Proximal Policy Optimisation for a Private Equity Recommitment System

Emmanuel Kieffer[1], Frédéric Pinel[1], Thomas Meyer[3], Georges Gloukoviezoff[2], Hakan Lucius[2], and Pascal Bouvry[1]

[1] University of Luxembourg, Esch-sur-Alzette, Luxembourg
firstname.lastname@uni.lu
[2] European Investment Bank, Luxembourg, Luxembourg
{g.gloukoviezoff,h.lucius}@eib.org
[3] SimCorp Luxembourg SA, Luxembourg, Luxembourg
thomas.meyer@simcorp.com

Abstract. Recommitments are essential for limited partner investors to maintain a target exposure to private equity. However, recommitting to new funds is irrevocable and expose investors to cashflow uncertainty and illiquidity. Maintaining a specific target allocation is therefore a tedious and critical task. Unfortunately, recommitment strategies are still manually designed and few works in the literature have endeavored to develop a recommitment system balancing opportunity cost and risk of default. Due to its strong similarities to a control system, we propose to "learn how to recommit" with Reinforcement Learning (RL) and, more specifically, using Proximal Policy Optimisation (PPO). To the best of our knowledge, this is the first attempt a RL algorithm is applied to private equity with the aim to solve the recommitment problematic. After training the RL model on simulated portfolios, the resulting recommitment policy is compared to state-of-the-art strategies. Numerical results suggest that the trained policy can achieve high target allocation while bounding the risk of being overinvested.

Keywords: Reinforcement learning · Private Equity · Control system.

1 Introduction

Private equity is an alternative asset class which refers to direct investments in non-listed companies made at different stages of their development to create added value. These companies are then sold few years later with the expectation to obtain a significant capital gain. Early investments in strong performing companies help them to develop their business and make them more profitable. Contrary to the public equity market, private equity investments are not easily accessed as stocks and bonds. Recently, private equity has been included in the portfolios of institutional investors such as pension funds, sovereign wealth funds, etc. These institutional investors have been building sizable allocation by investing "indirectly" to private companies through private equity funds. Indeed,

managing such a less traditional asset class requires a high level of expertise to properly enter and exit direct investments. This explains their preferred modus operandi to invest indirectly as so-called limited partners (LP) through limited partnership funds in which they commit a certain amount of capital for a given period of time. Commitments are irrevocable and called at the discretion of the fund's management, i.e., the general partner (GP), to decide how investments should be realised. The committed capital is gradually draw down during the so-called investment period which last several years. To complicate matters, stakes in these funds are illiquid [7] which enforce LP investors to be extremely cautious when it comes to recommit into new funds to limit the risk of default. Generally, the committed capital is an upper-bound of the total capital finally called by a fund. A significant part ($\approx 10\%$) of the initial capital is generally never invested as described in [18]. Furthermore, committed capital waiting to be called is generally pictured as dry powder. Prequin [4] reported in November 2020 that North American private equity firms are sitting on almost $980bn in reserves. This uncalled capital dramatically impacts investors' exposure (see [12]). In practice, LP investors therefore run so-called overcommitment strategies, i.e., committing more capital in aggregate than actually available as dedicated resources, with the gap expected to be filled by future distributions from investments made in other existing funds. These strategies thus increase the liquidity risk when the fund is only few years old when the likelihood to be called is the highest. LP investors need to setup a commitment-pacing strategy, i.e., on how to size and time their commitments, in order to achieve and maintain a target allocation while complying with the liquidity constraints imposed by the uncalled capital. As reported in [3] and [9], few investigations have been engaged to evaluate the cost of maintaining uncalled capital. This is the reason why the current existing models still remain rudimentary and depend on spreadsheet-based and "trial-and-error" approaches. These manually-designed strategies are often error-prone and naive although the opportunity cost, i.e., the cost of being underinvested, and the risk of default in case of overinvestment can be very damaging for LP investors.

In this work, we propose to investigate an approach relying on Reinforcement Learning to learn how to size and time dynamic recommitments. The latter can be formulated as a RL problem to discover reliable recommitment policies using a Proximal Policy Optimisation algorithm. Recommitment policies can be assimilated as control policies which should maintain a target allocation minimizing the opportunity cost while preserving investors from the risk of default.

The remainder of this paper is organized as follows. The next section provides a state of the art on existing recommitment strategies. Section 3 introduces formally the Private Equity Recommitment Problem (PERP). Section 4 described the Proximal Policy Optimisation algorithm applied on the RL version of the PERP introduced in Section 5. Experiment setups and results are discussed in

[4] https://www.preqin.com/insights/research/blogs/what-private-equitys-record-dry-powder-haul-means-for-the-industry

Section 6 and 7. Finally, the last section provides our conclusions and proposes some possible perspectives.

