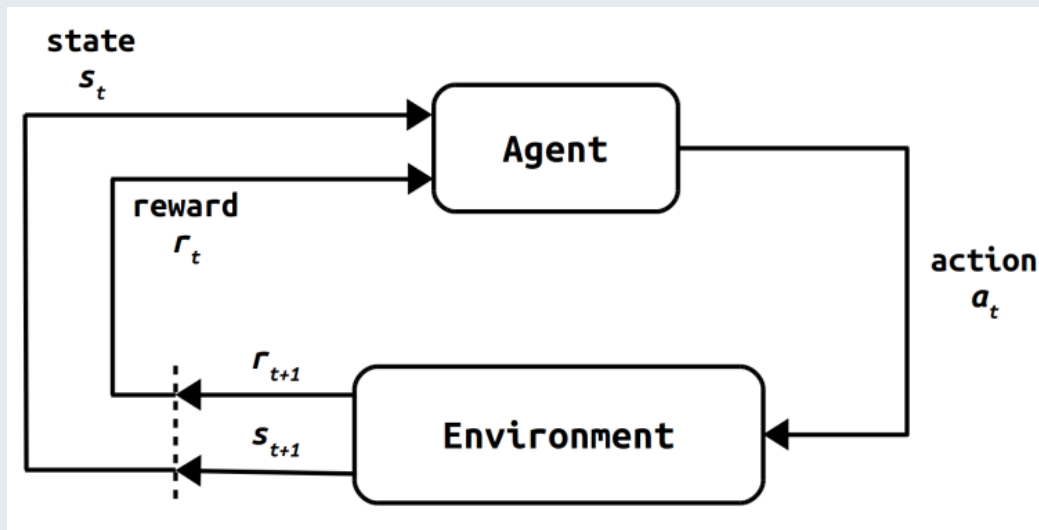


Reinforcement Learning for Portfolio Management



强化学习

- 环境
- 智能体



R.S. Sutton, A.G. Barto,
Reinforcement Learning: An
Introduction,
MIT Press, Cambridge, MA, 1998

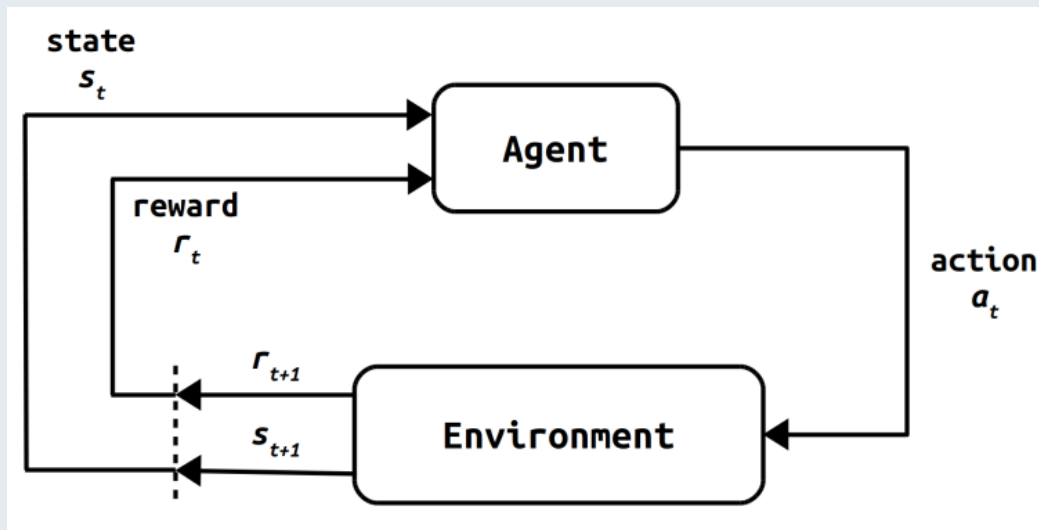


状态空间

- 50×11
- 50: 可投资的股票数目
- 11: 股票的维度: ['zopen', 'zhhigh', 'zlow', 'zadjcp', 'zclose', 'zd_5', 'zd_10', 'zd_15', 'zd_20', 'zd_25', 'zd_30']

