these correlations may nonetheless be spurious. They are likely to hold out-of-sample only if they reflect genuine causality relationships, which are much harder to uncover.[2]

Beyond supervised learning, researchers have resorted to another powerful family of techniques to understand and predict returns: reinforcement learning (RL). Contributions range from early

*EMLYON Business School, 23 avenue Guy de Collongue, 69130 Ecully, FRANCE. E-mail: eandre@em-lyon.com
†EMLYON Business School, 23 avenue Guy de Collongue, 69130 Ecully, FRANCE. E-mail: coqueret@em-lyon.com

[1]See for instance Chen et al. (2019), Feng et al. (2019), Gu et al. (2020a) and Gu et al. (2020b).

[2]We refer to Pearl (2009) for an exhaustive treatment on causal models and to Arjovsky et al. (2019) and Pfister et al. (2019) for recent perspectives on causality in machine learning models.

tests of Neuneier (1996) and Moody et al. (1998) to the more recent work of Deng et al. (2016), Li et al. (2019) and Wang and Zhou (2020).[3] Two common threads between these studies is that they often originate from the field of computer science and that they work with price data only (at high frequencies most of the time). To the best of our knowledge, there are no contributions that seek to harvest the information contained in firm specific attributes and combine it with reinforcement routines to produce factor-based portfolios. One goal of the present paper is to fill this void.

The main challenge when implementing RL algorithms for the purpose of trading is the modelling of the environment. The infinite dimensions of the state space (firm attributes) and action space (investment policies) make many approaches relying on Markov decision processes (MPD) inadequate. This is because a focal tool in MDP analysis is the value function, which measures the expected gain or reward for any given action or state. In the framework of factor investing, these states and actions cannot be properly discretized without either making overly simplistic assumptions, or rendering the computations intractable.

In order to bypass these technical hurdles, one solution is to resort to the so-called policy gradient approach. In this case, the decisions are made according to a parametric function which probabilistically determines which actions (i.e., investments) to perform. The agent then learns by sequentially updating the policy parameters after receiving flows of rewards (e.g., returns). Most of the time, the policy is modelled by neural networks (NNs), which is a convenient choice, given their flexibility. It is for instance the option chosen by Deng et al. (2016) and Zhang et al. (2019). One drawback of general purpose NNs is that their output cannot be directly translated into portfolio weights, because it violates the budget constraint. The core idea of the present paper is to resort to a special class of distributions that circumvent this issue by directly yielding the investment allocations.

Indeed, Dirichlet distributions have the opportune property of being defined on simplexes, which makes them appropriate to model long-only portfolio compositions. In fact, Dirichlet distributions have already been used in related studies. Cover and Ordentlich (1996) find that two such distributions yield portfolio allocations with interesting theoretical properties. More recently, Le Courtois and Xu (2019) rely on Dirichlet distributions to derive robust estimates of the efficient frontier, and Korsos (2013) uses them to estimate the composition of hedge fund portfolio holdings. In a similar vein, Sosnovskiy (2015) shows that Dirichlet laws can be used to approximate the distribution of stock weights in aggregate market indices.

One of the simple but novel contributions of the paper is to link the RL policy to firm-specific attributes. To this purpose, the inspiration comes from earlier work on characteristics-based investing.[4] The idea is to map a linear combination of the characteristics into portfolio weights. While the traditional models aim to optimize expected utility functions, our approach seeks to maximize expected gains. The simplest definition of gain is a portfolio return but it is possible to adjust it to risk via the sequential Sharpe ratio computations presented in Moody et al. (1998).

Our contribution is threefold. First, we propose a tractable formulation of the reinforcement learning problem when designing portfolio allocations based on firm specific attributes. To the best

[3]These references are by no means an exhaustive account of the literature on this subject. On the arXiv repository only, more than 20 papers including the terms "reinforcement learning" in their title have been posted in the *quantitative finance* (qfin) section in 2019 only. We also direct to the survey Sato (2019) for more references on RL applied to portfolio optimization.

[4]See, e.g., Haugen and Baker (1996), Daniel and Titman (1997), Brandt et al. (2009), Hjalmarsson and Manchev (2012) and Ammann et al. (2016)). Our approach is closer in spirit to the most recent of these references.

of our knowledge our approach is the first to articulate the combination between factor investing and RL in such a simple fashion. Second, we employ our methodology on a large dataset of US equities. Our results are qualitatively homogeneous, despite the numerous degrees of freedom in the implementation, and they indicate that the agent should be better of by *ignoring* the informational content provided by firm-specific attributes. Finally, we compare the learning process to a simple factor-based quadratic optimization. The two are hard to reconcile, except for one salient stylized fact: both methods recognize a strong common factor within the cross-section of stock returns. Consequently, portfolios allocate almost uniformly across assets.

The paper is structured as follows. In Section 2, we lay out the theoretical foundations of RL-based factor investing. Section 3 is dedicated to a detailed presentation of the dataset and the implementation protocol. Our empirical results are outlined in Section 4. In Section 5, we compare our baseline findings to those stemming from a more classical, asset pricing model. Finally, 6 concludes.

2. Reinforcement learning meets factor investing

This section is dedicated to the presentation of all concepts and theoretical apparatus developed and required in the paper.

2.1. The framework

We study a dynamic discrete time investment problem with finite horizon T. The investable universe consists of N assets indexed by $n = 1, \dots, N$. There are K characteristics associated to each asset. We refer to section 3.1 for a list of those retained in the empirical section of this study. To allow for a bias or non-zero intercept in our model, we add a constant characteristic equal to 1. Therefore, at time $t \in \{0, 1, \dots, T\}$, asset n is described by a $(K+1)$-dimensional vector $\boldsymbol{x}_{t,n} = \left[x_{t,n}^{(0)} \dots x_{t,n}^{(k)} \dots x_{t,n}^{(K)}\right]^\intercal$ where $x_{t,n}^{(0)} = 1$ is an indicator that is kept fixed through the cross-section of assets.

Among these characteristics are $p_{t,n}$, the time-t price of asset n, and $d_{t,n}$, the dividend per share issued between time $t-1$ and t. The total return of asset n between $t-1$ and t is therefore

$$r_{t,n} = \frac{p_{t,n} + d_{t,n}}{p_{t-1,n}} - 1. \tag{1}$$

In our setting, we can work with price returns (omitting dividends) or total returns interchangeably, as they are simply two different drivers of rewards for the investor. We use the bold notations $\boldsymbol{r}_t$ for the vector of the returns of all assets at time t and $\boldsymbol{X}_t$ for the $N \times (K+1)$ matrix of characteristics at time t:

$$\boldsymbol{X}_t = \begin{bmatrix} (\boldsymbol{x}_{t,1})^\intercal \\ \vdots \\ (\boldsymbol{x}_{t,n})^\intercal \\ \vdots \\ (\boldsymbol{x}_{t,N})^\intercal \end{bmatrix}.$$

We denote by $\mathcal{M}$ the set of these $N \times (K+1)$ matrices whose first column is $\mathbb{1}$, the vector of 1.

The agent posits a factor model for the returns of the assets

$$\boldsymbol{r}_{t+1} = f(\boldsymbol{X}_t) + \boldsymbol{\epsilon}_{t+1} \tag{2}$$

where f is a function from $\mathcal{M}$ to $\mathbb{R}^N$ and $\boldsymbol{\epsilon}_t$ is an i.i.d. White Noise with mean vector equal to 0 and a diagonal correlation matrix $\boldsymbol{\Sigma}_\epsilon$. The diagonal elements of $\boldsymbol{\Sigma}_\epsilon$ are the idiosyncratic variances of the assets σ_n^2. Let $\boldsymbol{P}_\epsilon$ denote the law of the r.v. $\boldsymbol{\epsilon}_t$, defined on $\mathbb{R}^N$.

A standard assumption in the finance literature is that the function f is a linear map that can be represented by a $(K+1)$ vector $\boldsymbol{\beta}$, that is $f(\boldsymbol{X}_t) = \boldsymbol{X}_t\boldsymbol{\beta}$. Note however that we do not need this assumption in our study.

2.2. Markov Decision Process

We assume that the investment problem can be formulated as a finite horizon Markov Decision Process (MDP). At each time t, the agent observes the state S_t of the system (the characteristics of the investable universe and of her portfolio) and then takes an action A_t (a choice of a composition for her portfolio). Finally, the agent obtains a time-$(t+1)$ reward, which is linked to the return of her portfolio between t and $t+1$ and the system transition to the next state. We now describe formally this MDP[5].

Actions. The *action* A_t taken by the agent at time t is the choice of a vector $\boldsymbol{w}_t \in \mathbb{R}^N$, which is the composition of her portfolio. We consider the case where there is no short selling. The restriction to positive weights is realistic since most asset managers have long-only constraints. This is typically the case of institutional investors (see Koijen and Yogo, 2019). Therefore $\boldsymbol{w}_t$ must be in the $N-1$ simplex Δ (we omit the dimension superscript to lighten notations), which is then the *action space*:

$$\Delta = \left\{ (w_1, \ldots, w_N) \in \mathbb{R}^N \colon \sum_{n=1}^{N} w_n = 1 \text{ and } w_n \geq 0 \text{ for all } n \right\}. \tag{3}$$

Seen as a subset of $\mathbb{R}^{N-1}$, it is endowed with the inherited Borel σ-algebra that we denote $\mathscr{B}(\Delta)$.

Rewards. The agent's objective is to maximise her utility of, or some performance measure of, the terminal value of her portfolio $V_T = V_0 + \sum_{t=0}^{T-1} V_t \rho_{t+1}$, where $\rho_{t+1} \coloneqq \boldsymbol{w}_t^\intercal \boldsymbol{r}_{t+1}$ is the return of her portfolio between t and $t+1$. We will consider two cases: a risk insensitive agent who seeks to maximize her profit and a risk sensitive agent whose goal is to maximize the differential Sharpe Ratio proposed by Moody et al. (1998).

In the first case, the agent's reward at time t for the action taken at time $t-1$ is simply

$$R_t = \rho_t \tag{4}$$

In the second case, it is

$$R_t \coloneqq \mathrm{SR}_t = \frac{\hat{\mu}_t}{K_\kappa \sqrt{\hat{\sigma}_t^2 - \hat{\mu}_t^2}} \tag{5}$$

[5]See, e.g., Bäuerle and Rieder (2011).

with

$$\begin{array}{ll} \hat{\mu}_t = \kappa\rho_t + (1-\kappa)\hat{\mu}_{t-1} & \text{exponentially weighted (EW) moving average of returns;} \\ \hat{\sigma}_t^2 = \kappa\rho_t^2 + (1-\kappa)\hat{\sigma}_{t-1}^2 & \text{EW moving average of squared returns;} \\ K_\kappa = \sqrt{\frac{1-\kappa/2}{1-\kappa}} & \text{scaling factor.} \end{array}$$

We underline that we use uppercase R_t for rewards and lowercase r_t for returns. The two are closely linked, but not equal.

States. As we want the agent to choose an action using the asset characteristics, $\boldsymbol{X}_t$ must be included in the *state*. The chosen reward defines which data must be added to the state: for the risk-insensitive agent ρ_t is enough, for the differential Sharpe Ratio, we should also include $\hat{\mu}_t$ and $\hat{\sigma}_t^2$. In the latter case, it is still ρ_t that drives deterministically the additional data, therefore we will henceforth consider without loss of generality the case where a state corresponds to the couple

$$S_t = (\rho_t, \boldsymbol{X}_t)$$

The *state space* is then $\mathcal{S} = \mathbb{R} \times \mathcal{M}$. The set $\mathcal{M}$, seen as a subset of the space of the $N \times (K+1)$ matrices which can be identified with $\mathbb{R}^{N\times(K+1)}$, is endowed with the inherited product of Borel σ-algebras that we denote $\mathscr{B}(\mathcal{M})$. The state space itself is endowed with the product σ-algebra $\mathscr{B}(\mathcal{S}) = \mathscr{B}(\mathbb{R}) \otimes \mathscr{B}(\mathcal{M})$.

Episodes. The agent having observed the state of the system takes an action, then the system transitions to the next state and the agent receives a reward. A given realization of this interaction between the agent and her environment is an *episode*:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$$

At any date t, the *cumulative discounted return* can be computed. It is the sum of the future rewards in this episode, possibly discounted at a discount rate $0 < \gamma \leq 1$:

$$G_t = \sum_{l=1}^{T-t} \gamma^{l-1} R_{t+l} = R_{t+1} + \gamma G_{t+1}$$

For the differential Sharpe Ratio, this additive structure is obtained by construction. In the case of the risk insensitive agent, let us rewrite the growth of the portfolio from t to T as $V_T = V_t \prod_{l=1}^{T-t}(1+\rho_{t+l})$. Maximizing the profit on this horizon is the same as maximizing $\sum_{l=1}^{T-t} \log(1+\rho_{t+l})$ which is approximately equal to $\sum_{l=1}^{T-t} \rho_{t+l} = \sum_{l=1}^{T-t} R_{t+l}$ when the horizons of returns is not too long (which implies that the magnitude of the portfolio returns ρ_{t+l} is small).

Transition probability. How the system transitions to the next state S_{t+1} given some previous state S_t and action A_t is given by the *state transition probability*

$$\operatorname{Prob}\left(S_{t+1} \in B \mid S_t, A_t\right), \quad B \in \mathscr{B}(\mathcal{S}).$$

We assume that the matrix of asset characteristics is a Markov process whose evolution is driven by the transition probabilities

$$\mathbb{P}_t^u\left(M \mid \boldsymbol{X}\right) = \operatorname{Prob}\left(\boldsymbol{X}_u \in M \mid \boldsymbol{X}_t = \boldsymbol{X}\right), \qquad u > t$$

which are independent of the value of the portfolio and of the choice of the action. Therefore, if B is a Cartesian product of Borel sets, $B = C \times M$, where $C \in \mathscr{B}(\mathbb{R})$ and $M \in \mathscr{B}(\mathcal{M})$, we obtain the factorization

$$\text{Prob}\left(S_{t+1} \in B \mid S_t, A_t\right) = \text{Prob}\left(\rho_{t+1} \in C \mid S_t, A_t\right) \cdot \mathbb{P}_t^{t+1}\left(M \mid \boldsymbol{X}_t\right).$$

In our specific setting with a factor model, we have a transition function $\mathbb{T}$ that gives the next value of ρ_{t+1} given the state and action at t. This is $\rho_{t+1} = \mathbb{T}(\boldsymbol{X}_t, \boldsymbol{w}_t, \boldsymbol{\epsilon}_{t+1}) = \boldsymbol{w}_t^\intercal(f(\boldsymbol{X}_t) + \boldsymbol{\epsilon}_{t+1})$. When S_t and A_t are known, the value of ρ_{t+1} is driven by $\boldsymbol{\epsilon}_{t+1}$ and conversely, if $r \in \mathbb{R}$, then $\mathbb{T}^{-1}(r \mid \boldsymbol{X}_t, \boldsymbol{w}_t)$ is the hyperplane orthogonal to the vector $\boldsymbol{w}_t$ translated by the vector $-f(\boldsymbol{X}_t)$ and by any vector $\boldsymbol{\alpha}$ such that $\boldsymbol{w}_t^\intercal \boldsymbol{\alpha} = r$. Finally, we can write

$$\text{Prob}\left(S_{t+1} \in C \times M \mid S_t, A_t\right) = \boldsymbol{P}_\epsilon(\boldsymbol{\epsilon}_{t+1} \in \mathbb{T}^{-1}(C \mid \boldsymbol{X}_t, \boldsymbol{w}_t)) \cdot \mathbb{P}_t^{t+1}\left(M \mid \boldsymbol{X}_t\right)$$

Of special interest for our study is the computation of future periods expected returns given some state and action at date t. It is given by the following lemma.