2 Related works

Recommitment strategies are essential to keep investors constantly invested at some target allocation. To the best of our knowledge, few studies have tried to model this as an optimisation problem. They generally rely on some rules of thumb lacking robustness and flexibility. In [4], authors considered that the entire private equity allocation should be recommitted to new funds every year without taking into account past portfolios evolution. Nevin et al. in [11] based their recommitment strategy on average rates of distributions and commitments. New commitments should be made if the committed capital does not reach a target threshold to compensate the difference. This strategy assumes constants rates which seems very illusory over time. In [18], de Zwart et al. proposed recommitment strategies for funds aiming to maintain stable the exposure to PE. The strategy's key feature is the level of new commitments in a given period which depends on the current portfolio's characteristics. Importantly, de Zwart's strategies does not require to forecast funds'cashflows. Although they consider 100% PE portfolios, their last suggested strategy is a first attempt to design dynamic recommitment strategies relying on past portfolio development. Finally, Oberli et al. in [12] extended de Zwart's work to multi-asset class portfolios including stocks and bonds. These two last contributions solely rely on handcrafted recommitment strategies to control the investment degree (ID), i.e., PE exposure. While they are innovative and improving attempts without the need to forecast future cashflows, they have been built on specific and limited datasets with given market conditions. Building recommitment strategies in various market conditions is a challenging task. In this work, we investigate Reinforcement Learning to discover promising recommitment policies using the policy-based PPO algorithm. Policy-based algorithms [13, 15] have been motivated by the fact that solving a RL problem is all about finding a sequences of actions even for value-based algorithms [10, 6]. Discovering and predicting the best actions avoid the computational burden to compute all state values. Besides, when the action space is continuous or very large, policy-based approaches are more attractive than values as we do not need to solve an optimisation problem to select the best action.

3 Problem description

This section describes the Private Equity Recommitment (PERP) by considering a single LP investor owning a 100% private equity portfolio. To minimize the opportunity cost, the investor's primary target is to remain fully invested while avoiding cash shortage. Let us define $\mathcal{P}(t) = \{f\}_{i=1}^{M}$ the set of active funds in the portfolio at time t. In order to measure its degree of investment, the fraction of total allocated capital that is actually invested can be computed as follows:

$$ID(\mathcal{P},t) = \frac{\sum\limits_{f\in\mathcal{P}(t)} NAV(f,t)}{\sum\limits_{f\in\mathcal{P}(t)} NAV(f,t) + Cash(\mathcal{P},t)} \tag{1}$$

where $\sum\limits_{f\in\mathcal{P}(t)} NAV(f,t)$ represents the sum of all Net Asset Value (NAV) for the underlying funds in the portfolio at period t. $Cash(\mathcal{P},t)$ accounts for the global uninvested cash in the portfolio, i.e., uncalled capital and possible distributions. Ideally, the investment degree ID should be as close as possible to 1. A trivial but not viable solution would be to bring $Cash(\mathcal{P},t)$ to 0 but this is without counting on future and inopportune capital calls exceeding the investor resources capacities. Becoming a defaulting investor once capital has been committed is subject to strong financial and reputational penalties. The PERP is therefore a challenging problematic for LP investors as they constantly need to stay close to the boundary without over-crossing it. In [18], authors modelled the problem as a sequence of single-period portfolio optimisation problems maximizing subsequent investment degrees using the following formulation:

$$\min_{C(\mathcal{P},t)} E_t\left[(1 - ID(\mathcal{P},t+1))^2\right] \tag{2}$$

where the $C(\mathcal{P},t)$ represents the optimal amount of capital to be recommitted at t. Note that this model only determines the optimal recommitment level with regards to the next period. This is debatable as the committed capital is called progressively over the investment period, i.e., roughly during the first 6 years. With respect to formulation (2), the optimal level of commitment at period t is therefore:

$$C(\mathcal{P},t) = E_t\left(\frac{Cash(\mathcal{P},t) + D(\mathcal{P},t+1) - \sum_{i=1}^{\tau}\gamma_{t+1,i+1}C_(\mathcal{P},t-i)}{\gamma_{t+1,1}}\right) \tag{3}$$

with E_t the conditional expectation, $Cash(\mathcal{P},t)$ the uninvested cash in the portfolio, $D(\mathcal{P},t)$ representing distributions for the next period, $C(\mathcal{P},t-i)$ the capital committed i period ago and $\gamma_{t+1,i+1}$ is the fraction of the capital committed i periods ago. $\gamma_{t+1,i+1}$ enables to compute the total capital called at the end of quarter $t+1$, i.e.,
$CC(\mathcal{P},t-i) = \sum\limits_{i=0}^{\tau}\gamma_{t+1,i+1}C(\mathcal{P},t-i)$ with τ representing the maximum fund age at which capital can still be called. Interested readers can refer to [18] for more details about the proof.

One can observe that the analytical solution requires to forecast distributions (see [16, 8]) at $t+1$ and the fraction of the capital committed in the past that will be called. Although prediction models can be developed to approximate future distributions, it is very unlikely to *guess* future capital calls as direct investments in private companies are made at the discretion of the fund's management.

Some works [18, 12] in the literature have tried to cope with this issue by engineering strategies using only available and past quantities. These strategies can be likened "heuristics" to approximate the optimal amount to be recommitted at each period and are defined as follows:

- $DZ^1(\mathcal{P}, t) = D(\mathcal{P}, t)$;
- $DZ^2(\mathcal{P}, t) = D(\mathcal{P}, t) + UC(\mathcal{P}, t - 24)$;
- $DZ^3(\mathcal{P}, t) = \frac{1}{ID(\mathcal{P},t)} \times (D(\mathcal{P}, t) + UC(\mathcal{P}, t - 24))$

Strategy $DZ^1(\mathcal{P}, t)$ recommits only current distributions at t while the strategy $DZ^2(\mathcal{P}, t)$ incorporates the uncalled capital made 24 quarters ago, i.e., $UC(\mathcal{P}, t - 24)$. The inclusion of this quantity is based on the observation that unallocated but committed capital for older funds that already passed their maximal NAV's peak is unlikely to be called. These funds are typically in the divestment period. The last strategy $DZ^3(\mathcal{P}, t)$ scales recommitments obtained from $DZ^2(\mathcal{P}, t)$ with the inverse of the current investment degree. If the investment degree is high, the recommitted capital will be decreased. Conversely, a low investment degree will amplify the recommitted capital. This allows to perform some kind of active control to adjust the level of recommitment to reach and remain stable at a target allocation .

