

DMUU Assignment 2 – Solution

Part 1. Guided Reading.

You did not have to submit answers to these questions. Short answers from the Medium article:

1. *Why are nearly all RL trading systems trained in simulators before they trade real money? (Section 1)*

Exploration in a live market is expensive. Every exploratory “let’s see what happens” order has real gains or losses. In a simulator, the agent explores without losing real money. It trades real money only after its policy has converged.

Learning on historic data does not have the disadvantage of losing real money, but the amount of (recorded) data may not be sufficient for certain algorithms/products.

2. *Why does the author call finance “almost perfectly structured as an RL problem”? (Section 1)*

The author gives four reasons:

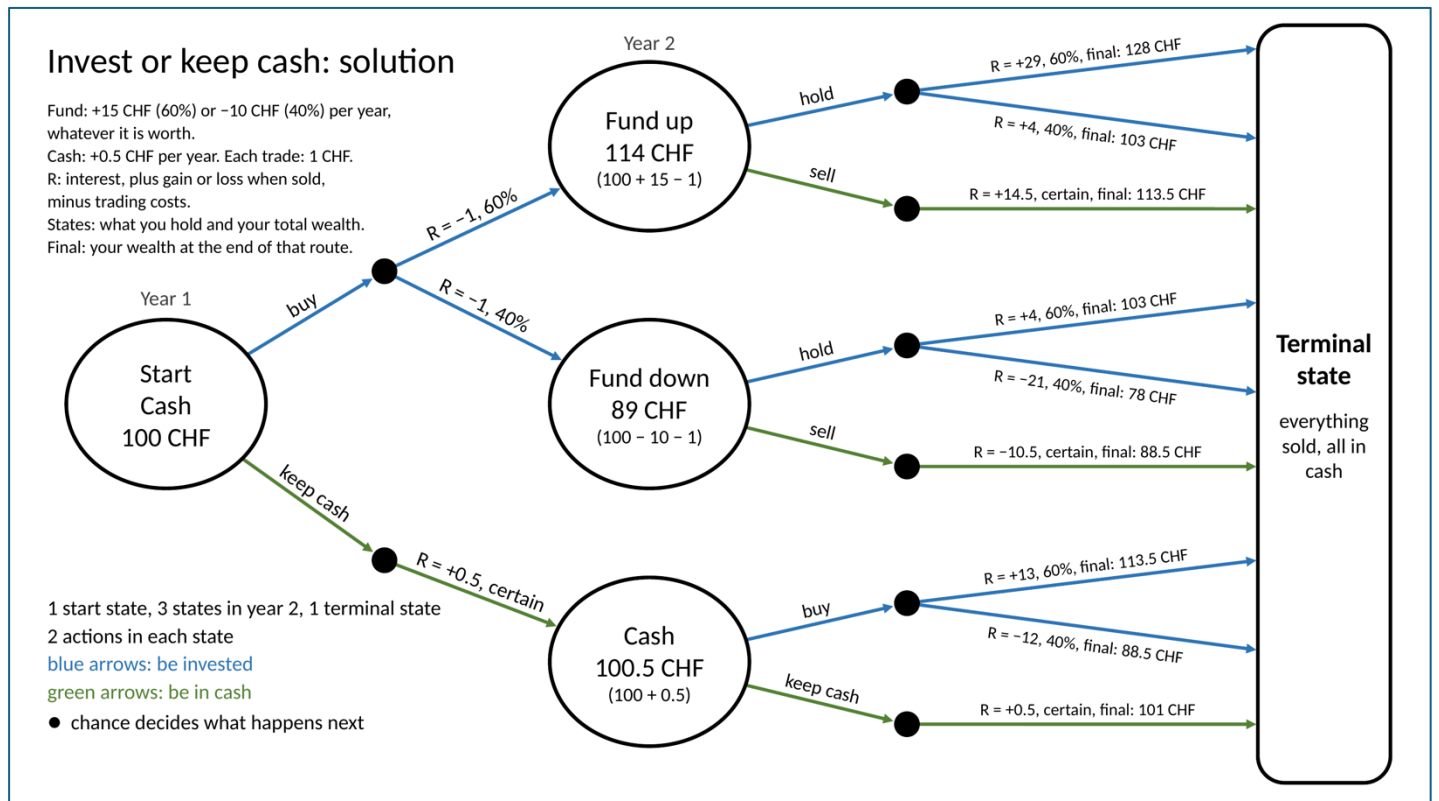
- Decisions are sequential.
- The data the agent sees depends on its own actions.
- Rewards can be measured in money.
- The market changes over time. No fixed formula stays the best strategy forever.

3. *Why do many impressive results of RL portfolio strategies disappear when tested on later periods? (Section 6e)*

The later period often contains a regime change, for example 2008, COVID or the 2022 interest rate shock. The market then behaves differently than in the training period. The policy has learned the old market, so it no longer works.

Part 2. The scenario: Invest or keep cash

Task 1. Complete the backup diagram.