Lemma 1.

$$\int_{\mathcal{S}} \rho_{t+l} \text{Prob}(ds \mid S_t = (\rho_t, \boldsymbol{X}_t), A_t = \boldsymbol{w}_t) = \begin{cases} \boldsymbol{w}_t^\intercal f\left(\boldsymbol{X}_t\right) & l = 1 \\ \int_{\mathcal{M}} \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+l-1} \mid \xi\right]^\intercal f\left(\xi\right) \mathbb{P}_t^{t+l-1}(d\xi \mid \boldsymbol{X}_t) & l \geq 2 \end{cases}$$

where $\mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+l-1} \mid \xi\right]$ *is the vector expectation of the portfolio composition.*

Proof. See Appendix A.1. □

2.3. Policy

To allow for the exploration of all actions versus the exploitation of the optimal action, we will use a stochastic policy that gives the probability of choosing an action A_t given the state S_t. Specifically, we will study policies $\pi_{\boldsymbol{\theta}}$ defined on $\mathscr{B}(\Delta)$ with parameter $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(k)})$. At each time step, we will draw from this distribution to select an action:

$$A_t = \boldsymbol{w}_t \sim \pi_{\boldsymbol{\theta}}\left(\cdot \mid S_t\right).$$

More precisely, we are looking in this study for a policy that only takes into account the asset characteristics, hence we restrict ourselves to policies that takes the form

$$\boldsymbol{w}_t \sim \pi_{\boldsymbol{\theta}}\left(\cdot \mid \boldsymbol{X}_t\right)$$

We will use the notations $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ or $\mathbb{E}_\pi[\cdot \mid \boldsymbol{\theta}]$ for the expectation under the policy $\pi_{\boldsymbol{\theta}}$. When there is no ambiguity, we will omit to write that a result holds conditionally on a given a policy.

Value function. The *value function* at t of the state S_t under the policy $\pi_{\boldsymbol{\theta}}$, is the expected value of the cumulative discounted return from t onward, when this policy is chosen to select the actions at each future time steps:

$$V^{\boldsymbol{\theta}}(t, S_t) = \mathbb{E}_{\boldsymbol{\theta}}\left[G_t \mid S_t\right] = \sum_{l=1}^{T-t} \gamma^{l-1} \mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+l} \mid S_t\right]$$

To find the optimal policy, the standard tool is dynamic programming, for which the value function must satisfy the recursive Bellman equation (see Chapter 4 in Sutton and Barto (2018)). However, for the differential Sharpe Ratio, it is known that the introduction of the variance in the reward renders the problem time inconsistent. In this paper, we will use RL algorithms to explore the optimal policies. Nonetheless, in the case of the risk insensitive agent, the problem can also be solved with dynamic programming as the next result shows.

Proposition 2. *For the risk insensitive agent, the time t expected values of the future rewards are given by*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+l} \mid S_t = (\rho_t, \boldsymbol{X}_t)\right] = \begin{cases} \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_t \mid \boldsymbol{X}_t\right]^\intercal f\left(\boldsymbol{X}_t\right) & l = 1, \\ \int_{\mathcal{M}} \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+l-1} \mid \xi\right]^\intercal f\left(\xi\right) \mathbb{P}_t^{t+l-1}(d\xi \mid \boldsymbol{X}_t) & l \geq 2. \end{cases}$$

The policy value satisfies the recursive Bellman equation

$$\begin{aligned} V^{\boldsymbol{\theta}}(t, \boldsymbol{X}_t) &= \mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+1} \mid S_t\right] + \int_{\mathcal{M}} V^{\boldsymbol{\theta}}(t+1, \xi) \mathbb{P}_t^{t+1}(d\xi \mid \boldsymbol{X}_t), \\ V^{\boldsymbol{\theta}}(T-1, \boldsymbol{X}_{T-1}) &= \mathbb{E}_{\boldsymbol{\theta}}\left[R_T \mid S_{T-1}\right]. \end{aligned}$$

Proof. See Appendix A.2. □

Performance measure. The *performance measure* of the policy is its value from some initial state S_0: $J(\boldsymbol{\theta}) = V^{\boldsymbol{\theta}}(0, S_0)$. Our aim is to find a parameter of the policy that maximizes this performance measure

$$\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}). \tag{6}$$

Before taking on this task, we specify the parametrized form of the policy that we use in this study.

2.4. Dirichlet policies

One of the main contribution of this paper is the use of Dirichlet distributions to define the policy of the agent. We find it particularly well suited for describing portfolio weights when short selling is proscribed. We first briefly recall its definition and some of its properties that are used thereafter.

Definition. The Dirichlet distribution is defined on the $N-1$ simplex (3) and its density is zero outside Δ. It is parametrized by a vector $\boldsymbol{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_N \end{bmatrix}^\intercal$ of *concentration parameters* where $a_n > 0$ for all $n = 1, \ldots, N$. We will use the notation σ for the *scale parameter*

$$\sigma = \sum_{n=1}^{N} a_n = \mathbb{1}^\intercal \boldsymbol{a}.$$

The probability density function (pdf) is given by

$$f(w_1, \dots, w_N \mid \boldsymbol{a}) = \frac{1}{B(\boldsymbol{a})} \prod_{n=1}^{N} w_n^{a_n - 1},$$

where the normalizing constant is the Multivariate Beta function, which can be written with the Gamma function as follows

$$B(\boldsymbol{a}) = \frac{\prod_{n=1}^{N} \Gamma(a_n)}{\Gamma(\sigma)}.$$

A note on Order and Dimension. The Lebesgue measure of Δ^{N-1} in $\mathbb{R}^N$ being 0, for the pdf to be meaningful and integrate to 1, it must be the density with respect to the Lebesgue measure on $\mathbb{R}^{N-1}$. Hence it is a function defined on $\mathbb{R}^{N-1}$ and we should write

$$f(w_1, \dots, w_{N-1} \mid \boldsymbol{a}) = \frac{1}{B(\boldsymbol{a})} \left(\prod_{n=1}^{N-1} w_n^{a_n - 1} \right) \left(1 - \sum_{n=1}^{N-1} w_n \right)^{a_N - 1}.$$

As a consequence, we say that a random vector $\boldsymbol{W} = [W_1 \ \cdots \ W_{N-1}]^\intercal$ of dimension $N-1$ has the Dirichlet distribution of order N with concentration parameters $\boldsymbol{a} = [a_1 \ a_2 \ \cdots \ a_N]^\intercal$, and we write

$$\boldsymbol{W} \sim \mathrm{Dir}_N(\boldsymbol{a}).$$

Alternatively, we could write

$$\left[W_1 \quad \cdots \quad W_{N-1} \quad 1 - \sum_{n=1}^{N-1} W_n \right]^\intercal \sim \mathrm{Dir}_N(\boldsymbol{a}).$$

In the case of a 1-dimensional vector, we write $W \sim \mathrm{Dir}_2(a_1, a_2)$ or $[W \ 1-W] \sim \mathrm{Dir}_2(a_1, a_2)$, the density is then

$$f(w \mid a_1, a_2) = \frac{1}{B(a_1, a_2)} w^{a_1 - 1} (1 - w)^{a_2 - 1},$$

that is, $W \sim \mathrm{Beta}(a_1, a_2)$ and the Dirichlet distribution is sometimes called the Multivariate Beta distribution.

Some properties. Let $\boldsymbol{W} \sim \mathrm{Dir}_N(\mathbf{a})$. The marginal distributions are Beta distributions: for $n = 1, \dots, N-1$,

$$W_n \sim \mathrm{Beta}(a_n, \sigma - a_n),$$

and the two-dimensional marginal distributions are Dirichlet: for $1 \le n < m \le N-1$,

$$\left[W_n \quad W_m \right]^\intercal \sim \mathrm{Dir}_3(a_n, a_m, \sigma - a_n - a_m).$$

Let $\digamma$ denotes the Digamma function, the derivative of the natural logarithm of the Gamma function: for $a > 0$,

$$\digamma(a) = \frac{\mathrm{d}}{\mathrm{d}a} \Gamma(a) = \frac{\Gamma'(a)}{\Gamma(a)}.$$

The first equality below is well documented and we prove the three additional results in appendix A.3.

Proposition 3. *Let* $\mathbf{W} \sim \mathrm{Dir}_N(\mathbf{a})$*. Then for* $1 \leq n \leq N-1$*,*

$$\begin{aligned}
\mathbb{E}[W_n] &= \frac{a_n}{\sigma}, \\
\mathbb{E}[\ln W_n] &= \digamma(a_n) - \digamma(\sigma), \\
\mathbb{E}[W_n \ln W_n] &= \frac{a_n}{\sigma}\left(\digamma(a_n) + \frac{1}{a_n} - \digamma(\sigma) - \frac{1}{\sigma}\right),
\end{aligned}$$

and for $1 \leq m \leq N-1$*,* $n \neq m$*,*

$$\mathbb{E}[W_n \ln W_m] = \frac{a_n}{\sigma}\left(\digamma(a_m) - \digamma(\sigma) - \frac{1}{\sigma}\right).$$

Link with the asset characteristics. In this paper, we study the policy for which the probability of choosing action $\boldsymbol{w}_t$ at time t has the Dirichlet distribution with concentration parameters $\mathbf{a}_t = \begin{bmatrix} a_{t,1} & a_{t,2} & \cdots & a_{t,N} \end{bmatrix}^\intercal$ where $a_{t,n} > 0$ for all n. We posit that the concentration parameters are functions of the asset characteristics. Two possible forms are studied:

$$\mathbf{a}_t = \begin{cases} \boldsymbol{X}_t \boldsymbol{\theta}_t & (\mathbf{F1}) \\ e^{\boldsymbol{X}_t \boldsymbol{\theta}_t} & (\mathbf{F2}). \end{cases} \tag{7}$$

The first form is a simple linear combination which is highly tractable, but may violate the condition that $a_{t,n} > 0$ for some values of $\theta_t^{(k)}$. Indeed, during the learning process, an update in $\boldsymbol{\theta}$ might yield values that are out of the feasible set of $\boldsymbol{a}_t$. In this case, it is possible to resort to a trick that is widely used in online learning (see, e.g., Section 2.3.1 in Hoi et al., 2018). The idea is simply to find the acceptable solution that is closest to the suggestion from the algorithm. If we call $\boldsymbol{\theta}^*$ the result of an update rule from a given algorithm, then the closest feasible vector is

$$\boldsymbol{\theta} = \min_{\boldsymbol{z} \in \Theta(\boldsymbol{X}_t)} ||\boldsymbol{\theta}^* - \boldsymbol{z}||^2, \tag{8}$$

where $||\cdot||$ is the Euclidean norm and $\Theta(\boldsymbol{X}_t)$ is the feasible set, that is, the set of vectors $\boldsymbol{\theta}$ such that the $a_{t,n} = \theta_t^{(0)} + \sum_{k=1}^K \theta_t^{(k)} x_{t,n}^{(k)}$ are all nonnegative.

The second form of the policy is slightly more complex but remains always valid.

The combination of the Dirichlet distribution with time-varying weights $\boldsymbol{w}_t$ and parameters $\boldsymbol{a}_t$ defined above yields a policy $\pi := \pi_{\boldsymbol{\theta}}$ that depends on exogenous characteristics $\boldsymbol{X}_t$ as well as $K+1$ parameters, stacked in the vector $\boldsymbol{\theta}_t$. There is a very strong link between this formulation and other methods that link financial performance to firm-specific characteristics like Brandt et al. (2009) and Ammann et al. (2016). One common feature is that for any $k \neq 0$, the parameter $\theta^{(k)}$ synthesizes the impact of feature k on the whole cross-section of returns. If $\theta^{(k)}$ is positive (*resp.*, negative), then, on average, the corresponding feature is expected to have a positive (*resp.*, negative) effect on returns. The parameter $\theta^{(0)}$ is intended to reflect some idiosyncrasy that is not rendered by the characteristics but that is shared by all assets.

Note that if the idiosyncratic parameter $\theta^{(0)}$ is much larger than the sum $\sum_{k=1}^K \theta_t^{(k)} x_{t,n}^{(k)}$ for all n, we converge to the limiting case where $\mathbf{a}_t = \theta^{(0)}\mathbb{1}$. In this case the Dirichlet distribution is symmetric and the portfolio converges to the Equally Weighted portfolio as $\mathbb{E}[W_n] = 1/N$ for all n.

2.5. *The policy gradient method*

The optimization problem (6) cannot be solved by dynamic programming when the reward is the differential Sharpe Ratio. We thus search for an approximate solution using the method named Policy Gradient (Sutton and Barto, 2018, Chapter 13). This method can deal with the infinite state space $\mathcal{S}$ and seeks to learn a parametrized policy by updating the parameter via gradient ascent in J:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \widehat{\nabla J(\boldsymbol{\theta}_t)} \tag{9}$$

where $\widehat{\nabla J(\boldsymbol{\theta}_t)}$ is a stochastic estimate of the gradient of the performance measure (with respect to $\boldsymbol{\theta}_t$) and $\alpha \in (0,1)$ is a learning rate.

The core result when implementing policy gradient learning is the so-called Policy Gradient Theorem:

$$\nabla J(\boldsymbol{\theta}_t) = \mathbb{E}_\pi \left[G_t \nabla \ln \pi \left(\boldsymbol{w}_t \mid S_t, \boldsymbol{\theta}_t \right) \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t \right] \tag{10}$$

which is incredibly convenient because the two terms in the expectation are disentangled. We refer to Section 13.3 in Sutton and Barto (2018) for a proof of this result. It is thus imperative to derive analytical expressions for $\nabla ln \pi_{\boldsymbol{\theta}}$.