In this paper, we propose to learn an active control system to recommit at each period. Instead of relying on cashflow predictions and strategies' engineering which require strong expert knowledge, we posit that recommitment policies could be learnt using a policy-based algorithm introduced in the next section.

4 Proximal Policy Optimisation

As aforementioned in section 2, the number of approaches relying on policy learning has flourished since recent years. They all try to find a trade-off between fast training and stability. Making large steps in the policy update can be disastrous, especially for on-policy algorithms which could never recover from subsequent updates. Among all existing alternatives in the literature, we considered the Proximal Policy Optimisation (PPO) algorithm [15] due to its simplicity. Although the PPO algorithm was released long after the Trust Region Policy Optimisation (TRPO) [13] which was the first of its kind, the PPO policy update is simpler but empirically seems to perform at least as well as TRPO relying on a second-order approach. But before diving into the stability improvement proposed in the PPO algorithm, let us recall the foundations, i.e., the vanilla policy gradient. Let π_θ represents a policy as a function of the parameter θ, the current state s_t, the taken action a_t and the received reward r_t at time t. A trajectory τ is a sequence of states and actions representing the path taken by an agent. In Reinforcement Learning, the goal is to discover the trajectory maximizing the expected return $J(\theta) = \mathbf{E}_{\pi_\theta}[R(\tau)]$ by updating sequentially the weights θ as follows: $\theta_{k+1} = \theta_k + \alpha * \nabla_\theta J(\theta_k)$ where $\nabla_\theta J(\theta_k)$ represents the policy gradient and is expressed as $\nabla_\theta J(\theta) = \mathbf{E}[R(\tau)\nabla_\theta \log \pi_\theta(a_t|s_t)]$. $R(\tau)$ can take different forms as suggested in [14]:

- the total reward trajectory: $\sum_{t=0}^{\infty} r_t$
- the future reward from action a_t or rewards-to-go: $\sum_{t=t'}^{\infty} r'_t$
- Future reward with baseline: $\sum_{t=t'}^{\infty} r'_t - b(s_t)$
- State-action value function: $Q^{\pi_\theta}(s_t, a_t)$
- Advantage function: $A^{\pi_\theta}(s_t, a_t) = Q^{\pi_\theta}(s_t, a_t) - V^{\pi}(s_t)$

All the previous choices lead to the same expected value but have different variance. The formulation using the advantage function is extremely common as it uses the state-action value function and the estimation value of the state as baseline to reduce the variance of the gradient. The PPO algorithm relies on an estimation of the advantage function and tries to avoid parameter updates that change the policy too much at one step. In the same way as TRPO, the loss function is built to measure of how policy π_θ performs relatively to an old policy $\pi_{\theta_{old}}$:

$$\mathcal{L}(\theta, \theta_{old}) = \mathbf{E}\left[A^{\pi_\theta}(s_t, a_t)\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\right] \tag{4}$$

While the TRPO algorithm uses the hard constraint $D_{KL}(\theta||\theta_{old}) < \lambda$ to limit the KL-divergence between both policies, the PPO algorithm relaxes the hard constraints and:

- either penalizes the KL-divergence directly in the loss function. This is the PPO-penalty version which we did not consider in this work.
- or clips the ratio $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ in the loss function to remove incentives for the new policy to get far from the old policy. Note that the KL-divergence is not used anymore as constraints nor as a penalty.

The PPO-clip algorithm considered in this work is depicted in Algorithm 1. Contrary to the penalty version in which penalty coefficients are adjusted automatically during training, PPO-clip requires a static hyper-parameter ϵ use to clip the ratio between the policies. Due to space restriction, we will not go further into details but more explanations can be obtained from the original paper [15].

5 Private Equity Recommitment as RL problem

As described in Section 3, the PERP can be solved using two main methodologies. While the first one relies on cashflow forecasting, the second one engineers recommitment functions only using past and current quantities from portfolios. Instead of building explicitly these functions, one could consider a Markov Decision Processes (MDP) to model a recommitment system and searches for the best policy in order to maintain a target investment degree while minimizing the risk of default.