How the added values are calculated

States in year 2

- Fund down: $100 - 10 - 1 = 89$ CHF
- Cash: $100 + 0.5 = 100.5$ CHF

Arrows in year 1

- buy, fund down: $R = -1$ (trading cost), 40%
- keep cash: $R = +0.5$ (interest), certain

Actions in "Fund down" (the fund is worth 90 CHF)

- hold, fund up: $R = 105 - 100 - 1 = +4$, 60%, final: 103 CHF
- hold, fund down: $R = 80 - 100 - 1 = -21$, 40%, final: 78 CHF
- sell: $R = 90 - 100 - 1 + 0.5 = -10.5$, certain, final: 88.5 CHF

Actions in "Cash" (buying and the final sale both happen in year 2)

- buy, fund up: $R = 15 - 1 - 1 = +13$, 60%, final: 113.5 CHF
- buy, fund down: $R = -10 - 1 - 1 = -12$, 40%, final: 88.5 CHF
- keep cash: $R = +0.5$, certain, final: 101 CHF

Check: final wealth = 100 + sum of R along the route. Example: keep cash, buy, fund up:
 $100 + 0.5 + 13 = 113.5$ CHF.

Task 2. Q-values and the optimal policy.

- a) Compute $Q(\text{state}=\text{"cash"}, \text{action}=\text{"buy"})$.

$$Q(\text{Cash}, \text{buy}) = 0.6 \times 13 + 0.4 \times (-12) = 7.8 - 4.8 = 3.0$$

- b) Calculate all Q-values in the diagram.

Work backwards: first year 2, then year 1.

- Year 2: the next state is the terminal state. The Q-value is the expected reward of this step.
- Year 1: the Q-value is the reward of this step plus the expected Q-value of the better action in year 2.

State	Action	Calculation	Q-value
Fund up	hold	$0.6 \times 29 + 0.4 \times 4$	19.0
Fund up	sell	+14.5, certain	14.5
Fund down	hold	$0.6 \times 4 + 0.4 \times (-21)$	-6.0
Fund down	sell	-10.5, certain	-10.5
Cash	buy	$0.6 \times 13 + 0.4 \times (-12)$	3.0
Cash	keep cash	+0.5, certain	0.5
Start	buy	$-1 + 0.6 \times 19.0 + 0.4 \times (-6.0)$	8.0
Start	keep cash	$+0.5 + 3.0$	3.5

The better action in each state is in bold.

Example: $Q(\text{Start}, \text{buy}) = -1 + 0.6 \times Q(\text{Fund up}, \text{hold}) + 0.4 \times Q(\text{Fund down}, \text{hold})$
 $= -1 + 11.4 - 2.4 = 8.0$

- c) What is the best action in the first year?

Buy. $Q(\text{Start}, \text{buy}) = 8.0$ is larger than $Q(\text{Start}, \text{keep cash}) = 3.5$.

Optimal policy: Start: buy. Fund up: hold. Fund down: hold. Cash: buy.

Task 3. Learning from experience.

- a) How many Q-values does the agent have to learn in our scenario?

8. There are 4 states with 2 actions each: Start, Fund up, Fund down, Cash. The terminal state has no actions.

- b) Roughly how many years of investing would the agent need? Are 20 years of historic market data enough?

$$8 \text{ Q-values} \times 10 \text{ tries} = 80 \text{ tries.}$$

One episode gives 2 tries: one action in year 1 and one action in year 2. The agent needs at least 40 episodes, which is 80 years.

In practice, the agent needs more. Chance decides which state the agent reaches in year 2. Example: the agent chooses buy and keep cash equally often in year 1. It then reaches Fund down in only $0.5 \times 0.4 = 20\%$ of the episodes. The 20 tries in Fund down need about 100 episodes, which is 200 years.

Answer: roughly 80 to 200 years. 20 years of data give only 10 episodes. This is not enough.



- c) Give one advantage and one risk of historical market data and of simulated markets.

Examples:

	Advantage	Risk
Historical market data	Real market behaviour, including real crises.	Little data (Task 3b). The past may not repeat (Part 1, question 3).
Simulated markets	Unlimited episodes. The agent explores without losing real money (Part 1, question 1).	The simulator may be wrong. The agent then learns the simulator, not the market.

- d) Is learning from experience easier or harder for an agent that places orders within fractions of a second? Explain.

Both answers can be argued.

- **Easier:** the agent makes thousands of decisions per day. Our agent makes 2 decisions in 2 years. The fast agent collects far more experience.
- **Harder:** the agent's own orders move the price. Other traders react to these orders. Price changes within fractions of a second contain a lot of noise.
- **Risk reduction:** In algorithmic high-frequency trading, the agents often hold a position only for a short period of time. This has the advantage that the risk of a large change (geopolitical event, pandemy outbreak, crash, ...) is extremely low. The macro-economic context can be assumed constant during a short period.
- **High volume:** Algorithms often exploit very small changes in price. In order to still make profit, they buy/sell large volumes. This makes learning not harder but riskier.