Closed forms for the Dirichlet policy gradients We recall the two policy options derived from (**F1**) and (**F2**) in (7):

$$\pi(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t) = \frac{\Gamma\left(\sigma_t\right)}{\prod_{n=1}^{N} \Gamma\left(a_{t,n}\right)} \prod_{n=1}^{N} w_{t,n}^{a_{t,n}-1}, \tag{11}$$

where

$$a_{t,n} = \begin{cases} \boldsymbol{x}_{t,n}\boldsymbol{\theta}_t & (\mathbf{F1}) \\ \exp(\boldsymbol{x}_{t,n}\boldsymbol{\theta}_t) & (\mathbf{F2}) \end{cases} \quad \text{for all } 1 \leq n \leq N, \text{ and } \quad \sigma_t = \sum_{n=1}^{N} a_{t,n}. \tag{12}$$

Proposition 4. *For a Dirichlet policy, the gradients are given by*

$$\nabla \ln \pi \left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t \right) = \sum_{n=1}^{N} \left(\digamma\left(\sigma_t\right) - \digamma\left(a_{t,n}\right) + \ln w_{t,n} \right) \nabla a_{t,n},$$

where

$$\nabla a_{t,n} = \begin{cases} \boldsymbol{x}_{t,n}^\intercal & (\mathbf{F1}) \\ \exp(\boldsymbol{x}_{t,n}\boldsymbol{\theta}_t)\boldsymbol{x}_{t,n}^\intercal & (\mathbf{F2}) \end{cases}.$$

Proof. From equation (11) we obtain

$$\ln \pi \left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t \right) = \ln \Gamma\left(\sigma_t\right) - \sum_{n=1}^{N} \ln \Gamma(a_{t,n}) + \sum_{n=1}^{N} (a_{t,n} - 1) \ln w_{t,n}$$

Then

$$\begin{aligned} \nabla \ln \pi \left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t \right) &= \digamma\left(\sigma_t\right) \sum_{n=1}^{N} \nabla a_{t,n} - \sum_{n=1}^{N} \digamma(a_{t,n}) \nabla a_{t,n} + \sum_{n=1}^{N} \ln w_n \nabla a_{t,n} \\ &= \sum_{n=1}^{N} \left(\digamma\left(\sigma_t\right) - \digamma\left(a_{t,n}\right) + \ln w_n \right) \nabla a_{t,n} \end{aligned}$$

where the gradients of the concentration parameters $a_{t,n}$ are computed for both case given in equation (12). $\square$

It can be seen that the policy gradient is the weighted sum of the concentration parameters' gradients. The weight applied to each parameter's gradient is given by the difference between the actual value of the logarithm of the weight of asset n in the portfolio and the expected value of this logarithm under the policy $\pi_{\boldsymbol{\theta}_t}$. Indeed, using Proposition 3, we have

$$F(\sigma_t) - F(a_{t,n}) + \ln w_{t,n} = \ln w_{t,n} - \mathbb{E}_\pi \left[\ln w_{t,n} \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t \right]. \tag{13}$$

By Proposition 3, the average weight of asset n in the portfolio is

$$\mathbb{E}_\pi \left[w_{t,n} \mid S_t, \boldsymbol{\theta}_t \right] = \frac{a_{t,n}}{\sigma_t}, \tag{14}$$

hence, the gradient of the concentration parameter $\nabla a_{t,n}$ gives the direction in the parameter space that will increase the relative importance of asset n in the portfolio. Therefore, learning with the policy gradient given by equation (10) will, when G_t is positive, increase the expected weights at time $t+1$ of those assets which had, at time t, their realized log weights above their expected values. In this way, the stochastic policy allows to explore the action space through random deviations from the mean, that are reinforced if they generate a profit.

Gradient of the performance measure For the risk insensitive agent, the gradient of the performance measure takes a simple form.

Proposition 5. *For the risk insensitive agent, under a Dirichlet policy,*

$$\nabla J(\boldsymbol{\theta}_t) = \sum_{n=1}^{N} \left(\mathbb{E}\left[r_{t+1,n} \mid \boldsymbol{X}_t \right] - \mathbb{E}_\pi \left[R_{t+1} \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t \right] \right) \frac{\nabla a_{t,n}}{\sigma_t},$$

where we recall that $r_{t+1,n}$ is the return of asset n between time t and $t+1$ (while R_{t+1} is the reward).

Proof. See Appendix A.4. $\square$

We see that the learning process is "*myopic*", as the one step ahead return is the only one taken into consideration for the update of the parameters. The learning process will, on average, increase (*resp.* decrease) the weights of those assets whose expected returns are higher (*resp.* lower) than the portfolio's expected return. This behavior could be expected from the risk-insensitive agent.

3. Data and protocols

In this section, we describe the data on which we carry out our empirical analysis. Additionally, we discuss many implementation issues related to the RL framework.

3.1. Data

The dataset comprises firms listed in the US between January 2000 and June 2020 downloaded from Bloomberg. The number of firms through time is depicted in Figure 1. Observations are sampled at a monthly frequency. Each stock is characterized by twelve attributes that correspond to documented predictors (accounting-based, risk-based and momentum-based). These variables are summarized in Table 1. We restrict our analysis to these twelve indicators to be able to easily comment on the associated values of $\boldsymbol{\theta}$. These features naturally serve as the non constant components of the $\boldsymbol{X}_t$ matrices in our model.

Fig. 1. **Number of firms in the sample, through time**.

The features (predictors) are cross-sectionally processed so that for a fixed t and given predictor j, $\boldsymbol{x}_t^{(j)}$ is uniformly distributed (over the [-0.5,0.5] interval) across firms. Scaling predictors is standard practice both in the machine learning literature and in some recent asset pricing models (e.g., Koijen and Yogo (2019), Kelly et al. (2019) and Freyberger et al. (2020)). In characteristics-based investing (e.g., in Brandt et al. (2009)), the indicators are for instance also demeaned and standardized.

3.2. The REINFORCE algorithm and implementation issues

The learning process at the core of our method is the so-called REINFORCE algorithm, which is the most straightforward route towards the policy gradient approach (see Chapter 13 of Sutton and Barto (2018)). We briefly recall the steps in Table 2.

In spite of its apparent simplicity, the REINFORCE algorithm leaves a lot of room for implementation design. Below, we discuss open options (highlighted in bold font) for the steps defined in Table 2:

- **Step 0**: the policy is defined by Equation (11) along with one of the specifications (**F1**) or (**F2**). The levels of the two rates γ and η will be discussed below.
- **Step 1**: the central question is: what is an episode? More precisely, do we need to impose a chronological ordering of events? In traditional RL, this is imperative because actions can

Short name	Long name	Academic references (alphab. order)
cst	Constant	-
cap	Market Capitalization	Banz (1981); Fama and French (1992)
pb	Price-to-Book ratio	Asness et al. (2013); Fama and French (1992)
de	Debt-to-Equity ratio	Barbee Jr et al. (1996); Bhandari (1988)
vol	Realized volatility in the past 30 days	Baker et al. (2011)
prof	Profitability	Fama and French (2015)
inv	Asset growth	Cooper et al. (2008); Fama and French (2015)
eps	Earnings per share	Ball and Brown (1968, 2019)
liq	Trading volume	Chordia and Swaminathan (2000)
rsi	Relative strength index	Han et al. (2013)
pe	Price-earnings ratio	Basu (1983); Easton (2004)
dy	Dividend yield	Litzenberger and Ramaswamy (1982), Naranjo et al. (1998)
mom	12-1M momentum	Asness et al. (2013); Jegadeesh and Titman (1993)

Table 1: **List of predictors and associated academic references**. The Bloomberg fields are, in order, CUR_MKT_CAP, PX_TO_BOOK_RATIO, TOT_DEBT_TO_TOT_EQY, VOLATILITY_30D, PROF_MARGIN, ASSET_GROWTH, IS_EPS, PX_VOLUME, RSI_30D, PE_RATIO. The dividend yield is evaluated as EQY_DPS divided by the lagged value of the closing price field PX_LAST. Momentum is also computed via the closing price (lagged 12 month value divided by lagged one month value, minus one).

Step	**REINFORCE**
0	Given a policy $\pi_{\boldsymbol{\theta}}$, a discount rate $\gamma \in (0,1)$ and a learning rate $\eta \in (0,1)$;
1	**For** i = 1, 2, ..., number of episodes, do:
2	Generate sequence S_0, A_0, R_1, ..., S_{T-1}, A_{T-1}, R_T, with actions driven by $\pi_{\boldsymbol{\theta}}$
3	**For** t = 0, 1, ..., $T-1$, do:
4	$G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ (compute the gain)
5	$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta\gamma^t G \nabla \log(\pi_{\boldsymbol{\theta}})$. (update the parameters)

Table 2: Steps of the baseline REINFORCE algorithm.

have an impact on the environment (the states). This is rarely the case in finance, except when taking very large orders, which large institutions usually avoid to limit the odds of market shifts. The generation of the SARSA sequences in Step 2 can thus be either **chronological** indeed, or independently drawn from samples of features and returns, akin to **bootstrapping**. In the latter case, the discounting rate γ would lose its meaning and should be set at one.

- **Step 2 & 4**: the definition of the reward R_t is not unambiguous. A natural choice is to take raw **returns**. The most prominent extension is when returns need to be adjusted by some risk measure, like in the **Sharpe ratio** (SR). However, the computation of the rewards in this case is not straightforward. Luckily, Moody et al. (1998) provide a solution to this obstacle. Their idea is to sequentially update the reward using the exponential moving average SR given in Equation (5).

  With this convention, the first steps of the SARSA sequence rapidly yield a risk adjusted reward.

- **Step 5**: this is not an option, but in the case of the linear policy (**F1**), the updated $\boldsymbol{\theta}$ has to be adjusted so that the a_n lie in the intervals discussed in Section B in the Appendix. Since we will work with bundles of $N = 100$ assets, we choose $a_- = 0.02$ and $a_+ = 1.6$. Thus, the feasible set in the projection (8) is

$$\Theta(\boldsymbol{X}_t) = \{\boldsymbol{\theta},\ a_- \leq \boldsymbol{\theta}\boldsymbol{X}_t \leq a_+\}.$$

3.3. Protocols

The above discussion gives rise to two dichotomies: **chronological** versus **bootstrapped** sequences and **return** versus **Sharpe ratio** rewards. Below, we explain how to incorporate these design choices in a series of thorough backtests that rely on market data (and not on simulated samples).

Chronological sequences require temporal depth. Every January (time t), the preceding 12 months of data are gathered and the sequences will consist of portfolios held during each month of the sample. The number of episodes is E and their length is 12. Each action at month s (A_s) consists in randomly choosing (with replacement) N assets and sampling their portfolio weights according to the current policy $\pi_{\boldsymbol{\theta}}$. The weights depend both on $\boldsymbol{\theta}$ and on the characteristics of the assets at month s. Rewards can either be returns, or Sharpe ratios, computed iteratively as defined in Equation (5).

Bootstrapped sequences do not require much depth. They can be performed every month. The number of episodes is E and their length is one: the learning is performed over values that originate from the past month only. This could be relaxed, but it creates more reactive portfolios, as opposed to learning on chronological sequences. Again, actions A_s consist in randomly choosing (with replacement) N assets and sampling their portfolio weights according to the current policy $\pi_{\boldsymbol{\theta}}$. Rewards can only be returns.

Both types of learning processes are summarized in Table 3. Because of the numerous degrees of freedom (γ and η rates, initialization values, random seeds, number of episodes, etc. - see Section 4.1 below), we restrict our study to two alternatives only. The first one links bootstrapped sequences with simple returns, while the second combines chronological sequences with Sharpe ratio rewards.

Step	**Chronological method**	Step	**Bootstrap method**
0	**For** every January, do:	0	**For** every date $t = 2, \dots, T-1$, do:
1	Randomly pick N assets	1	Randomly pick N assets
2	Extract data from previous year	2	Extract data from previous month
3	Initialize $\boldsymbol{\theta}$	3	Initialize $\boldsymbol{\theta}$
4	**For** i $= 1, \dots$ episodes, do	4	**For** i $= 1, \dots$ episodes, do
5	Sample N stocks randomly	5	Sample N stocks randomly
6	Generate streams A_t and R_t via $\pi_{\boldsymbol{\theta}}$	6	Generate action and reward
7	Update $\boldsymbol{\theta}$ via (9)	7	Update $\boldsymbol{\theta}$ via (9)
8	**For** next 12 months, do:	8	**For** date $t+1$, do:
9	allocate via average policy, Eq. (14)	9	allocate via average policy, Eq. (14)
10	store realized returns	10	store realized returns

Table 3: Macro view of backtest stages. The differences in the two REINFORCE implementations are outlined in Section 3.3.

4. Results

4.1. Degrees of freedom

Before we move towards a presentation of our results, we expose the richness of the flexibility of the modelling approach. Below, we list the different choices we need to make to run one batch of learning over our whole dataset:

- Choices that we will always compare:
  1. Whether to learn form chronological sequences or bootstrap (see Table 3).
  2. Whether to resort to a linear (**F1**) or an exponential (**F2**) policy.
- Choices that we will discuss:
  1. The number of episodes.
  2. The initialization values of $\boldsymbol{\theta}_0$.
  3. The seed for the quasi-random number generator.
  4. η, the learning rate in the update of the policy parameters. To simplify scale issues, the gradient in the update is divided by the maximum absolute value of gradient values. This makes the learning rate easier to interpret.
- Choices that are fixed throughout the entire study:
  1. γ, the reward discounting factor. For bootstrap learning, this parameter is irrelevant. For chronological sequences, since they only last 12 months, there is no major nor obvious gain in using a discount. Thus we set $\gamma = 1$.
  2. N, the number of stocks that are integrated in the portfolio (used to compute the reward). As discussed in Section B, it is impossible to consider very large portfolios because of the asymptotic behavior of the functions required in the Dirichlet forms. The most obvious choice is $N = 100$. Larger portfolios impose stringent constraints on the Dirichlet parameters, making the approach impractical. By construction, smaller portfolios lack diversification and may reflect cross-sectional information insufficiently.
  3. The bounds on the Dirichlet parameters (see Section B in the Appendix). They are fixed to $a_- = 0.2$ and $a_+ = 1.6$. These bounds are optimal empirically: going beyond leads to numerical errors.
  4. *Rewards*. Bootstrapped sequences can only work with simple return rewards. Chronological sequences are more flexible. To reduce the amount of results, we work with the differential Sharpe ratio for temporal learning.

4.2. Baseline output: factor coefficients and Dirichlet parameters

First and foremost, the Dirichlet policy depends on its parameter vector $\boldsymbol{\theta}$. It is thus natural to start by showing the evolution of the $\theta_t^{(k)}$ for the four specifications we work with. They are shown in Figure 2. While only a few cases are outlined, they are qualitatively representative of all the other parameter configurations studied below.

There is a clear discrepancy between the two learning schemes: chronological sequences (lower panels) lead to the hegemony of the constant variable while the bootstrapped sequences (upper plots) give more room to the firm characteristics. The latter are also much more volatile through time. Across both learning methods, the exponential policy (to the left) saturates the constant much more often, compared to the linear policy (to the right).