Algorithm 1 PPO-clip version

Initialize policy parameters θ_1 and value function parameters ϕ_1
for $k \in \{1, ..., M\}$ **do**
  Sample a set of trajectories $\{\tau_i\}_{i=1}^{M}$ using the policy π_{θ_k}
  Create a batch $\mathcal{B}$ of transitions (s_t^i, a_t^i, r_t^i) $\forall t \in \{1, ..., |\tau_i|\}$ $\forall i \in \{1, ..., M\}$
  Compute rewards-to-go $\hat{\mathcal{R}}_t^i$, i.e. rewards from action a_t^i, $\forall t \in \{1, ..., |\tau_i|\}$ $\forall i \in \{1, ..., M\}$
  Estimate the advantages $A^{\pi_{\theta_k}}(s_t^i, a_t^i)$ using the value function V_{ϕ_k}
  Perform policy update:

$$\theta_{k+1} = \arg\max_{\theta} \frac{1}{M} \sum_{i=1}^{M} \frac{1}{|\tau_i|} \sum_{t=1}^{T_i} \left[\min \left(A^{\pi_\theta}(s_t^i, a_t^i) \frac{\pi_\theta(a_t^i|s_t^i)}{\pi_{\theta_{old}}(a_t^i|s_t^i)}, g\left(\epsilon, A^{\pi_\theta}(s_t^i, a_t^i)\right) \right) \right]$$

$$\text{with } g\left(\epsilon, A^{\pi_\theta}(s_t^i, a_t^i)\right) = \text{clip}\left(\frac{\pi_\theta(a_t^i|s_t^i)}{\pi_{\theta_{old}}(a_t^i|s_t^i)}, 1-\epsilon, 1+\epsilon \right)$$

  Perform value function update by minimizing mean-squared error:

$$\phi_{k+1} = \arg\min_{\phi} \frac{1}{M} \sum_{i=1}^{M} \frac{1}{|\tau_i|} \sum_{t=1}^{T_i} \left[V_\phi(s_t^i) - \hat{\mathcal{R}}_t^i \right]^2$$

end for

5.1 Modelling

Fig. 1 illustrates how the PERP can be turned into a Reinforcement Learning problem. Each state s_t represents the portfolio position at time t and contains the following information:

- $ID(\mathcal{P}, t)$: Investment degree at time t
- $D(\mathcal{P}, t)$: Distributions obtained from divestments at time t
- $CC(\mathcal{P}, t)$: Capital called at time t
- $UC(\mathcal{P}, t-24)$: Uncalled capital from commitment made 24 quarters ago
- $Cash(\mathcal{P}, t)$: Portfolio cash at time t
- $NAV(\mathcal{P}, t)$: Net Asset Value at time t

The state s_t gives us the opportunity to control the amount of recommitted capital at time t, i.e., the continuous action a_t depicted in Fig. 1. So far, the RL model is trivial to obtain. However, we need to be extremely cautious regarding the reward provided to the agent. Although we could define the reward by minimizing the deviation to the ideal investment degree as done in Equation 2, there is no control on the risk of default. Two alternatives open to us: (1) either we train on multiple portfolios per episode and adjust the objective using the standard deviation or (2) we constrain the agent to remain below the fateful boundary, i.e., $ID(\mathcal{P}, t) = 1.0$. Needless to say, alternative (2) is more challenging for the agent but we argue that it will be more generalizable than alternative (1). For this purpose, we define a local reward r_t^{valid} and a global reward r_τ^{ID}. While the former is applied after each action(recommitment), the second one only occurs at the end of a valid episode. We recall that a valid episode ends when the maximum number of steps has been reached. The agent is rewarded after each action depending on whether the future state of the portfolio is valid:

$$r_t^{valid} = \begin{cases} 0 \text{ if } ID(\mathcal{P}, t+1) > 1 \\ 1 \text{ if } else \end{cases}$$

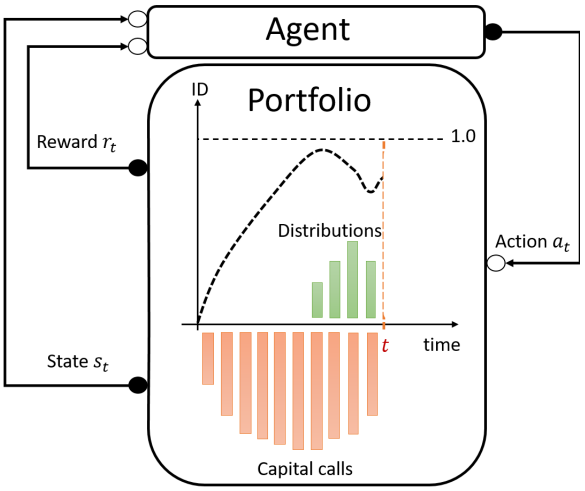

Fig. 1: Reinforcement Learning of private equity policies

If a situation of default happens, the episode is stopped and does not reach the maximum number of steps allowed. The accumulated reward obtained during the episode would finally correspond to the number of periods in which the portfolio remained valid. This reward function strictly increases monotonically to drive the agent to simply learn to provide valid episodes. Once the agent has learnt to recommit, i.e., it reaches the maximum number of steps per episode, it receives an additional and final global reward $r_\tau^{ID} = \sum_{t=1}^{T} ID(\mathcal{P}, t)$ where T is the maximal number of steps per episode. Note that the sum could be replaced by the min to maximize the worst investment degree obtained during an episode. Finally, the total reward of a valid episode is the accumulated local reward added to the shifted global reward:

$$r_\tau = r_\tau^{ID} \times 10^{(digits(T)+1)} + \sum_{t=1}^{T} r_t^{valid} \tag{5}$$

where $digits(T)$ is the number of digit of T. For an episode lasting 100 steps, $\#digit(100) = 2$. This shifting mechanism is a constraint handling approach to make sure that non-valid episodes are guaranteed to receive a total reward lower than valid ones.