奖励

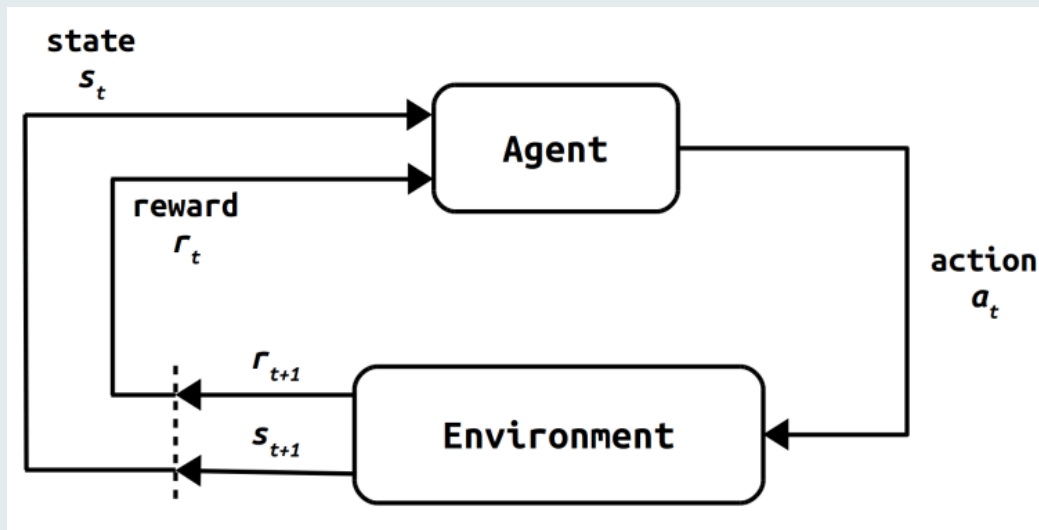
- reward = portfolio_value(t+1) - portfolio_value(t)



R.S. Sutton, A.G. Barto,
Reinforcement Learning: An
Introduction,
MIT Press, Cambridge, MA, 1998

行动空间

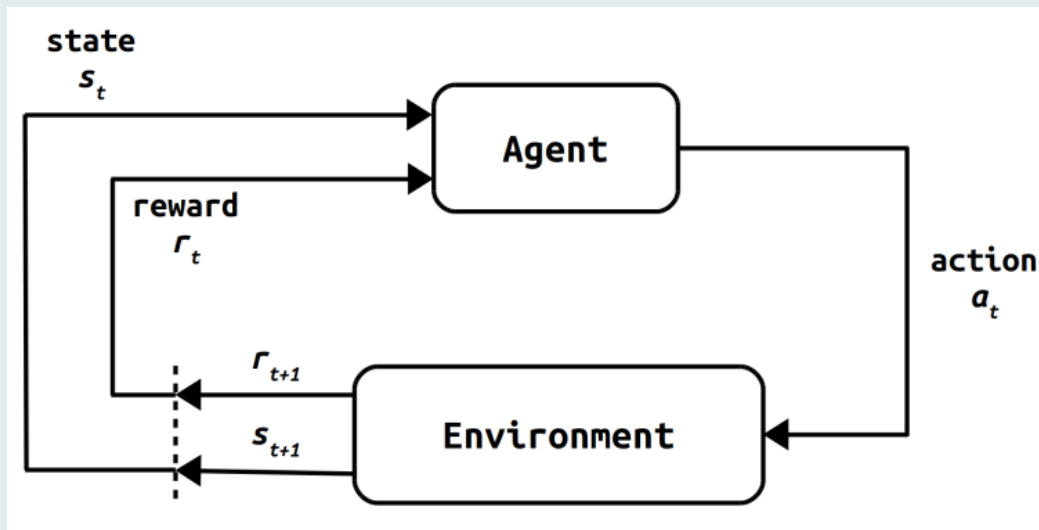
- (51,)
- 50: 可投资的股票数目+1:不投资



R.S. Sutton, A.G. Barto,
Reinforcement Learning: An
Introduction,
MIT Press, Cambridge, MA, 1998

智能体

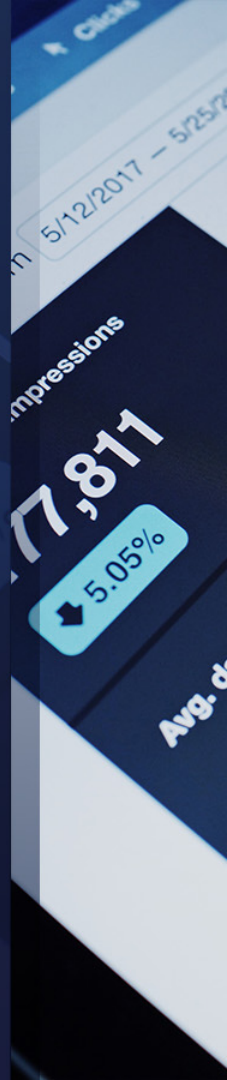
- 状态空间(50,11) ->行动空间 (51,)
- 用反馈的奖励来训练智能体