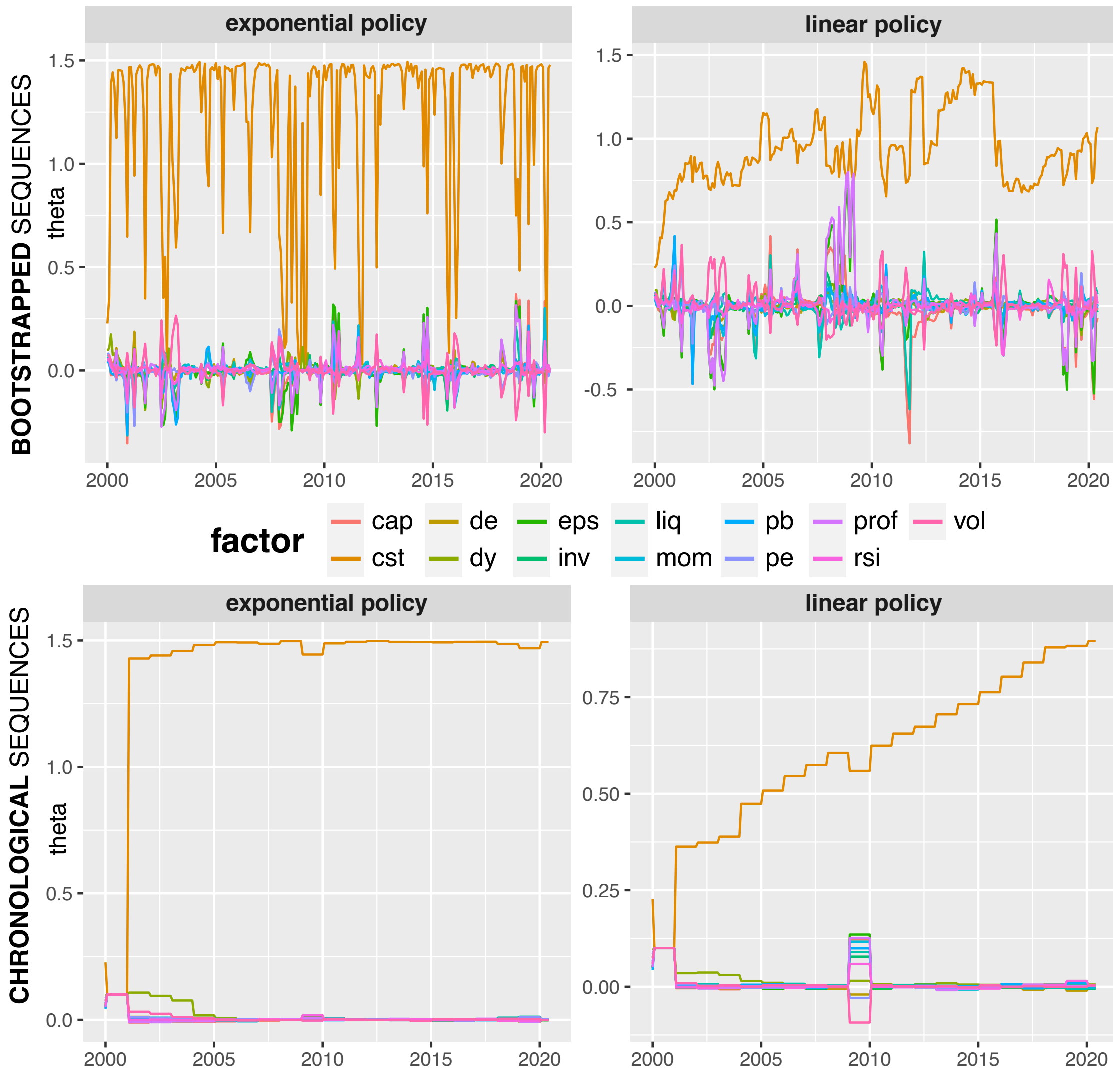

Fig. 2. **Values of $\theta_t^{(k)}$.** We plot the value of parameters through time for our two learning schemes (bootstrapped (upper panel) and chronological (middle panel)) and two policy schemes (linear (right) versus exponential (left) - see Equation (7)). The parameters are the following: the learning rate $\eta = 0.1$, the number of episodes $E = 500$, the bounds for the a_n are $[0.2, 1.6]$, the initial value for all θ^k is 1. Finally, the random seed in 42.

This has consequences on the optimal weights derived from the policy parameters via Equation (7). The chosen portfolio weights are simply chosen as the mean of the policy distributions: $w_{t,n} = a_{t,n}/\sigma_t$ (see (12) and (14)). In Figure 3, we plot the histogram of these weights. The distributions are grouped by year and then stacked on the graph. Because the number of assets changes through time (see Figure 1), we add two bounds on the plots. The full vertical black line marks the minimum uniform allocation ($1/N$), which is reached in 2016. The dotted line shows the maximum $1/N$ weights, which are implemented in 2001.

First of all, because of the increase in the number of stocks, there is a temporal shift in the

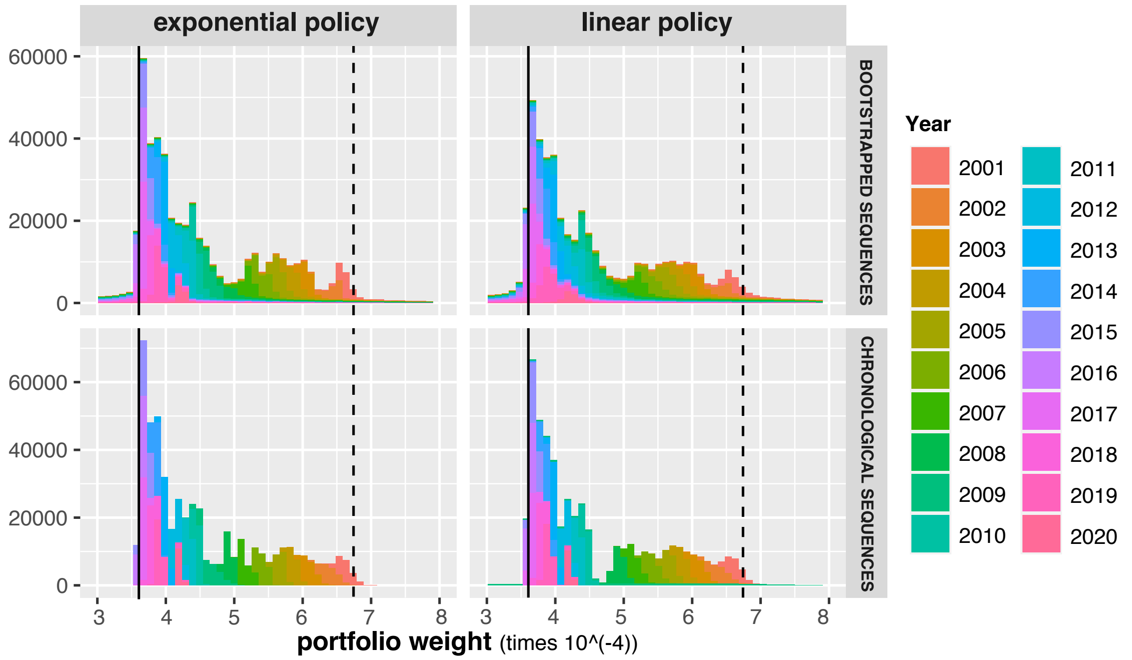

Fig. 3. **Distribution of weights**. We plot the histogram of portfolio weights for our two learning schemes (bootstrapped (upper panel) and chronological (middle panel)) and two policy schemes (linear (right) versus exponential (left) - see Equation (7)). The histograms are stacked and each color stands for a given year. The full vertical line marks the minimum uniform allocation ($1/N$) over all dates, while the dotted line shows the maximum $1/N$ value. The parameters are the following: the learning rate $\eta = 0.1$, the number of episodes $E = 500$, the bounds for the a_n are $[0.2, 1.6]$, the initial value for all θ^k is 1. Finally, the random seed in 42.

distribution of weights. Average weights are smaller in the later years and portfolios are more diversified. Moreover, weights are not very dispersed and appear concentrated around their means, which implies that allocations are relatively close to the EW benchmark and do not make strong bets towards some assets. This is especially true for the lower panel (chronological sequences), where there are almost no outliers beyond the vertical lines. This is consistent with the prominence of the constant in the lower panels of Figure 2.

Again, we underline that these results depend only marginally on the parametric choices described in the caption of the figures. The concentration of portfolios does not depend much implementation choices, as long as they are realistic (e.g., sufficiently many episodes, or moderate learning rate).[6]

4.3. Portfolio performance

The ultimate yardstick for sophisticated portfolio construction methods is out-of-sample performance. It is usually presented in several steps: starting with a pure return indicator, and complementing it by other risk-adjusted metrics, like the Sharpe ratio. In Figure 4, we display average

[6] Additional results are available upon request.

realized returns (left panels) and Sharpe ratios (right plots) of the mean policy for couples of values for random seeds, learning rate, and parameter initialization.

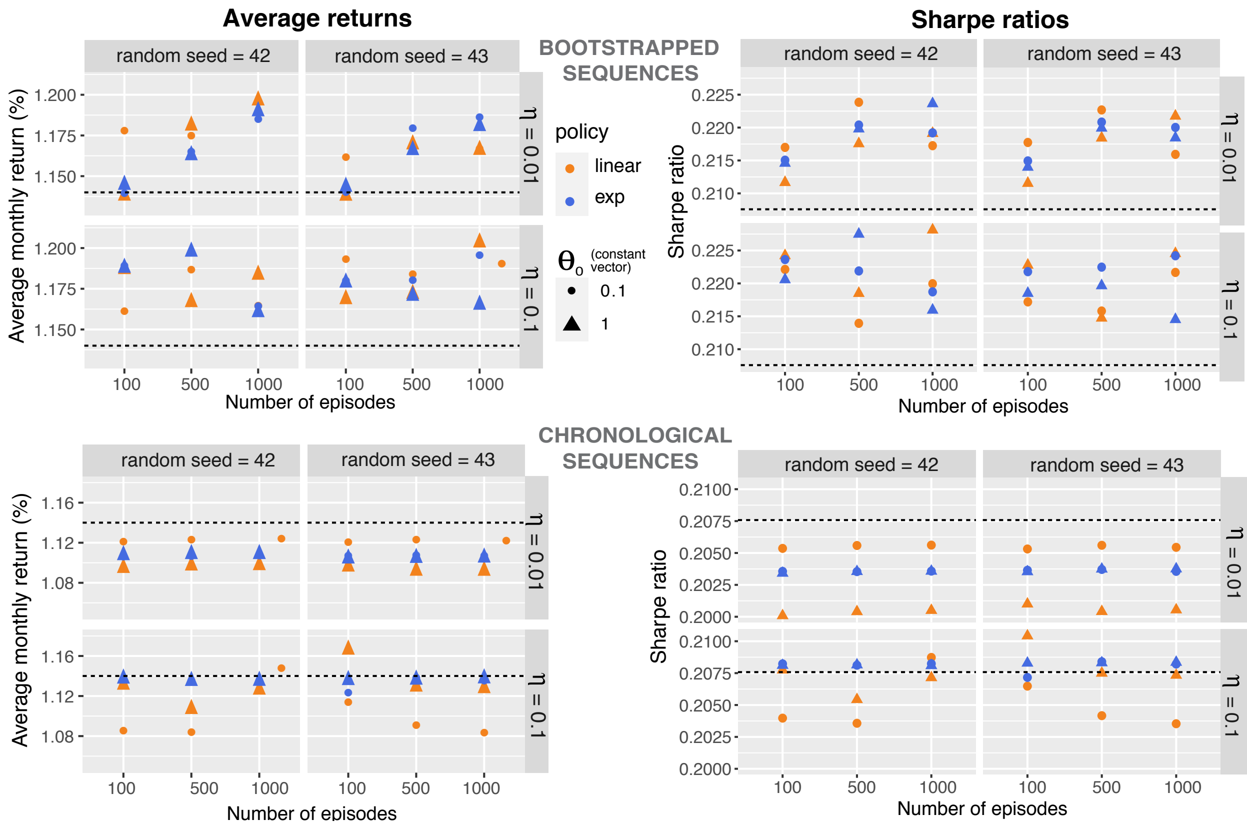

Fig. 4. **Performance.** We plot the average returns (left quadrants) and Sharpe ratios (right quadrants) of the portfolio allocations for our two learning schemes (bootstrapped (upper panel) and chronological (lower panel)), and two policy schemes (linear (orange points) versus exponential (blue points) - see Equation (7)). The dotted horizontal line marks the performance of the EW ($1/N$) portfolio.

The only robust conclusion is that bootstrap sequences perform better than sequential ones. Almost all bootstrap configurations surpass the EW benchmark, while it is the opposite for the portfolios based on chronological learning. The four degrees of freedom (random seed, number of episodes, learning rate and $\boldsymbol{\theta}_0$) have an impact on average returns that is not consistent across configurations. Notably, this corroborates the sensitivity of RL algorithms to random seeds (see, e.g., Henderson et al. (2017), Islam et al. (2017), and Colas et al. (2018)). Nevertheless, the magnitude of changes is small overall: average returns are scattered between 1.08% and 1.2%, so that the difference with the uniform allocation (1.14%) is not significant.[7] Thus, even though parameter configurations alter the results, the changes are very limited in magnitude and all portfolios remain somewhat in the vicinity of the $1/N$ benchmark.

[7] A simple t-test of series of RL-based portfolio returns versus $1/N$ portfolio returns yields $p-$values between 0.82 and 0.998.

5. Insights from a toy factor model

The purpose of this section is to elaborate on the patterns observed in Figure 2, which shows that the most important driver of the policies is the constant term. This preponderance implies that RL-based allocations remain in close to the equally-weighted portfolio. We propose a factor-based allocation model that tries to replicate this stylized property. In particular, we investigate if RL-driven decisions can be reproduced by simpler models. We start with a theoretical contribution and subsequently move towards simple statistical estimates to illustrate the forces at work.

5.1. The model

We assume that there are N assets on the market. Their future returns are driven by a linear model

$$\boldsymbol{r} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{15}$$

where $\boldsymbol{X} = X_{nk}$ is a $N \times (K+1)$ matrix of firm-specific characteristics with $N > K+1$. We omit the time index for notational simplicity. The first column of the matrix is constant with all elements equal to one. The innovations $\boldsymbol{\epsilon}$ and loadings $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_K)$ are random and mutually independent. Moreover, we posit that the errors are independent across assets (and independent of loadings), and have zero means and uniform variance:

$$\bar{\boldsymbol{\epsilon}} = \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}_N \tag{16}$$

$$\boldsymbol{\Sigma}_\epsilon = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal] = \sigma_\epsilon^2 \boldsymbol{I}_N, \tag{17}$$

where $\mathbf{0}_N$ is a N-dimensional vector of zeroes and $\boldsymbol{I}_N$ the corresponding identity matrix. For analytical tractability concerns, we also need to be more specific with regard to the covariance structure of loadings and characteristics. We assume that both

$$\boldsymbol{\Sigma}_\beta := \mathbb{E}[(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\intercal] \quad \text{and} \tag{18}$$

$$\hat{\boldsymbol{\Sigma}}_X := N^{-1}(\boldsymbol{X} - \mathbf{1}_N \bar{\boldsymbol{x}}^\intercal)^\intercal (\boldsymbol{X} - \mathbf{1}_N \bar{\boldsymbol{x}}^\intercal) \tag{19}$$

are diagonal with diagonal values equal to $\sigma^2_{\beta,k}$ and $\sigma^2_{X,k}$ respectively, where we have casually written $\bar{\boldsymbol{\beta}}$ for the mean vector of $\boldsymbol{\beta}$ and $\bar{\boldsymbol{x}} = N^{-1}\boldsymbol{X}^\intercal \mathbf{1}_N$ for the column vector of sample column means of $\boldsymbol{X}$. The function $\text{diag}(\cdot)$ maps a vector into the corresponding diagonal matrix. Note that since $\boldsymbol{X}$ is given and non-random, the matrix $\hat{\boldsymbol{\Sigma}}_X$ is its *sample* covariance matrix. The fact that $\hat{\boldsymbol{\Sigma}}_X$ is diagonal implies that the firm characteristics are uncorrelated, i.e., that they carry information pertaining to companies that is not redundant. The β_k are also unrelated, which means that each factor impacts returns regardless of the effect of other firm attributes.