5.2 Synthetic cashflows

Private equity data is a sensitive topic. Private equity players generally protect their rich cashflow histories. Although some financial data providers propose

commercial libraries for very specific periods and economies, their data are generally incomplete. Historical cashflows's data capture the fund's dynamics which is an essential information for training. Multiple works including [18] and [12] relied on commercial libraries to draw conclusions or train their own model. In this work, we adopt another strategy to simulate portfolio evolution over time. Since 1973, the Yale University's endowment has been investing in private equity using a methodology for modelling illiquid assets proposed by Takahashi and Alexander (see [16]). Referred to as the *mother of all cashflows's models*, this Yale-model can be applied to private equity and real asset funds (e.g. natural resources and infrastructures). Although, according to Takahasi and Alexander, the generated projections fit historical data, the cashflows are modelled as deterministic which limit their applicability.

Instead of depending on a commercial solution to acquire historical cashflows which are often expensive and incomplete, synthetic fund cashflows have been preferred in this work as they represent a more practical solution. This is the reason why we decide to rely on an alteration of the Yale-model to make it probabilistic. These synthetic cashflows are created by funnelling data generated by the robust and tried-and-tested, albeit over-simplistic, Yale-model through a noise-adding algorithm to construct a new dataset. The resulting dataset shows the statistical features and the useful patterns needed for capturing the liquidity risks associated with portfolio of funds. The synthetic cashflows considered in this work have been provided by T.Meyer, an expert in private equity and co-author of this paper.

6 Experimental setups

In order to fairly evaluate the resulting recommitment policies with the state of the art, simulations have been performed according to the parameters described in [18]. Due to the lack of secondary market, a portfolio cannot be bought instantaneously. We empirically created initial but mature portfolios over a year by committing equal capital to 16 randomly selected private equity funds. We also apply 30 % initial overcommitment in setting up all portfolios to be in line with the experiments performed in [18].

A portfolio simulation consists in recommitting some capital to new selected fund every quarter. The amount of capital is determined by the current policy sampled from the critic network (see Algorithm 1). Table 1a details the simulation parameters while Table 1b described the PPO-clip parameters. A single portfolio simulation last 104 quarters, i.e., 26 years. Capital is recommitted uniformly into 4 randomly selected funds. The number of portfolio simulations is therefore equal to the number of episodes:

$$\#episodes = \frac{steps\_per\_epoch \times epochs}{104} = 125000$$

Strategies $DZ^i(\mathcal{P}, t)$ for $i \in \{1, 2, 3\}$ proposed in [18] have been evaluated with the same parameters and over the same period. All experiments presented

in this paper were carried out using the HPC facility of the University of Luxembourg [17]. The python library SpinningUp [1] has been considered for the PPO-clip implementation. A distributed implementation using OpenMPI [5] has been considered to work with multiple environment in parallel. The discount parameter γ has been set to 1.0 since an episode's length is finite and last 26 years. The clip ratio ϵ has been set to 0.2 and represents how far can the new policy go from the old policy while still improving the objective. PPO-clip 's networks, i.e., actor and critic have both two hidden layers of 64 nodes. The ReLU function [2] has been chosen as activation function.

Table 1: Parameters

Parameters	**Training**	**Validation**
Cashflows frequency	quarterly	quarterly
Investment period	26 years	26 years
Funds per recommitment	4	4
Fund selection	random	random
Number of simulated portfolios	#episodes	1000

(a) Simulation parameters

Parameters	**Value**
steps_per_epoch	26000
gamma	1
epochs	500
# episodes	125000
clip_ratio ϵ	0.2
pi_lr / vf_lr	$3e^{-4}$ / $1e^{-4}$
hidden layers	$[64, 64]$

(b) PPO-clip parameters

7 Experimental results

With regards to the experimental setups described in the previous section, Fig. 2 illustrates the average rewards recorded during policy optimisation/training. One can easily observe that the PPO-clip algorithm required few epochs to generate valid policies. The average rewards curve then steadily increases to reach what we can consider as a plateau in terms of improvements. Indeed, we can note periodic falls indicating that the algorithm have strong difficulties to improve more significantly the investment degree without breaking the cash constraint. When arrived at the rupture point, a policy yielding non-valid episodes is more likely to be generated leading to a steep fall in terms of overall rewards. When a fall occurs, the algorithm tries to recover until the next rupture. This pattern can be easily oberved in Fig. 2. Due to the shifting constraint handling approach implemented in this work, non-valid and valid episodes do not have the same reward scale which explains these deep reward falls every time the algorithm encounters a non-valid episode.

The best policy obtained after training is depicted in Fig. 2. In order to validate results, the obtained policy has been applied on a test set of 1000 portfolios. After recording the investment degree evolution and the validity of each portfolio, the average investment degree as well as the surrounding 95% confidence