R.S. Sutton, A.G. Barto,
Reinforcement Learning: An
Introduction,
MIT Press, Cambridge, MA, 1998

智能体设计

- Policy Optimization: 策略优化
 - Policy Gradient: 策略梯度
- State Optimization: 状态优化
 - Q-learning: Q学习
- Actor-Critic Methods: 演员-评论家方法



随机策略揭示问题

```
def random_agent(env):  
    return env.action_space.sample()
```

```
def one_agent(env):  
    return np.ones_like(env.action_space_shape)
```

- Fully random vs Evenly distributed



portfolio_value -93.676294%



portfolio_value +323.096709%

Policy Gradient: 策略梯度

- 策略网络结构
 - input 状态空间
 - out 行动空间
- 训练策略网络
 - Step1. MC 采样
 - Step2. batch 训练

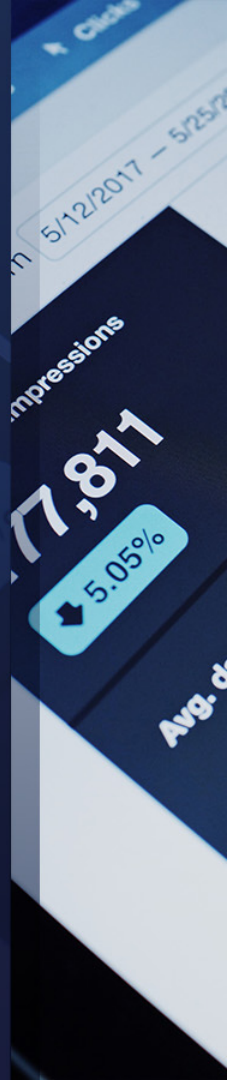
Algorithm 5: Model-Carlo Policy Gradient (REINFORCE).

inputs : trading universe of M -assets
initial portfolio vector $w_1 = a_0$
initial asset prices $p_0 = o_0$
objective function \mathcal{J}
initial agent weights θ_0
output: optimal agent policy parameters θ_*

```
1 initialize buffers:  $G, \Delta\theta_c \leftarrow 0$  repeat
2   for  $t = 1, 2, \dots T$  do
3     observe tuple  $\langle o_t, r_t \rangle$ 
4     sample and take action:  $a_t \sim \pi_\theta(\cdot | s_t; \theta)$  // portfolio
      rebalance
5     cache rewards:  $G \leftarrow G + r_t$  // (6.19)
6     cache log gradients:  $\Delta\theta_c \leftarrow \Delta\theta_c + \nabla_\theta \log[\pi_\theta(s, a)] G$  // (6.20)
7   end
8   update policy parameters  $\theta$  using buffered
9     Monte-Carlo estimates via adaptive optimization // (6.18),
      ADAM
10  empty buffers:  $G, \Delta\theta_c \leftarrow 0$ 
11 until convergence
12 set  $\theta_* \leftarrow \theta$ 
```

网络设计

- 简单全连接
- Lstm



The header image is a collage of financial data. It includes a line graph with a blue line showing an upward trend, with labels '150,000' and '120,000'. There is also a bar chart with a blue bar labeled '7,811' and a percentage '25%'. The word 'Impressions' is visible in the top left. The date 'May 12' is partially visible on the right side of the chart.

简单全连接网络

The header image is a collage of financial data. On the left, there's a dark blue area with the number '7,811' in large white font, and '05%' in a light blue box below it. Above '7,811' is the word 'impressions' in white. To the right, there's a light blue area with a line graph showing an upward trend, with labels '120,000' and '150,000' on the y-axis, and 'May 12' on the x-axis.

Lstm



问题和未来方向

- 环境:
 - 数据集:
 - 考虑停牌股票
 - 考虑股票的基本面信息
- 智能体:其他的智能体结构
 - 网络结构: 可扩展性（现在只能支持50支股票）

- R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998
- Filos, A. (2019). Reinforcement Learning for Portfolio Management. ArXiv [q-Fin.PM]. Retrieved from <http://arxiv.org/abs/1909.09571>
- Sun, S., Qin, M., Wang, X., & An, B. (2023). PRUDEX-Compass: Towards Systematic Evaluation of Reinforcement Learning in Financial Markets. Transactions on Machine Learning Research. Retrieved from <https://openreview.net/forum?id=JjbslYOuNi>



谢谢！