Some representative agent seeks to maximize a standard quadratic function of expected returns. The portfolio is based on firms' characteristics in a linear fashion: $\boldsymbol{w} = \boldsymbol{X}\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_K)$ drives and reflects the agents beliefs and preferences with regard to the corresponding factors. This form is on purpose the same as **F1** in Equation (7) (Section 2.4), which drives the average portfolio allocation. The utility function is quadratic (as in the standard mean-variance formulation), hence, the optimization program is the following:

$$\max_{\boldsymbol{\theta}} \mathbb{E}\left[\boldsymbol{\theta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X}\boldsymbol{\beta} - \frac{\gamma}{2}\boldsymbol{\theta}^\intercal \boldsymbol{X}^\intercal(\boldsymbol{X}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) + \boldsymbol{\epsilon})(\boldsymbol{\epsilon}^\intercal + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\intercal \boldsymbol{X}^\intercal)\boldsymbol{X}\boldsymbol{\theta}\right], \quad \text{s.t.} \quad \boldsymbol{\theta}^\intercal \boldsymbol{X}^\intercal \mathbf{1}_N = 1. \tag{20}$$

The lemma below provides the solution to this problem.

Lemma 6. *If $\boldsymbol{\Sigma}_\beta$, $\boldsymbol{\Sigma}_\epsilon$ are diagonal and assuming (16)-(17), the solution to (20) is*

$$\tilde{\boldsymbol{\theta}}^* = \gamma^{-1}\text{diag}(\boldsymbol{\sigma}^2)^{-1}\left(\boldsymbol{I}_K - \frac{\text{diag}(\boldsymbol{\sigma}^2_\beta)\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'\text{diag}(\boldsymbol{\sigma}^2)^{-1}}{1+\bar{\boldsymbol{x}}'\text{diag}(\boldsymbol{\sigma}^2_\beta)\text{diag}(\boldsymbol{\sigma}^2)^{-1}\bar{\boldsymbol{x}}}\right)\left(N^{-1}\bar{\boldsymbol{\beta}} + c(\boldsymbol{X}'\boldsymbol{X})^{-1}\bar{\boldsymbol{x}}\right), \quad (21)$$

where $\text{diag}(\boldsymbol{\sigma}^2) = \text{diag}(\boldsymbol{\sigma}^2_X)\text{diag}(\boldsymbol{\sigma}^2_\beta) + N^{-1}\sigma^2_\epsilon\boldsymbol{I}_K$ *and c is the scaling constant that warrants the budget constraint is satisfied. If, in addition,* $\bar{\boldsymbol{x}}' = [1 \quad \boldsymbol{0}'_K]$, *then*

$$\tilde{\boldsymbol{\theta}}^* = \begin{cases} \tilde{\theta}^*_0 & = N^{-1} \\ \tilde{\theta}^*_j & = (\gamma N)^{-1}\frac{\bar{\beta}_j}{\sigma^2_{X,j}\sigma^2_{\beta,j}+\sigma^2_\epsilon/N} \end{cases}. \quad (22)$$

The proof of the lemma is located in Appendix A.5. All other things equal, in the second part of the lemma, θ^*_j increases with $\bar{\beta}_j$, but decreases with all sources of risk: $\sigma^2_{X,j}$, $\sigma^2_{\beta,j}$, and σ^2_ϵ. More importantly, the relative importance of the non-constant factors are strongly linked to γ. When risk aversion is low, the non-constant factors play a prominent role in the allocation choice. If, however, risk aversion is high, then the $\tilde{\theta}^*_j$ are negligible and the $1/N$ portfolio is appealing to the investor. Based on our empirical results, the latter situation seems more likely.

One particular subcase of the lemma is when the budget constraint (to the right of Equation (20)) is removed. In the general case, this implies $c = 0$ in (21). If non constant predictors have a zero mean, then $\tilde{\theta}^*_0$, like the other $\tilde{\theta}^*_j$, is given by the second part of (22). This configuration is interesting because in practice, θ_j values that are derived from RL algorithms are not subject to the budget constraint (see, e.g., step 5 in Table 2).

5.2. *Cross-sectional betas*

We illustrate the implications of Lemma 6 by running monthly regressions to estimate the loadings in Equation (15). To ease interpretability, we restrict the analysis to the three most common factors in the literature: size (market capitalization), value (price-to-book) and momentum (12 month to 1 month return). The largest correlation between them is 0.18 on the whole sample, thus the hypothesis of diagonal covariance matrix is not too far-fetched. Each month, we report the OLS coefficients for Equation (15), where $\boldsymbol{r}$ is the vector of *future* one month returns.

In the upper panel of Figure 5, we depict the estimated $\hat{\beta}_j$ for the three factors plus the constant. In addition, in the lower panel, we plot the scaled unconstrained theta values $\tilde{\theta}_j = \frac{\hat{\beta}_j/(2N)}{\hat{\sigma}^2_{X,j}\hat{\sigma}^2_{\beta,j}+\hat{\sigma}^2_\epsilon/N}$ (i.e., with $\gamma = 2$, which is without loss of generality, as γ is only a normalizing constant). One additional reason we resort to unconstrained θ_j values is that they do not depend on the risk aversion parameter, which only plays the role of a scaling factor.

All betas and unconstrained thetas oscillate strongly, but their means and deviations are telling. The θ_0 associated to the constant is volatile, but solidly positive on average, and by far dominating in magnitude, while the values for market capitalization are negative (which tends to be consistent with the so-called size anomaly).

5.3. *Comparing learning schemes*

Let us briefly summarize how agents allocate in the two frameworks (RL versus *loadings*-based):

Fig. 5. **Panel betas and scaled unconstrained thetas**. We plot the estimated panel betas $\hat{\beta}_j$ for each month in the dataset (upper panel). For each of the four features (3 factors + constant), we show the scaled values $\tilde{\theta}_j = \frac{\hat{\beta}_j/(2N)}{\hat{\sigma}^2_{X,j}\hat{\sigma}^2_{\beta,j}+\hat{\sigma}^2_{\epsilon}/N}$ in the lower panel. Note: the risk aversion parameter (scaling constant) is $\gamma = 2$.

- When resorting to RL, the agent learns (via the policy gradient) by computing the **sensitivity** of the performance metric (average return or Sharpe ratio) with respect to variations in the parameters of the policy ($\boldsymbol{\theta}$).
- In a more conventional characteristics-based asset pricing approach, the econometrician will evaluate the **exposure** of returns to firm specific attributes ($\boldsymbol{\beta}$). These exposures can then be translated into portfolio weights, when optimizing the average of a given utility function (see Lemma 6 when the utility is quadratic).

Theoretically, there are no reasons why these two approaches should be linked (they are hard to reconcile analytically, even though both seek to give more weight to assets that are expected to outperform - see Section 2.5). However, empirically, we find some consistency between the two methods. One common feature is the dominance of the constant in the upper panel of Figure 2 and in the lower panel of Figure 5. This indicates that both approaches find a strong common factor in the cross-section of returns which cannot be explained by firm level idiosyncasies. Heuristically, when the estimated $\hat{\beta}_0$ (which drives $\tilde{\theta}_0$) is high, returns are on average high in the cross-section, thus, the sensitivity of performance to variations in θ_0 should also be positive. This is when the orange line in Figure 2 is either at its ceiling, or increasing. When the policy learns over longer longer samples (bottom panel of Figure 2), then the long-term positivity of $\hat{\beta}_0$ (which is linked to that of the equity premium) pushes θ_0 upwards.

We thus wish to further investigate the link between the two learning processes. On the one hand, the driving element in the RL allocation is the policy update $\Delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1} = \alpha\widehat{\nabla J(\boldsymbol{\theta}_t)}$ (see Equation (9)). On the other hand, we pick the optimal (unconstrained) $\tilde{\boldsymbol{\theta}}_t$ in Equation (22) to proxy for the information that is processed by the asset pricing model during the period between $t-1$ and t. In Figure 6, we plot the first values (y-axis) against the second ones (x-axis).

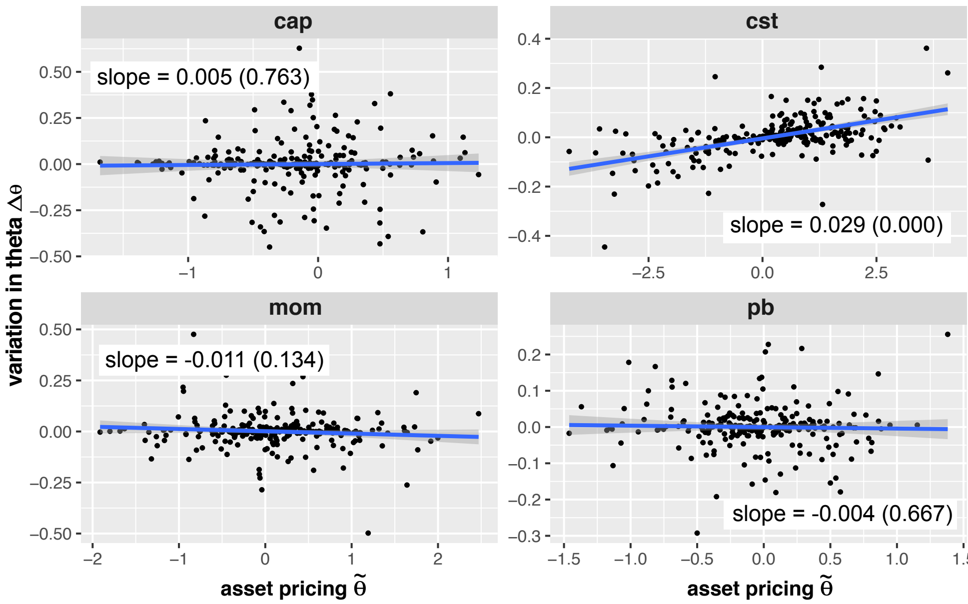

Fig. 6. **Policy gradients versus optimal asset pricing parameters**. We plot the variations in policy parameters $\Delta\boldsymbol{\theta}_t$ versus the unconstrained asset pricing based $\tilde{\boldsymbol{\theta}}_t^*$ defined in Equation (22) and plotted in Figure 5. The former correspond to the first policy form (**F1**) updated from bootstrapped returns. The latter are built with returns from month $t-1$ to t for consistency reasons (i.e., to match the informational set with which $\Delta\boldsymbol{\theta}_t$ is built). The slope of the fitted linear relationship is reported, along with the corresponding p-value (in brackets). Note: the risk aversion parameter (scaling constant) is $\gamma = 2$.

There is only one dimension for which the two schemes yield consistent values: the constant (upper right quadrant). On all purely firm-specific characteristics, the approaches seldom agree and the correlation between the two approaches is not significantly different from zero. Thus, apart from the relative agreement on the dominance of the common factor in the cross-section of stocks, it is hard to fully reconcile the two methods.

6. Conclusion

In this article, we combine reinforcement learning with factor investing. The investor learns from the impact of firm-specific characteristics on a chosen performance measure. The technical machinery relies on tractable expressions derived from analytical properties of the Dirichlet distribution. This allows to keep allocation decision inside a simplex, which is the unique requirement in long-only portfolios.

Empirically, the approach yields weights that are very diversified, akin to the 1/N portfolio. One interpretation is that the learning process captures the importance of a common factor that drives the cross-section of stocks beyond their factorial idiosyncrasies. We compare the RL decisions to those of a simple asset pricing model and find major difference for the non-constant characteristics. This shows the peculiarities stemming from the RL-based methodology.

Interestingly, and in spite of a wide range of implementation choices, the fact that the RL portfolios remain in the vicinity of uniform allocations underlines the efficiency of the latter.

Appendix A Proofs

A.1 Proof of Lemma 1

Case $l = 1$. We have

$$\begin{aligned}\int_{\mathcal{S}} \rho_{t+1} \mathrm{Prob}\left(ds \mid (\rho_t, \boldsymbol{X}_t), \boldsymbol{w}_t\right) &= \int_{\mathbb{R}} \int_{\mathcal{M}} r \boldsymbol{P}_\epsilon(\mathbb{T}^{-1}(dr \mid \boldsymbol{X}_t, \boldsymbol{w}_t)) \cdot \mathbb{P}_t^{t+1}\left(d\xi \mid \boldsymbol{X}_t\right) \\ &= \int_{\mathbb{R}} r \boldsymbol{P}_\epsilon(\mathbb{T}^{-1}(dr \mid \boldsymbol{X}_t, \boldsymbol{w}_t)) \\ &= \int_{\mathbb{R}^N} \mathbb{T}(\boldsymbol{e}) \boldsymbol{P}_\epsilon(d\boldsymbol{e}) \\ &= \int_{\mathbb{R}^N} (\boldsymbol{w}_t^\intercal (f(\boldsymbol{X}_t) + \boldsymbol{e}) \boldsymbol{P}_\epsilon(d\boldsymbol{e}) \\ &= \boldsymbol{w}_t^\intercal f(\boldsymbol{X}_t) + \boldsymbol{w}_t^\intercal \left(\int_{\mathbb{R}^N} \boldsymbol{e} \boldsymbol{P}_\epsilon(d\boldsymbol{e})\right)\end{aligned}$$

which gives the result as $\boldsymbol{\epsilon}_t$ is a zero mean White Noise.

Case $l \geq 2$. Note that $\rho_{t+l} = \mathbb{T}(\boldsymbol{X}_{t+l-1}, \boldsymbol{w}_{t+l-1}, \boldsymbol{\epsilon}_{t+1})$. Therefore,

$$\begin{aligned}\int_{\mathcal{S}} \rho_{t+l} \mathrm{Prob}\left(ds \mid (\rho_t, \boldsymbol{X}_t), \boldsymbol{w}_t\right) &= \int_{\mathbb{R}} \int_{\mathcal{M}} \rho_{t+l} \boldsymbol{P}_\epsilon(\mathbb{T}^{-1}(dr \mid \boldsymbol{X}_t, \boldsymbol{w}_t)) \cdot \mathbb{P}_t^{t+1}\left(d\xi \mid \boldsymbol{X}_t\right) \\ &= \int_{\mathcal{M}} \rho_{t+l} \mathbb{P}_t^{t+1}\left(d\xi \mid \boldsymbol{X}_t\right).\end{aligned}$$

If $l = 2$, we go to the next step, otherwise we iterate

$$\begin{aligned}\int_{\mathcal{S}} \rho_{t+l} \mathrm{Prob}\left(ds \mid (\rho_t, \boldsymbol{X}_t), \boldsymbol{w}_t\right) &= \int_{\mathcal{M}} \left(\int_{\Delta} \left(\int_{\mathcal{S}} \rho_{t+l} \mathrm{Prob}\left(ds' \mid r, \xi, \boldsymbol{w}_{t+1}\right)\right) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_{t+1} \mid \xi\right)\right) \mathbb{P}_t^{t+1}\left(d\xi \mid \boldsymbol{X}_t\right) \\ &= \int_{\mathcal{M}} \left(\int_{\mathcal{S}} \rho_{t+l} \mathrm{Prob}\left(ds' \mid r, \xi, \boldsymbol{w}_{t+1}\right)\right) \mathbb{P}_t^{t+1}(d\xi \mid \boldsymbol{X}_t) \\ &= \int_{\mathcal{M}} \left(\int_{\mathbb{R}} \int_{\mathcal{M}} \rho_{t+l} \boldsymbol{P}_\epsilon(\mathbb{T}^{-1}(dr \mid \xi, \boldsymbol{w}_{t+1})) \mathbb{P}_{t+1}^{t+2}(d\xi' \mid \xi)\right) \mathbb{P}_t^{t+1}(d\xi \mid \boldsymbol{X}_t) \\ &= \int_{\mathcal{M}} \left(\int_{\mathcal{M}} \rho_{t+l} \mathbb{P}_{t+1}^{t+2}(d\xi' \mid \xi)\right) \mathbb{P}_t^{t+1}(d\xi \mid \boldsymbol{X}_t) \\ &= \int_{\mathcal{M}} \rho_{t+l} \mathbb{P}_t^{t+2}(d\xi \mid \boldsymbol{X}_t)\end{aligned}$$

up to the point where

$$\int_{\mathcal{S}} \rho_{t+l} \mathrm{Prob}\left(ds \mid (\rho_t, \boldsymbol{X}_t), \boldsymbol{w}_t\right) = \int_{\mathcal{M}} \left(\int_{\Delta} \left(\int_{\mathcal{S}} \rho_{t+l} \mathrm{Prob}\left(ds' \mid r, \xi, \boldsymbol{w}_{t+l-1}\right)\right) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_{t+l-1} \mid \xi\right)\right) \mathbb{P}_t^{t+l-1}(d\xi \mid \boldsymbol{X}_t).$$

The innermost integral is given by the above case $l = 1$ hence

$$\begin{aligned}\int_{\mathcal{S}} \rho_{t+l} \mathrm{Prob}\left(ds \mid (\rho_t, \boldsymbol{X}_t), \boldsymbol{w}_t\right) &= \int_{\mathcal{M}} \left(\int_{\Delta} \boldsymbol{w}_{t+l-1}^\intercal f(\xi) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_{t+l-1} \mid \xi\right)\right) \mathbb{P}_t^{t+l-1}(d\xi \mid \boldsymbol{X}_t) \\ &= \int_{\mathcal{M}} \left(\int_{\Delta} \boldsymbol{w}_{t+l-1} \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_{t+l-1} \mid \xi\right)\right)^\intercal f(\xi) \mathbb{P}_t^{t+l-1}(d\xi \mid \boldsymbol{X}_t).\end{aligned}$$

A.2 Proof of Proposition 2

We first compute the expected values of the reward.