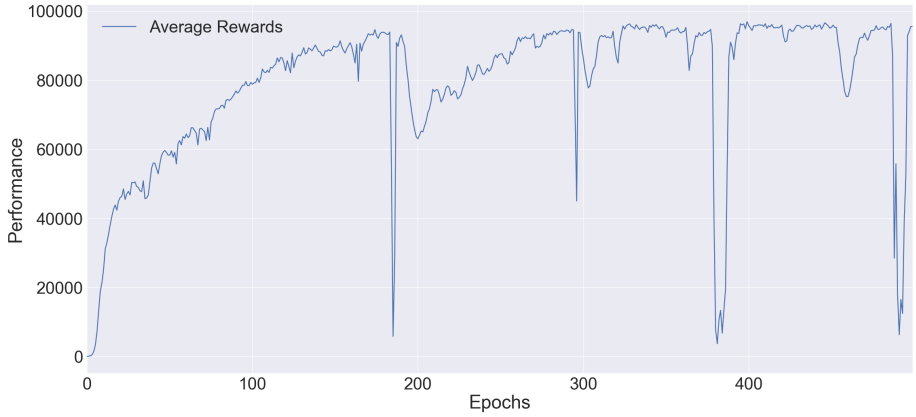

Fig. 2: Evolution of the average rewards per epoch

interval have been computed and are depicted in Fig. 3. We first observe that the percentage of overinvested portfolios remains extremely low, i.e. $\approx 0.7\%$. The investment degree varies strongly during the first 6 years going from 0.4 to almost 1.0. After the first 6 years, the average investment degree slightly increases to remain stable around 0.9.

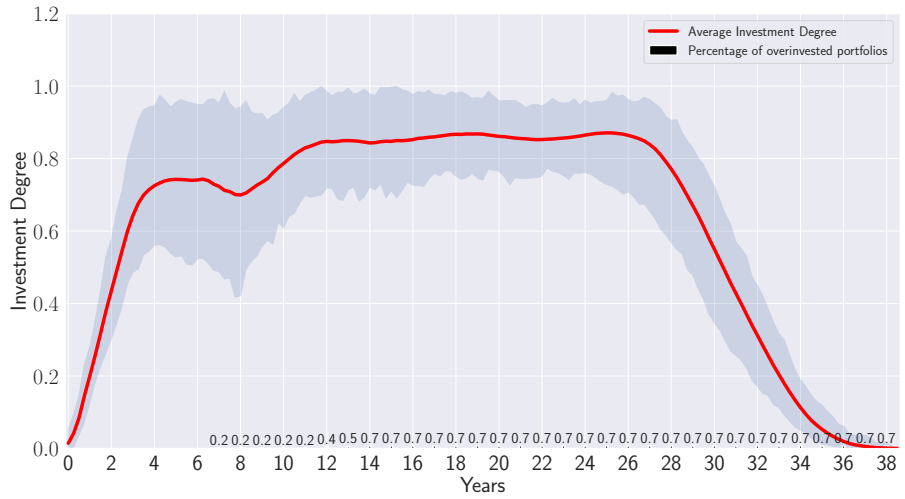

Fig. 3: Best policy obtained with the PPO-clip algorithm

We now compare the investment degree obtained with state-of-the-art strategies engineered in [18], namely DZ^i for all $i \in \{1, 2, 3\}$. Each DZ^i have been

applied on the same test set. Table 2 reports the average investment degree, the standard deviation of the investment degree and the fraction of overinvested portfolios obtained for each strategy including the best policy recommitment $PPO - clip^{best}$. Although the active recommitment period only lasts 26 years, we have still recorded the investment degree until portfolios were totally divested (38 years) to observe if there is no delay effect when applying a specific strategy. None of the 3 strategies have generated invalid portfolios. The investment degree reached by DZ^1 and DZ^2 remains low, i.e., below 0.6. Nevertheless, DZ^3 obtained the best results among the 3 strategies as reported in [18]. The recommitment policy $PPO - clip^{best}$ outperforms the 3 strategies by reaching a maximum investment degree above 0.8. Nonetheless, the DZ^3 reports better results during the first years as show in Fig. 4. The initial condition of the portfolio seems to be a challenge for the recommitment policy. Nevertheless, it is well-known in the literature that portfolio inception is a problem on its own. Therefore, we are not surprised by this under-performance at the beginning of the portfolio lifetime. In [18], authors discarded the first three years of the portfolio's lifetime to avoid the influence from the initial portfolio formation period.

Regarding the percentage of overinvested portfolios, it comes as no surprise to encounter some invalid portfolios when getting closer to $ID(\mathcal{P}, t) = 1.0$. This is due to cashflow variability which is very difficult to predict. An alternative would be to replace the strong cash constraint by a soft one taking the form of an additional objective. Most of the LP investors generally own multi-class asset portfolios. If liquidity is missing due to an unexpected capital calls, more liquid assets could be sold. Of course, such a situation should be tempered and the injected cash required to satisfy capital calls should be minimized. For this purpose, one could consider a multi-objective reinforcement learning algorithm.

8 Conclusion

Recommitting into new PE funds is crucial for LP investors to maintain high allocation to private equity. Current methodologies rely on cashflow forecasting and over-simplistic approaches which are lacking of flexibility. Although this problem is a key of major importance, few works have attempted to develop a robust and flexible recommitment system. Perhaps, this is due to the lack of data. This is the reason why we adopted a different strategy consisting in learning recommitment policies through Reinforcement Learning. Using synthetic cashflows build from the traditionnal but proven Yale-model, we applied Proximal Policy Optimisation to the Private Equity Recommitment Problem to maximise the investment degree while avoiding cash shortage situations by constraining the agent. Results obtained after training confirm that the recommitment policy outperform the strategies engineered in [18] while limiting the fractions of invalid portfolios. This work was a first proof of concept and subsequent experiments will be performed using different RL algorithms. Future works will investigate a strategy to handle the cash constraint more efficiently. Another avenue for research would be to model the cash constraint as a soft constraint, typically by