For the risk-insensitive agent, $R_t = \rho_t$, so that

$$\mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+l} \mid S_t = (\rho_t, \boldsymbol{X}_t)\right] = \int_\Delta \left(\int_{\mathcal{S}} \rho_{t+l} \mathrm{Prob}\left(ds \mid S_t, \boldsymbol{w}_t\right)\right) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_t \mid \boldsymbol{X}_t\right).$$

Lemma 1 gives the value of the inner integral and we have two cases. If $l = 1$, then

$$\mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+1} \mid S_t\right] = \int_\Delta \boldsymbol{w}_t^\intercal f\left(\boldsymbol{X}_t\right) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_t \mid \boldsymbol{X}_t\right) = \left(\int_\Delta \boldsymbol{w}_t \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_t \mid \boldsymbol{X}_t\right)\right)^\intercal f\left(\boldsymbol{X}_t\right).$$

If $l \geq 2$, we have

$$\mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+l} \mid S_t\right] = \int_\Delta \left(\int_{\mathcal{M}} \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+l-1} \mid \xi\right]^\intercal f\left(\xi\right) \mathbb{P}_t^{t+l-1}(d\xi \mid \boldsymbol{X}_t)\right) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_t \mid \boldsymbol{X}_t\right).$$

$\boldsymbol{w}_{t+l-1}$ is sampled under policy $\pi_{\boldsymbol{\theta}}$ when S_{t+l-1} is known and is independent of the previous states, while $\boldsymbol{r}_{t+l}$ is given by the state S_{t+l-1} and the realization of $\boldsymbol{\epsilon}_{t+l}$ and is independent of the choice of the action. Therefore the inner integral is not a function of $\boldsymbol{w}_t$ and

$$\mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+l} \mid S_t\right] = \int_{\mathcal{M}} \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+l-1} \mid \xi\right]^\intercal f\left(\xi\right) \mathbb{P}_t^{t+l-1}(d\xi \mid \boldsymbol{X}_t).$$

Now we show that the policy value takes a recursive form. First rewrite the policy value for the risk insensitive agent as

$$V^{\boldsymbol{\theta}}(t, S_t) = \sum_{l=1}^{T-t} \mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+l} \mid S_t\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+1} \mid S_t\right] + \sum_{l=1}^{T-(t+1)} \mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+1+l} \mid S_t\right].$$

Using the expression for the expected values of the reward for $l \geq 2$ that were computed in the proof of Lemma 1:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+1+l} \mid S_t\right] &= \int_{\mathcal{M}} \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+l} \mid \xi\right]^\intercal f\left(\xi\right) \mathbb{P}_t^{t+l}(d\xi \mid \boldsymbol{X}_t) \\
&= \int_{\mathcal{M}} \left(\int_{\mathcal{M}} \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+l} \mid \xi'\right]^\intercal f\left(\xi'\right) \mathbb{P}_{t+1}^{t+l}(d\xi' \mid \xi)\right) \mathbb{P}_t^{t+1}(d\xi \mid \boldsymbol{X}_t).
\end{aligned}$$

For $l = 1$, the inner integral is simply

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+1} \mid \xi\right]^\intercal f\left(\xi\right) = \mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+1+l} \mid S_{t+1} = \xi\right]$$

For $l \geq 2$, it is also

$$\int_{\mathcal{M}} \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+1+l-1} \mid \xi'\right]^\intercal f\left(\xi'\right) \mathbb{P}_{t+1}^{t+1+l-1}(d\xi' \mid \xi) = \mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+1+l} \mid S_{t+1} = \xi\right].$$

The rightmost part of the policy value expression can now be written as

$$\sum_{l=1}^{T-(t+1)} \mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+1+l} \mid S_t\right] = \int_{\mathcal{M}} \left(\sum_{l=1}^{T-(t+1)} \mathbb{E}_{\boldsymbol{\theta}}\left[R_{t+1+l} \mid S_{t+1} = \xi\right]\right) \mathbb{P}_t^{t+1}(d\xi \mid \boldsymbol{X}_t),$$

where the inner sum is $V^{\boldsymbol{\theta}}(t+1, \xi)$.

A.3 Proof of Proposition 3

A.3.1 Expectation of $\ln W_n$

We have $W_n \sim \text{Beta}\left(a_n, \sigma - a_n\right)$. Consequently,

$$\mathbb{E}\left[\ln W_n\right] = \frac{1}{B\left(a_n, \sigma - a_n\right)} \int_0^1 \ln w_n w_n^{a_n - 1}\left(1 - w_n\right)^{\sigma - a_n - 1} dw_n \quad = \frac{1}{B\left(a_n, \sigma - a_n\right)} \int_0^1 \ln w_n \varphi(w_n, a_n) dw_n,$$

where

$$\varphi(w_n, a_n) = w_n^{a_n - 1}\left(1 - w_n\right)^{\sigma - a_n - 1} = e^{(a_n - 1)\ln w_n + (\sigma - a_n - 1)\ln(1 - w_n)}.$$

As a_n cancels out in $\sigma - a_n$,

$$\frac{\partial}{\partial a_n}\varphi\left(w_n, a_n\right) = \ln w_n \varphi\left(w_n, a_n\right),$$

and

$$\begin{aligned}
\mathbb{E}\left[\ln W_n\right] &= \frac{1}{B\left(a_n, \sigma - a_n\right)} \int_0^1 \frac{\partial}{\partial a_n}\varphi\left(w_n, a_n\right) dw_n \\
&= \frac{1}{B\left(a_n, \sigma - a_n\right)} \frac{\partial}{\partial a_n} \int_0^1 \varphi\left(w_n, a_n\right) dw_n \\
&= \frac{1}{B\left(a_n, \sigma - a_n\right)} \frac{\partial}{\partial a_n} B\left(a_n, \sigma - a_n\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}a_n} \ln B\left(a_n, \sigma - a_n\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}a_n}\left(\ln\Gamma(a_n) + \ln\Gamma(\sigma - a_n) - \ln\Gamma(\sigma)\right) \\
&= \digamma\left(a_n\right) - \digamma\left(\sigma\right).
\end{aligned}$$

A.3.2 Expectation of $W_n \ln W_n$

We have $W_n \sim \text{Beta}\left(a_n, \sigma - a_n\right)$. Then,

$$\begin{aligned}
\mathbb{E}\left[W_n \ln W_n\right] &= \frac{1}{B\left(a_n, \sigma - a_n\right)} \int_0^1 w_n \ln w_n w_n^{a_n - 1}\left(1 - w_n\right)^{\sigma - a_n - 1} dw_n \\
&= \frac{1}{B\left(a_n, \sigma - a_n\right)} \int_0^1 \ln w_n \varphi(w_n, a_n) dw_n,
\end{aligned}$$

where

$$\varphi(w_n, a_n) = w_n^{a_n}\left(1 - w_n\right)^{\sigma - a_n - 1}.$$

In addition,

$$
\begin{aligned}
\mathbb{E}\left[W_n \ln W_n\right] &= \frac{1}{B\left(a_n, \sigma - a_n\right)} \int_0^1 \frac{\partial}{\partial a_n} \varphi\left(w_n, a_n\right) dw_n \\
&= \frac{1}{B\left(a_n, \sigma - a_n\right)} \frac{\partial}{\partial a_n} \int_0^1 \varphi\left(w_n, a_n\right) dw_n \\
&= \frac{1}{B\left(a_n, \sigma - a_n\right)} \frac{\partial}{\partial a_n} B\left(a_n + 1, \sigma - a_n\right) \\
&= \frac{B\left(a_n + 1, \sigma - a_n\right)}{B\left(a_n, \sigma - a_n\right)} \frac{\mathrm{d}}{\mathrm{d}a_n} \ln B\left(a_n + 1, \sigma - a_n\right).
\end{aligned}
$$

The ratio of Betas simplifies to

$$
\frac{B\left(a_n + 1, \sigma - a_n\right)}{B\left(a_n, \sigma - a_n\right)} = \frac{\Gamma\left(a_n + 1\right)\Gamma\left(\sigma - a_n\right)\Gamma\left(\sigma\right)}{\Gamma\left(\sigma + 1\right)\Gamma\left(a_n\right)\Gamma\left(\sigma - a_n\right)} = \frac{a_n \Gamma\left(a_n\right)\Gamma\left(\sigma\right)}{\sigma\Gamma\left(\sigma\right)\Gamma\left(a_n\right)} = \frac{a_n}{\sigma}.
$$

Therefore we can conclude

$$
\begin{aligned}
\mathbb{E}\left[W_n \ln W_n\right] &= \frac{a_n}{\sigma} \frac{\mathrm{d}}{\mathrm{d}a_n} \ln B\left(a_n + 1, \sigma - a_n\right) \\
&= \frac{a_n}{\sigma} \frac{\mathrm{d}}{\mathrm{d}a_n} \left(\ln\Gamma(a_n + 1) + \ln\Gamma(\sigma - a_n) - \ln\Gamma(\sigma + 1)\right) \\
&= \frac{a_n}{\sigma} \left(\digamma\left(a_n + 1\right) - \digamma\left(\sigma + 1\right)\right) \\
&= \frac{a_n}{\sigma} \left(\digamma\left(a_n\right) + \frac{1}{a_n} - \digamma\left(\sigma\right) - \frac{1}{\sigma}\right).
\end{aligned}
$$

A.3.3 Expectation of $W_n \ln W_m$

We have $\begin{bmatrix} W_n & W_m \end{bmatrix} \sim \mathrm{Dir}_3\left(\boldsymbol{a}_{n,m}\right)$, where $\boldsymbol{a}_{n,m} = \begin{bmatrix} a_n & a_m & \sigma - a_n - a_m \end{bmatrix}$. Also,

$$
\begin{aligned}
\mathbb{E}\left[W_n \ln W_m\right] &= \frac{1}{B\left(\boldsymbol{a}_{n,m}\right)} \int_0^1 \int_0^{1-w_n} w_n \ln w_m w_n^{a_n - 1} w_m^{a_m - 1} \left(1 - w_n - w_m\right)^{\sigma - a_n - a_m - 1} dw_n dw_m \\
&= \frac{1}{B\left(\boldsymbol{a}_{n,m}\right)} \int_0^1 w_n^{a_n} \left(\int_0^{1-w_n} \ln w_m w_m^{a_m - 1} \left(1 - w_n - w_m\right)^{\sigma - a_n - a_m - 1} dw_m\right) dw_n.
\end{aligned}
$$

The inner integral is

$$
I = \int_0^{\lambda} \ln w w^{a_m - 1} \left(\lambda - w\right)^{\sigma - a_n - a_m - 1} dw.
$$

With the change of variable $\lambda t = w$

$$
\begin{aligned}
I &= \int_0^1 \left(\ln\lambda + \ln t\right) \lambda^{a_m - 1} t^{a_m - 1} \lambda^{\sigma - a_n - a_m - 1} \left(1 - t\right)^{\sigma - a_n - a_m - 1} \lambda dt \\
&= \lambda^{\sigma - a_n - 1} \left(\ln\lambda \int_0^1 t^{a_m - 1} \left(1 - t\right)^{\sigma - a_n - a_m - 1} dt + \int_0^1 \ln t \; t^{a_m - 1} \left(1 - t\right)^{\sigma - a_n - a_m - 1} dt\right) \\
&= \lambda^{\sigma - a_n - 1} \left(\ln\lambda B\left(a_m, \sigma - a_n - a_m\right) + \int_0^1 \ln t \varphi\left(t, a_m\right) dt\right),
\end{aligned}
$$

where

$$\varphi\left(t,a_m\right)=t^{a_m-1}\left(1-t\right)^{\sigma-a_n-a_m-1}.$$

As a_m cancels out in $\sigma-a_n-a_m$,

$$\frac{\partial}{\partial a_m}\varphi\left(t,a_m\right)=\ln t\varphi\left(t,a_m\right).$$

Therefore,

$$\begin{aligned}\int_0^1\ln t\varphi\left(t,a_m\right)dt=\int_0^1\frac{\partial}{\partial a_m}\varphi\left(t,a_m\right)dt=\frac{\partial}{\partial a_m}\int_0^1\varphi\left(t,a_m\right)dt\\
=\frac{\mathrm{d}}{\mathrm{d}a_m}B\left(a_m,\sigma-a_n-a_m\right)\\
=B\left(a_m,\sigma-a_n-a_m\right)\frac{\mathrm{d}}{\mathrm{d}a_m}\ln B\left(a_m,\sigma-a_n-a_m\right),\end{aligned}$$

and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}a_m}\ln B\left(a_m,\sigma-a_n-a_m\right)&=\frac{\mathrm{d}}{\mathrm{d}a_m}\left(\ln\Gamma\left(a_m\right)+\ln\Gamma\left(\sigma-a_n-a_m\right)-\ln\Gamma\left(\sigma-a_n\right)\right)\\
&=\digamma\left(a_m\right)-\digamma\left(\sigma-a_n\right).\end{aligned}$$

This gives

$$I=\left(1-w_n\right)^{\sigma-a_n-1}B\left(a_m,\sigma-a_n-a_m\right)\left(\ln\left(1-w_n\right)+\digamma\left(a_m\right)-\digamma\left(\sigma-a_n\right)\right).$$

Back to the expectation,

$$\mathbb{E}\left[W_n\ln W_m\right]=\frac{B\left(a_m,\sigma-a_n-a_m\right)}{B\left(\mathbf{a}_{n,m}\right)}\int_0^1w_n^{a_n}\left(1-w_n\right)^{\sigma-a_n-1}\left(\ln\left(1-w_n\right)+\digamma\left(a_m\right)-\digamma\left(\sigma-a_n\right)\right)dw_n.$$

The ratio of Betas simplifies to

$$\begin{aligned}\frac{B\left(a_m,\sigma-a_n-a_m\right)}{B\left(\mathbf{a}_{n,m}\right)}&=\frac{\Gamma\left(a_m\right)\Gamma\left(\sigma-a_n-a_m\right)\Gamma\left(\sigma\right)}{\Gamma\left(\sigma-a_n\right)\Gamma\left(a_n\right)\Gamma\left(a_m\right)\Gamma\left(\sigma-a_n-a_m\right)}\\
&=\frac{\Gamma\left(\sigma\right)}{\Gamma\left(\sigma-a_n\right)\Gamma\left(a_n\right)}=\frac{1}{B\left(a_n,\sigma-a_n\right)},\end{aligned}$$

and the integral splits in two parts which are

$$\begin{aligned}I_1&=\left(\digamma\left(a_m\right)-\digamma\left(\sigma-a_n\right)\right)\int_0^1w_n^{a_n}\left(1-w_n\right)^{\sigma-a_n-1}dw_n\\
&=\left(\digamma\left(a_m\right)-\digamma\left(\sigma-a_n\right)\right)B(a_n+1,\sigma-a_n)\end{aligned}$$

and

$$\begin{aligned}I_2&=\int_0^1\ln\left(1-w_n\right)w_n^{a_n}\left(1-w_n\right)^{\sigma-a_n-1}dw_n\\
&=\int_0^1\frac{\partial}{\partial a_k}\left(w_n^{a_n}\left(1-w_n\right)^{\sigma-a_n-1}\right)dw_n\end{aligned}$$

for any $1 \leq k \leq N$, $k \neq n$. Thus,

$$\begin{aligned} I_2 &= \frac{\partial}{\partial a_k}\left(\int_0^1 w_n^{a_n}\left(1-w_n\right)^{\sigma-a_n-1} dw_n\right) \\ &= B\left(a_n+1, \sigma-a_n\right) \frac{\mathrm{d}}{\mathrm{d} a_k} \ln B\left(a_n+1, \sigma-a_n\right) \\ &= B\left(a_n+1, \sigma-a_n\right)\left(\digamma\left(\sigma-a_n\right)-\digamma(\sigma+1)\right). \end{aligned}$$

Putting everything together

$$\mathbb{E}\left[W_n \ln W_m\right]=\frac{B(a_n+1, \sigma-a_n)}{B\left(a_n, \sigma-a_n\right)}\left(\digamma\left(a_m\right)-\digamma\left(\sigma-a_n\right)+\digamma\left(\sigma-a_n\right)-\digamma(\sigma+1)\right).$$

Again, the ratio of Betas simplifies to

$$\frac{B(a_n+1, \sigma-a_n)}{B\left(a_n, \sigma-a_n\right)}=\frac{\Gamma\left(a_n+1\right) \Gamma\left(\sigma-a_n\right) \Gamma(\sigma)}{\Gamma(\sigma+1) \Gamma\left(a_n\right) \Gamma\left(\sigma-a_n\right)}=\frac{a_n \Gamma\left(a_n\right) \Gamma(\sigma)}{\sigma \Gamma(\sigma) \Gamma\left(a_n\right)}=\frac{a_n}{\sigma}.$$

Finally

$$\mathbb{E}\left[W_n \ln W_m\right]=\frac{a_n}{\sigma}\left(\digamma\left(a_m\right)-\digamma(\sigma)-\frac{1}{\sigma}\right).$$

A.4 Proof of Proposition 5

We have

$$\begin{aligned} \nabla J\left(\boldsymbol{\theta}_t\right) &= \mathbb{E}_\pi\left[G_t \nabla \ln \pi\left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right) \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right] \\ &= \sum_{l=1}^{T-t} \mathbb{E}_\pi\left[\rho_{t+l} \nabla \ln \pi\left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right) \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right] \\ &= \sum_{l=1}^{T-t} \int_\Delta\left(\int_{\mathcal{S}} \rho_{t+l} \nabla \ln \pi\left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right) \operatorname{Prob}\left(ds \mid S_t, \boldsymbol{w}_t\right)\right) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_t \mid \boldsymbol{X}_t\right) \\ &= \sum_{l=1}^{T-t} \int_\Delta\left(\int_{\mathcal{S}} \rho_{t+l} \operatorname{Prob}\left(ds \mid S_t, \boldsymbol{w}_t\right)\right) \nabla \ln \pi\left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_t \mid \boldsymbol{X}_t\right). \end{aligned}$$

The inner integrals are given by Lemma 1 and

$$\begin{aligned} \nabla J\left(\boldsymbol{\theta}_t\right) = & \int_\Delta \boldsymbol{w}_t^{\intercal} f\left(\boldsymbol{X}_t\right) \nabla \ln \pi\left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_t \mid \boldsymbol{X}_t\right) \\ & +\sum_{l=2}^{T-t} \int_\Delta\left(\int_{\mathcal{M}} \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+l-1} \mid \xi\right]^{\intercal} f(\xi)\, \mathbb{P}_t^{t+l-1}(d\xi \mid \boldsymbol{X}_t)\right) \nabla \ln \pi\left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_t \mid \boldsymbol{X}_t\right), \end{aligned}$$

as the inner integrals in the second part of the expression are independent of the choice of $\boldsymbol{w}_t$, we can write this second part as

$$\left(\int_\Delta \nabla \ln \pi\left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right) \pi_{\boldsymbol{\theta}}\left(d\boldsymbol{w}_t \mid \boldsymbol{X}_t\right)\right) \sum_{l=2}^{T-t} \int_{\mathcal{M}} \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{w}_{t+l-1} \mid \xi\right]^{\intercal} f(\xi)\, \mathbb{P}_t^{t+l-1}(d\xi \mid \boldsymbol{X}_t).$$

But it is immediate from equation (13) that

$$\int_{\Delta} \nabla \ln \pi \left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t \right) \pi_{\boldsymbol{\theta}} \left(d\boldsymbol{w}_t \mid \boldsymbol{X}_t \right) = 0.$$

Therefore,

$$\nabla J \left(\boldsymbol{\theta}_t \right) = \mathbb{E}_\pi \left[\boldsymbol{w}_t^\intercal f \left(\boldsymbol{X}_t \right) \nabla \ln \pi \left(\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t \right) \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t \right].$$

Let $f_{t,n}$ denotes the n-th element of the vector $f \left(\boldsymbol{X}_t \right)$. Then,

$$\begin{aligned}
\nabla J \left(\boldsymbol{\theta}_t \right) &= \mathbb{E}_\pi \left[\left(\sum_{m=1}^N w_{t,m} f_{t,m} \right) \left(\sum_{n=1}^N \left(F \left(\sigma_t \right) - F \left(a_{t,n} \right) + \ln w_{t,n} \right) \nabla a_{t,n} \right) \mid S_t, \boldsymbol{\theta}_t \right] \\
&= \sum_{n=1}^N \left(\mathbb{E}_\pi \left[\left(\sum_{m=1}^N w_{t,m} f_{t,m} \right) \left(F \left(\sigma_t \right) - F \left(a_{t,n} \right) + \ln w_{t,n} \right) \mid S_t, \boldsymbol{\theta}_t \right] \right) \nabla a_{t,n} \\
&= \sum_{n=1}^N \left(F \left(\sigma_t \right) - F \left(a_{t,n} \right) \right) \mathbb{E}_\pi \left[\sum_{m=1}^N w_{t,m} f_{t,m} \mid S_t, \boldsymbol{\theta}_t \right] \nabla a_{t,n} \\
&\quad + \sum_{n=1}^N \left(\mathbb{E}_\pi \left[\left(\sum_{m=1}^N w_{t,m} f_{t,m} \right) \ln w_{t,n} \mid S_t, \boldsymbol{\theta}_t \right] \right) \nabla a_{t,n}.
\end{aligned}$$

On the one hand we have, by the expectation of a Dirichlet distributed random vector,

$$\mathbb{E}_\pi \left[\sum_{m=1}^N w_{t,m} f_{t,m} \mid S_t, \boldsymbol{\theta}_t \right] = \sum_{m=1}^N \mathbb{E}_\pi \left[w_{t,m} \mid S_t, \boldsymbol{\theta}_t \right] f_{t,m} = \sum_{m=1}^N \frac{a_{t,m}}{\sigma_t} f_{t,m}.$$

On the other hand, using Proposition 3, we obtain

$$\begin{aligned}
\mathbb{E}_\pi \left[\left(\sum_{m=1}^N w_{t,m} f_{t,m} \right) \ln w_{t,n} \mid S_t, \boldsymbol{\theta}_t \right] &= \sum_{m=1}^N \mathbb{E}_\pi \left[w_{t,m} \ln w_{t,n} \mid S_t, \boldsymbol{\theta}_t \right] f_{t,m} \\
&= \sum_{\substack{m=1 \\ m \neq n}}^N \mathbb{E}_\pi \left[w_{t,m} \ln w_{t,n} \mid S_t, \boldsymbol{\theta}_t \right] f_{t,m} + \mathbb{E}_\pi \left[w_{t,n} \ln w_{t,n} \mid S_t, \boldsymbol{\theta}_t \right] f_{t,n} \\
&= \sum_{\substack{m=1 \\ m \neq n}}^N \frac{a_{t,m}}{\sigma_t} \left(F \left(a_{t,n} \right) - F \left(\sigma_t \right) - \frac{1}{\sigma_t} \right) f_{t,m} \\
&\quad + \frac{a_{t,n}}{\sigma_t} \left(F \left(a_{t,n} \right) + \frac{1}{a_{t,n}} - F \left(\sigma_t \right) - \frac{1}{\sigma_t} \right) f_{t,n} \\
&= \left(F \left(a_{t,n} \right) - F \left(\sigma_t \right) - \frac{1}{\sigma_t} \right) \sum_{m=1}^N \frac{a_{t,m}}{\sigma_t} f_{t,m} + \frac{1}{\sigma_t} f_{t,n}.
\end{aligned}$$

This yields

$$
\begin{aligned}
\nabla J(\boldsymbol{\theta}_t) = & \sum_{n=1}^{N} \left(\digamma(\sigma_t) - \digamma(a_{t,n})\right) \sum_{m=1}^{N} \frac{a_{t,m}}{\sigma_t} f_{t,m} \nabla a_{t,n} \\
& + \sum_{n=1}^{N} \left(\left(\digamma(a_{t,n}) - \digamma(\sigma_t) - \frac{1}{\sigma_t} \right) \sum_{m=1}^{N} \frac{a_{t,m}}{\sigma_t} f_{t,m} + \frac{1}{\sigma_t} f_{t,n} \right) \nabla a_{t,n} \\
= & \sum_{n=1}^{N} \left(f_{t,n} - \sum_{m=1}^{N} \frac{a_{t,m}}{\sigma_t} f_{t,m} \right) \frac{\nabla a_{t,n}}{\sigma_t}.
\end{aligned}
$$

Now note that

$$
\sum_{m=1}^{N} \frac{a_{t,m}}{\sigma_t} f_{t,m} = \mathbb{E}_{\pi}\left[\boldsymbol{w}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right]^{\intercal} f(\boldsymbol{X}_t) = \mathbb{E}_{\pi}\left[R_{t+1} \mid \boldsymbol{X}_t, \boldsymbol{\theta}_t\right],
$$

and that $f_{t,n}$ is the expected return of asset n between t and $t+1$ (by equation (2)).

A.5 Proof of lemma 6

The Lagrange formulation of the problem,

$$
L(\boldsymbol{\theta}) = \boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\intercal}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \frac{\gamma}{2}\boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\intercal}\mathbb{E}\left[(\boldsymbol{X}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})+\boldsymbol{\epsilon})(\boldsymbol{\epsilon}+\boldsymbol{X}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}))^{\intercal}\right]\boldsymbol{X}\boldsymbol{\theta} + \lambda(\boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\intercal}\mathbf{1}_N - 1) \tag{23}
$$

$$
\frac{\partial L}{\partial \boldsymbol{\theta}} = \boldsymbol{X}^{\intercal}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \gamma\boldsymbol{X}^{\intercal}(\boldsymbol{X}\boldsymbol{\Sigma}_{\beta}\boldsymbol{X}^{\intercal} + \sigma_{\epsilon}^2\boldsymbol{I}_N)\boldsymbol{X}\boldsymbol{\theta} + \lambda\boldsymbol{X}^{\intercal}\mathbf{1}_N \tag{24}
$$

leads, via the first order conditions, to the standard solution

$$
\begin{aligned}
\boldsymbol{\theta}^* &= \gamma^{-1}(\boldsymbol{X}^{\intercal}(\boldsymbol{X}\boldsymbol{\Sigma}_{\beta}\boldsymbol{X}^{\intercal} + \sigma_{\epsilon}^2\boldsymbol{I}_N)\boldsymbol{X})^{-1}(\boldsymbol{X}^{\intercal}\boldsymbol{X}\bar{\boldsymbol{\beta}} + c\boldsymbol{X}^{\intercal}\mathbf{1}_N), \\
&= \gamma^{-1}(\boldsymbol{\Sigma}_{\beta}\boldsymbol{X}^{\intercal}\boldsymbol{X} + \sigma_{\epsilon}^2\boldsymbol{I}_K)^{-1}(\bar{\boldsymbol{\beta}} + c(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal}\mathbf{1}_N)
\end{aligned} \tag{25}
$$

where c is a constant which ensures that the budget constraint (to the right of Equation 20) is fulfilled. Note that $\boldsymbol{X}^{\intercal}\boldsymbol{X}$ is nonsingular because the characteristics are not redundant and because $N > K+1$. For the sake of completeness, we derive the expressions for the first inverse matrice below.

From (19) and the definition of $\bar{\boldsymbol{x}}$, it holds that

$$
\boldsymbol{X}^{\intercal}\boldsymbol{X} = N(\hat{\boldsymbol{\Sigma}}_X + \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\intercal}) = N(\mathrm{diag}(\boldsymbol{\sigma}_X^2) + \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\intercal}), \tag{26}
$$

so that by the Sherman-Morrison formula, and because $\boldsymbol{\Sigma}_{\beta} = \mathrm{diag}(\boldsymbol{\sigma}_{\beta}^2)$,

$$
\begin{aligned}
(\boldsymbol{\Sigma}_{\beta}\boldsymbol{X}^{\intercal}\boldsymbol{X} + \sigma_{\epsilon}^2\boldsymbol{I}_K)^{-1} &= \left(N\mathrm{diag}(\boldsymbol{\sigma}_X^2)\mathrm{diag}(\boldsymbol{\sigma}_{\beta}^2) + N\mathrm{diag}(\boldsymbol{\sigma}_{\beta}^2)\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\intercal} + \sigma_{\epsilon}^2\boldsymbol{I}_K\right)^{-1} \\
&= N^{-1}\mathrm{diag}(\boldsymbol{\sigma}^2)^{-1}\left(\boldsymbol{I}_K - \frac{\mathrm{diag}(\boldsymbol{\sigma}_{\beta}^2)\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\intercal}\mathrm{diag}(\boldsymbol{\sigma}^2)^{-1}}{1 + \bar{\boldsymbol{x}}^{\intercal}\mathrm{diag}(\boldsymbol{\sigma}^2)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_{\beta}^2)\bar{\boldsymbol{x}}}\right),
\end{aligned} \tag{27}
$$

where $\mathrm{diag}(\boldsymbol{\sigma}^2) = \mathrm{diag}(\boldsymbol{\sigma}_X^2)\mathrm{diag}(\boldsymbol{\sigma}_{\beta}^2) + N^{-1}\sigma_{\epsilon}^2\boldsymbol{I}_K$ - this form being a strong echo of the structure in Equation (15). This proves the first point.