	$PPO - clip^{best}$			DZ^1			DZ^2			DZ^3		
years	mean	std	invalid (%)	mean	std	invalid (%)	mean	std	invalid (%)	mean	std	invalid (%)
0	0.07	0.02	0.00	0.07	0.02	0.0	0.07	0.02	0.0	0.07	0.02	0.0
1	0.29	0.03	0.00	0.29	0.03	0.0	0.29	0.03	0.0	0.30	0.03	0.0
2	0.52	0.04	0.00	0.52	0.04	0.0	0.52	0.04	0.0	0.55	0.03	0.0
3	0.68	0.06	0.00	0.69	0.04	0.0	0.69	0.04	0.0	0.75	0.03	0.0
4	0.73	0.06	0.00	0.75	0.04	0.0	0.75	0.04	0.0	0.83	0.03	0.0
5	0.74	0.07	0.00	0.76	0.04	0.0	0.76	0.04	0.0	0.85	0.04	0.0
6	0.74	0.07	0.08	0.71	0.05	0.0	0.71	0.05	0.0	0.81	0.05	0.0
7	0.71	0.08	0.20	0.63	0.05	0.0	0.63	0.05	0.0	0.74	0.05	0.0
8	0.71	0.07	0.20	0.56	0.04	0.0	0.57	0.05	0.0	0.70	0.04	0.0
9	0.75	0.05	0.20	0.54	0.03	0.0	0.56	0.03	0.0	0.72	0.04	0.0
10	0.80	0.05	0.20	0.56	0.03	0.0	0.58	0.03	0.0	0.76	0.03	0.0
11	0.84	0.05	0.23	0.58	0.02	0.0	0.60	0.02	0.0	0.79	0.03	0.0
12	0.85	0.05	0.40	0.59	0.02	0.0	0.62	0.02	0.0	0.81	0.03	0.0
13	0.85	0.05	0.58	0.59	0.02	0.0	0.62	0.02	0.0	0.81	0.03	0.0
14	0.84	0.06	0.70	0.58	0.02	0.0	0.60	0.02	0.0	0.79	0.03	0.0
15	0.85	0.06	0.70	0.56	0.02	0.0	0.58	0.02	0.0	0.77	0.03	0.0
16	0.85	0.06	0.70	0.55	0.02	0.0	0.57	0.02	0.0	0.76	0.03	0.0
17	0.86	0.06	0.70	0.54	0.02	0.0	0.57	0.02	0.0	0.76	0.03	0.0
18	0.86	0.07	0.70	0.55	0.02	0.0	0.58	0.02	0.0	0.77	0.02	0.0
19	0.86	0.07	0.70	0.55	0.02	0.0	0.58	0.02	0.0	0.78	0.02	0.0
20	0.85	0.07	0.70	0.55	0.02	0.0	0.58	0.02	0.0	0.79	0.02	0.0
21	0.85	0.08	0.70	0.55	0.02	0.0	0.58	0.02	0.0	0.78	0.02	0.0
22	0.85	0.08	0.70	0.54	0.02	0.0	0.58	0.02	0.0	0.78	0.02	0.0
23	0.85	0.08	0.70	0.54	0.02	0.0	0.57	0.02	0.0	0.77	0.02	0.0
24	0.86	0.08	0.70	0.54	0.02	0.0	0.57	0.02	0.0	0.77	0.02	0.0
25	0.86	0.08	0.70	0.54	0.02	0.0	0.57	0.02	0.0	0.77	0.02	0.0
26	0.85	0.08	0.70	0.54	0.02	0.0	0.57	0.02	0.0	0.78	0.02	0.0
27	0.81	0.09	0.70	0.53	0.02	0.0	0.56	0.02	0.0	0.76	0.02	0.0
28	0.73	0.08	0.70	0.49	0.02	0.0	0.52	0.02	0.0	0.71	0.03	0.0
29	0.62	0.08	0.70	0.44	0.02	0.0	0.46	0.02	0.0	0.62	0.03	0.0
30	0.50	0.07	0.70	0.37	0.02	0.0	0.39	0.02	0.0	0.51	0.03	0.0
31	0.38	0.06	0.70	0.29	0.02	0.0	0.31	0.02	0.0	0.40	0.03	0.0
32	0.27	0.05	0.70	0.21	0.02	0.0	0.22	0.02	0.0	0.29	0.03	0.0
33	0.17	0.04	0.70	0.14	0.02	0.0	0.14	0.02	0.0	0.19	0.03	0.0
34	0.09	0.02	0.70	0.07	0.01	0.0	0.08	0.01	0.0	0.10	0.02	0.0
35	0.04	0.01	0.70	0.03	0.01	0.0	0.03	0.01	0.0	0.05	0.01	0.0
36	0.01	0.01	0.70	0.01	0.01	0.0	0.01	0.01	0.0	0.02	0.01	0.0
37	0.00	0.00	0.70	0.00	0.00	0.0	0.00	0.00	0.0	0.00	0.00	0.0
38	0.00	0.00	0.70	0.00	0.00	0.0	0.00	0.00	0.0	0.00	0.00	0.0

Table 2: Summary statistics of the investment degree in recommitment strategies

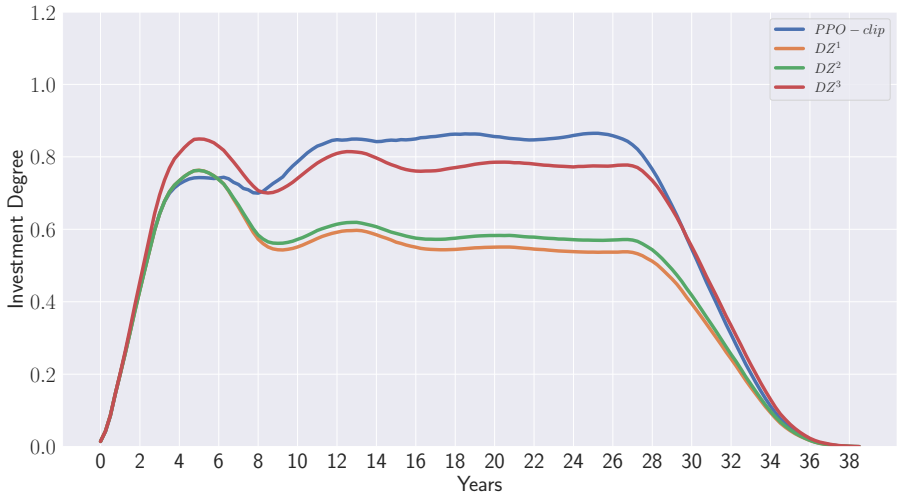

Fig. 4: Comparison between de Zwart's strategies [18] and the policy obtained with the PPO-clip algorithm

considering it as a second objective. Both opportunity cost and cash shortage are two conflicting objectives. Finally, this work could be extended to take into account multi-class asset portfolios.

Acknowledgment

E. Kieffer acknowledges the support of the European Investment Bank through its STAREBEI programme.

References

1. Achiam, J.: Spinning Up in Deep Reinforcement Learning (2018)
2. Agarap, A.F.: Deep learning using rectified linear units (relu) (2018), http://arxiv.org/abs/1803.08375, cite arxiv:1803.08375Comment: 7 pages, 11 figures, 9 tables
3. Arnold, T.R., Ling, D.C., Naranjo, A.: Waiting to be called: The impact of manager discretion and dry powder on private equity real estate returns. The Journal of Portfolio Management **43**(6), 23–43 (2017)
4. Cardie, J.H., Cattanach, K.A., Kelly, M.F.: How large should your commitment to private equity really be? The Journal of Wealth Management **3**(2), 39–45 (2000)
5. Gabriel, E., Fagg, G.E., Bosilca, G., Angskun, T., Dongarra, J.J., Squyres, J.M., Sahay, V., Kambadur, P., Barrett, B., Lumsdaine, A., Castain, R.H., Daniel, D.J., Graham, R.L., Woodall, T.S.: Open MPI: Goals, concept, and design of a next generation MPI implementation. In: Proceedings, 11th European PVM/MPI Users' Group Meeting. pp. 97–104. Budapest, Hungary (September 2004)

6. Hasselt, H.v., Guez, A., Silver, D.: Deep reinforcement learning with double q-learning. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. p. 2094–2100. AAAI'16, AAAI Press (2016)
7. Lerner, J., Schoar, A.: The illiquidity puzzle: theory and evidence from private equity. Journal of Financial Economics **72**(1), 3–40 (2004). https://doi.org/https://doi.org/10.1016/S0304-405X(03)00203-4, https://www.sciencedirect.com/science/article/pii/S0304405X03002034
8. de Malherbe, E.: Modeling private equity funds and private equity collateralised fund obligations. International Journal of Theoretical and Applied Finance **07**, 193–230 (2004)
9. Meyer, T.: Hidden in plain sight—the impact of undrawn commitments. The Journal of Alternative Investments **23**(2), 94–110 (2020)
10. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning (2013), http://arxiv.org/abs/1312.5602, cite arxiv:1312.5602Comment: NIPS Deep Learning Workshop 2013
11. Nevins, D., Conner, A., McIntire, G.: A portfolio management approach to determining private equity commitments. The Journal of Alternative Investments **6**(4), 32–46 (2004)
12. Oberli, A.: Private equity asset allocation: How to recommit? The Journal of Private Equity **18**(2), 9–22 (2015)
13. Schulman, J., Levine, S., Abbeel, P., Jordan, M.I., Moritz, P.: Trust region policy optimization. In: Bach, F.R., Blei, D.M. (eds.) ICML. JMLR Workshop and Conference Proceedings, vol. 37, pp. 1889–1897. JMLR.org (2015), http://dblp.uni-trier.de/db/conf/icml/icml2015.html#SchulmanLAJM15
14. Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P.: High-dimensional continuous control using generalized advantage estimation. In: Proceedings of the International Conference on Learning Representations (ICLR) (2016)
15. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. CoRR **abs/1707.06347** (2017), http://dblp.uni-trier.de/db/journals/corr/corr1707.html#SchulmanWDRK17
16. Takahashi, D., Alexander, S.: Illiquid alternative asset fund modeling. The Journal of Portfolio Management **28**(2), 90–100 (2002). https://doi.org/10.3905/jpm.2002.319836, https://jpm.pm-research.com/content/28/2/90
17. Varrette, S., Bouvry, P., Cartiaux, H., Georgatos, F.: Management of an academic hpc cluster: The ul experience. In: Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014). pp. 959–967. IEEE, Bologna, Italy (July 2014)
18. de Zwart, G., Frieser, B., van Dijk, D.: Private equity recommitment strategies for institutional investors. Financial Analysts Journal **68**(3), 81–99 (2012)