Now, let us make the extreme simplification, as in our empirical section, that $\bar{\boldsymbol{x}}^\intercal = [1 \quad \mathbf{0}_K^\intercal]$, so that firm characteristics have zero sample mean - apart for the first one. This is not uncommon in the recent literature as long as the data is preprocessed (see Freyberger et al. (2020), Gu et al. (2020b), Kelly et al. (2019) and Koijen and Yogo (2019)). Then, $\boldsymbol{X}^\intercal\boldsymbol{X} = N\text{diag}(\tilde{\boldsymbol{\sigma}}_X^2)$, where the modified vector of variances $\tilde{\boldsymbol{\sigma}}_X^2$ is simply $\boldsymbol{\sigma}_X^2$ with the first element equal to one (instead of zero). The ratio in (27) vanishes (because $\sigma_{\beta,1}^2 = 0$ and $\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal$ is empty apart for its unit first element) and

$$\boldsymbol{\theta}^* = (N\gamma)^{-1}\text{diag}(\boldsymbol{\sigma}^2)^{-1}(\bar{\boldsymbol{\beta}} + c\,\text{diag}(\tilde{\boldsymbol{\sigma}}_X^2)^{-1}\bar{\boldsymbol{x}}),$$

from which the lower part of (22) is derived (the constant c impacts only θ_0^*). In this case, the budget constraint is only binding for the first asset. Indeed, because $\mathbf{1}_N\boldsymbol{X} = \bar{\boldsymbol{x}} = [1 \quad \mathbf{0}_K^\intercal]^\intercal$, the sum of weights linked to the non-constant factors is always equal to zero. Thus, θ_0, which is linked to a constant column, must satisfy $\theta_0^* N = 1$.

Appendix B Dirichlet distributions and portfolios in high dimensions

One major issue with the Dirichlet distribution is the computation of the scaling constant in high dimension. More precisely, let us consider the log of this quantity:

$$c = \log(B(\boldsymbol{a})) = \sum_{n=1}^{N} \log(\Gamma(a_n)) - \log\left(\Gamma\left(\sum_{n=1}^{N} a_n\right)\right). \tag{28}$$

When N is large and the a_n are free, both terms can reach levels that are beyond what machines can handle when exponentiated. Thus, we need to impose restrictions. We do it in two steps. First, we set some lower and upper bound on the a_n. In a second stage, we compute an upper value for N that will depend on the range of the a_n. This second step is the most technical and we provide the details below. The third and last step is to determine a tradeoff.

Before we continue, we recall that the a_n dictate the allocation of the agent and that, on average, the position in asset n is equal to $a_n\left(\sum_{n=1}^N a_n\right)^{-1}$. For obvious risk-management reasons, it is preferable to diversify portfolios. In our framework, we assume that there exists a constant $\delta > 1$ such that:

$$\frac{1}{\delta N} \le a_n \left(\sum_{n=1}^{N} a_n\right)^{-1} \le \frac{\delta}{N}, \quad n = 1, \dots, N. \tag{29}$$

In practice, the minimum value of δ will be driven by the data, and we discuss realistic ranges below. This constraint helps measure if the portfolio is on average well balanced and does not make extreme bets. The smaller δ is, the higher the diversification of the positions. Notably, under condition (29), the mean of the a_n, m_a, is such that

$$\delta^{-1}a_+ \le m_a \le \delta a_-, \text{ with } a_+ = \max_n a_n, \quad a_- = \min_n a_n, . \tag{30}$$

To further explicit our idea, we fix a maximum threshold $\kappa_{\max}$ beyond which we consider that the two terms in Equation (28) have numerically exploded. The two terms in Equation (28) have very different asymptotics when the a_n are large or small, hence the treatment is not symmetric.

We start with problems when the a_n are large. Given the strong convexity of the Γ function, this is more impactful for the second term in (28). We seek an upper bound a_+ for the a_n such that this second term remains below $\kappa_{\max}$, i.e.,

$$\log\left(\Gamma\left(\sum_{n=1}^{N} a_n\right)\right) \leq \kappa_{\max}.$$

Although the inverse of the Γ function exists (at least when its argument is large enough, see Uchiyama (2012)), it is not straightforward to compute. We thus resort to Stirling's formula instead and seek to simplify

$$\log\left(\sqrt{2\pi\left(\sum_{n=1}^{N} a_n - 1\right)}\left(\frac{\sum_{n=1}^{N} a_n - 1}{e}\right)^{\sum_{n=1}^{N} a_n - 1}\right) \leq \kappa_{\max}. \tag{31}$$

If we omit the first negligible term inside the square root, this is equivalent to

$$\left(\sum_{n=1}^{N} a_n - 1\right)\left(\log\left(\sum_{n=1}^{N} a_n - 1\right) - 1\right) \leq \kappa_{\max}.$$

As a first order (rough) approximation, we reduce this expression to

$$\left(\sum_{n=1}^{N} a_n\right)\log\left(\sum_{n=1}^{N} a_n\right) \leq \kappa_{\max},$$

that is, $\sum_{n=1}^{N} a_n \leq \frac{\kappa_{\max}}{W(\kappa_{\max})} \sim \frac{\kappa_{\max}}{\log(\kappa_{\max})}$, where W is the principal branch of the Lambert function. Its asymptotic behavior for large arguments is indeed $W(z) \sim \log(z)$ (see Section 4.13 in Olver et al. (2010)). Given (30), a rule of thumb constraint that links N and a_+ is

$$a_+ \leq \frac{\delta}{N}\frac{\kappa_{\max}}{\log(\kappa_{\max})} \Leftrightarrow N \leq \frac{\delta\kappa_{\max}}{a_+\log(\kappa_{\max})}, \tag{32}$$

where we purposefully underline that the condition can also be viewed as a limit on the number of assets.

The first term in 28 relates to the lower bound on the a_n. Indeed, as z shrinks to zero, $\Gamma(z)$ is equivalent to z^{-1}. Thus, if the a_n are small and a_- is sufficiently close to zero,

$$\sum_{n=1}^{N} \log(\Gamma(a_n)) \leq N\log\left(\frac{1}{a_-}\right) \leq \kappa_{\max} \quad \Leftrightarrow \quad N \leq \kappa_{\max}/\log(a_-^{-1}) \quad \Leftrightarrow \quad a_- \geq e^{-\kappa_{\max}/N}. \tag{33}$$

Conditions (32) and (33) link the bounds of the a_n to the number of assets N. In Figure 7, we illustrate them by assigning values to $\kappa_{\max}$, δ and N. Taking $\kappa_{\max} = 100$ allows $B(\boldsymbol{a})$ to range from e^{-100} to e^{100}, which is a large magnitude. In the figure, as the number of assets increases, the range of the a_n shrinks.

For our empirical study, we pick $a_- = 0.02$ and $a_+ = 1.6$. These values are optimal empirically because we obtain errors outside this range.

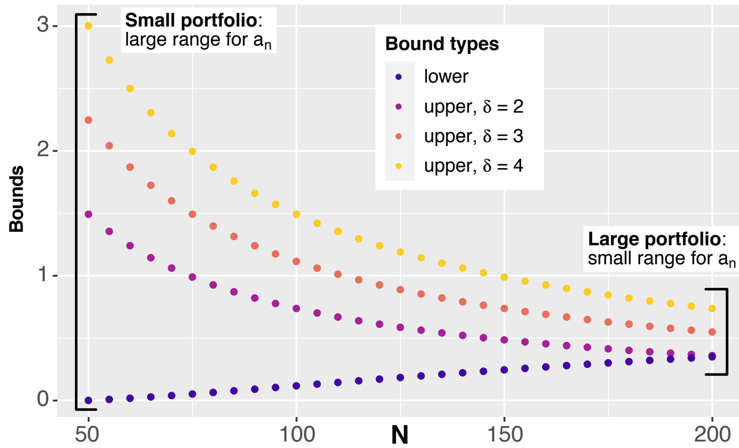

Fig. 7. **Intervals for the a_n.** We show the lower (a_-) and upper (a_+) bound for the a_n when the number of assets is fixed to 50, 100 or 200 and $\kappa_{\max} = 100$. They are derived from Equations (32) and (33). The black line is the Γ function.

References

Ammann, M., G. Coqueret, and J.-P. Schade (2016). Characteristics-based portfolio choice with leverage constraints. *Journal of Banking & Finance 70*, 23–37.

Arjovsky, M., L. Bottou, I. Gulrajani, and D. Lopez-Paz (2019). Invariant risk minimization. *arXiv Preprint* (1907.02893).

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen (2013). Value and momentum everywhere. *Journal of Finance 68*(3), 929–985.

Baker, M., B. Bradley, and J. Wurgler (2011). Benchmarks as limits to arbitrage: Understanding the low-volatility anomaly. *Financial Analysts Journal 67*(1), 40–54.

Ball, R. and P. Brown (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research*, 159–178.

Ball, R. and P. Brown (2019). Ball and Brown (1968) after fifty years. *Pacific-Basin Finance Journal 53*, 410–431.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics 9*(1), 3–18.

Barbee Jr, W. C., S. Mukherji, and G. A. Raines (1996). Do sales–price and debt–equity explain stock returns better than book–market and firm size? *Financial Analysts Journal 52*(2), 56–60.

Basu, S. (1983). The relationship between earnings' yield, market value and return for nyse common stocks: Further evidence. *Journal of Financial Economics 12*(1), 129–156.

Bäuerle, N. and U. Rieder (2011). *Markov Decision Processes with Applications to Finance.* Universitext. Berlin Heidelberg: Springer-Verlag.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *Journal of Finance 43*(2), 507–528.

Brandt, M. W., P. Santa-Clara, and R. Valkanov (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *Review of Financial Studies 22*(9), 3411–3447.

Chen, L., M. Pelger, and J. Zhu (2019). Deep learning in asset pricing. *SSRN Working Paper 3350138*.

Chordia, T. and B. Swaminathan (2000). Trading volume and cross-autocorrelations in stock returns. *Journal of Finance 55*(2), 913–935.

Colas, C., O. Sigaud, and P.-Y. Oudeyer (2018). How many random seeds? Statistical power analysis in deep reinforcement learning experiments. *arXiv Preprint* (1806.08295).

Cooper, M. J., H. Gulen, and M. J. Schill (2008). Asset growth and the cross-section of stock returns. *Journal of Finance 63*(4), 1609–1651.

Cover, T. M. and E. Ordentlich (1996). Universal portfolios with side information. *IEEE Transactions on Information Theory 42*(2), 348–363.

Daniel, K. and S. Titman (1997). Evidence on the characteristics of cross sectional variation in stock returns. *Journal of Finance 52*(1), 1–33.

Deng, Y., F. Bao, Y. Kong, Z. Ren, and Q. Dai (2016). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems 28*(3), 653–664.

Easton, P. D. (2004). Pe ratios, peg ratios, and estimating the implied expected rate of return on equity capital. *Accounting Review 79*(1), 73–95.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *Journal of Finance 47*(2), 427–465.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Feng, G., N. G. Polson, and J. Xu (2019). Deep learning in characteristics-sorted factor models. *SSRN Working Paper 3243683*.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies 33*(5), 2326–2377.

Gu, S., B. T. Kelly, and D. Xiu (2020a). Autoencoder asset pricing models. *Journal of Econometrics forthcoming*.

Gu, S., B. T. Kelly, and D. Xiu (2020b). Empirical asset pricing via machine learning. *Review of Financial Studies 33*(5), 2223–2273.

Han, Y., K. Yang, and G. Zhou (2013). A new anomaly: The cross-sectional profitability of technical analysis. *Journal of Financial and Quantitative Analysis 48*(5), 1433–1461.

Haugen, R. A. and N. L. Baker (1996). Commonality in the determinants of expected stock returns. *Journal of Financial Economics 41*(3), 401–439.

Henderson, P., R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger (2017). Deep reinforcement learning that matters. *arXiv Preprint* (1709.06560).

Hjalmarsson, E. and P. Manchev (2012). Characteristic-based mean-variance portfolio choice. *Journal of Banking & Finance 36*(5), 1392–1401.

Hoi, S. C., D. Sahoo, J. Lu, and P. Zhao (2018). Online learning: A comprehensive survey. *arXiv Preprint* (1802.02871).

Islam, R., P. Henderson, M. Gomrokchi, and D. Precup (2017). Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. *arXiv Preprint* (1708.04133).

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance 48*(1), 65–91.

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Koijen, R. S. and M. Yogo (2019). A demand system approach to asset pricing. *Journal of Political Economy 127*(4), 1475–1515.

Korsos, L. F. (2013). The Dirichlet portfolio model: Uncovering the hidden composition of hedge fund investments. *arXiv Preprint* (1306.0938).

Le Courtois, O. and X. Xu (2019). Portfolio optimization in the presence of extreme risks: A pareto-dirichlet approach. *SSRN Working Paper 3376921*.

Lettau, M. and M. Pelger (2020a). Estimating latent asset-pricing factors. *Journal of Econometrics 218*(1), 1–31.

Lettau, M. and M. Pelger (2020b). Factors that fit the time series and cross-section of stock returns. *Review of Financial Studies 33*(5), 2274–2325.

Li, Y., W. Zheng, and Z. Zheng (2019). Deep robust reinforcement learning for practical algorithmic trading. *IEEE Access 7*, 108014–108022.

Litzenberger, R. H. and K. Ramaswamy (1982). The effects of dividends on common stock prices tax effects or information effects? *Journal of Finance 37*(2), 429–443.

Moody, J., L. Wu, Y. Liao, and M. Saffell (1998). Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting 17*(5-6), 441–470.

Naranjo, A., M. Nimalendran, and M. Ryngaert (1998). Stock returns, dividend yields, and taxes. *Journal of Finance 53*(6), 2029–2057.

Neuneier, R. (1996). Optimal asset allocation using adaptive dynamic programming. In *Advances in Neural Information Processing Systems*, pp. 952–958.

Olver, F. W., D. W. Lozier, R. F. Boisvert, and C. W. Clark (2010). *NIST handbook of mathematical functions*. Cambridge university press.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference. Second Edition*, Volume 29. Cambridge University Press.

Pfister, N., P. Bühlmann, and J. Peters (2019). Invariant causal prediction for sequential data. *Journal of the American Statistical Association 114*(527), 1264–1276.

Sato, Y. (2019). Model-free reinforcement learning for financial portfolios: A brief survey. *arXiv Preprint* (1904.04973).

Sosnovskiy, S. (2015). On financial applications of the two-parameter Poisson-Dirichlet distribution. *arXiv Preprint* (1501.01954).

Sutton, R. S. and A. G. Barto (2018). *Reinforcement learning: An introduction (2nd Edition)*. MIT press.

Uchiyama, M. (2012). The principal inverse of the gamma function. *Proceedings of the American Mathematical Society 140*(4), 1343–1348.

Wang, H. and X. Y. Zhou (2020). Continuous-time mean-variance portfolio selection: A reinforcement learning framework. *Mathematical Finance Forthcoming.*

Zhang, Z., S. Zohren, and S. Roberts (2019). Deep reinforcement learning for trading. *arXiv Preprint* (1911.10107).