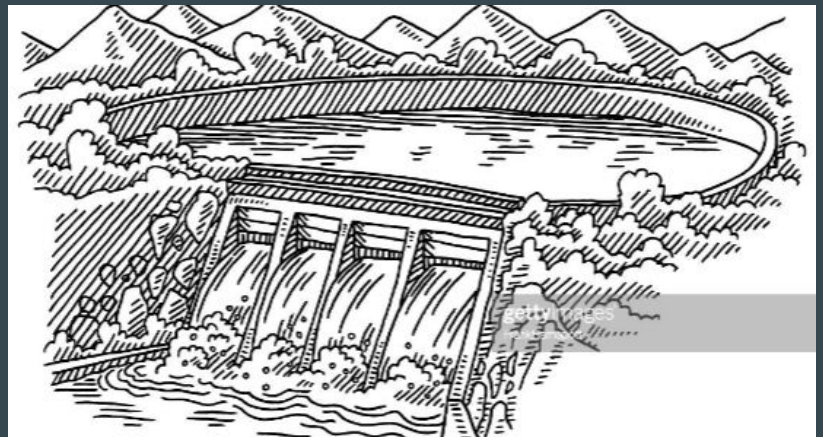
Project Reinforcement Learning

Group 11

Environment (MDP)

- Actions: idle · sell · buy
- State: storage level, electricity price, time features
- Reward: hourly trading profit
- Constraints: storage capacity · max flow

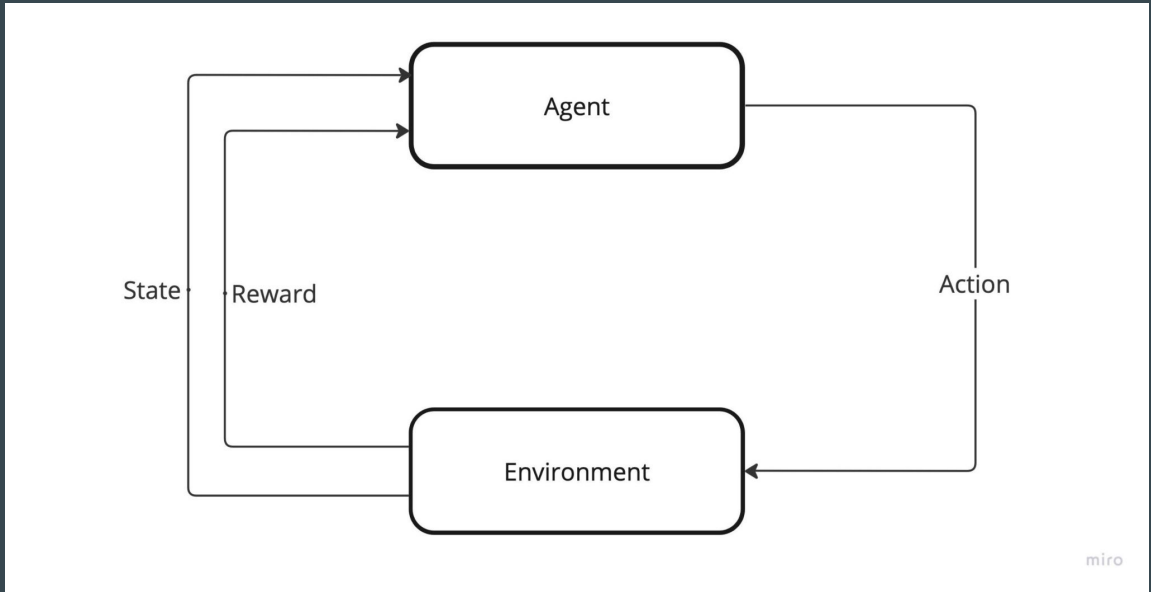


Baseline

- Uses current price only
- Buy / Sell at max volume
- Thresholds: 33rd & 67th price percentiles
- No learning. No future price access

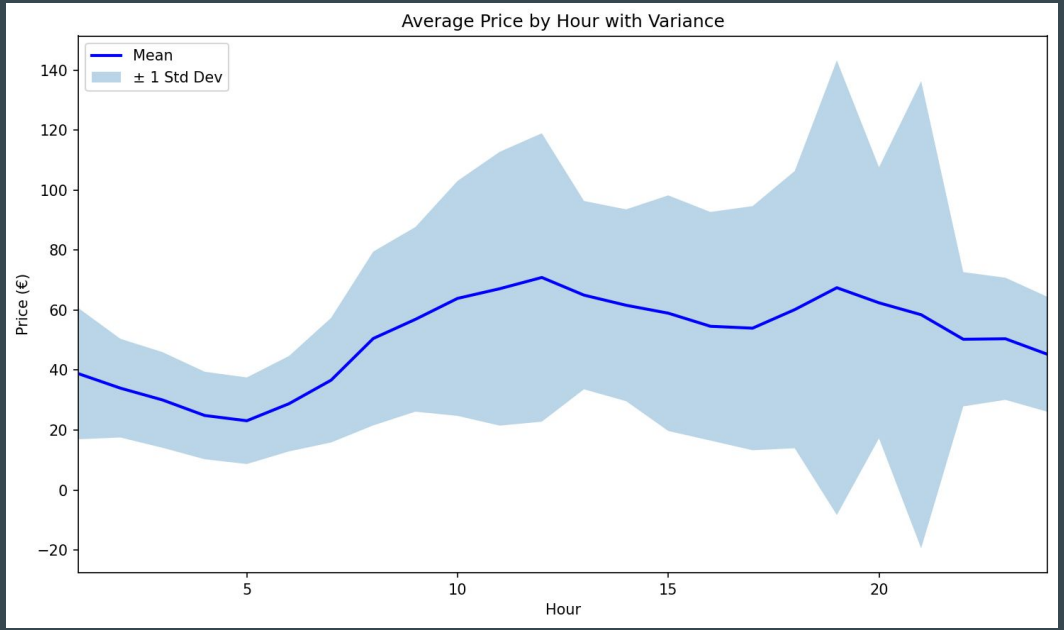
Tabular Q-Learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$



Data analysis: Variability & Patterns

Hourly

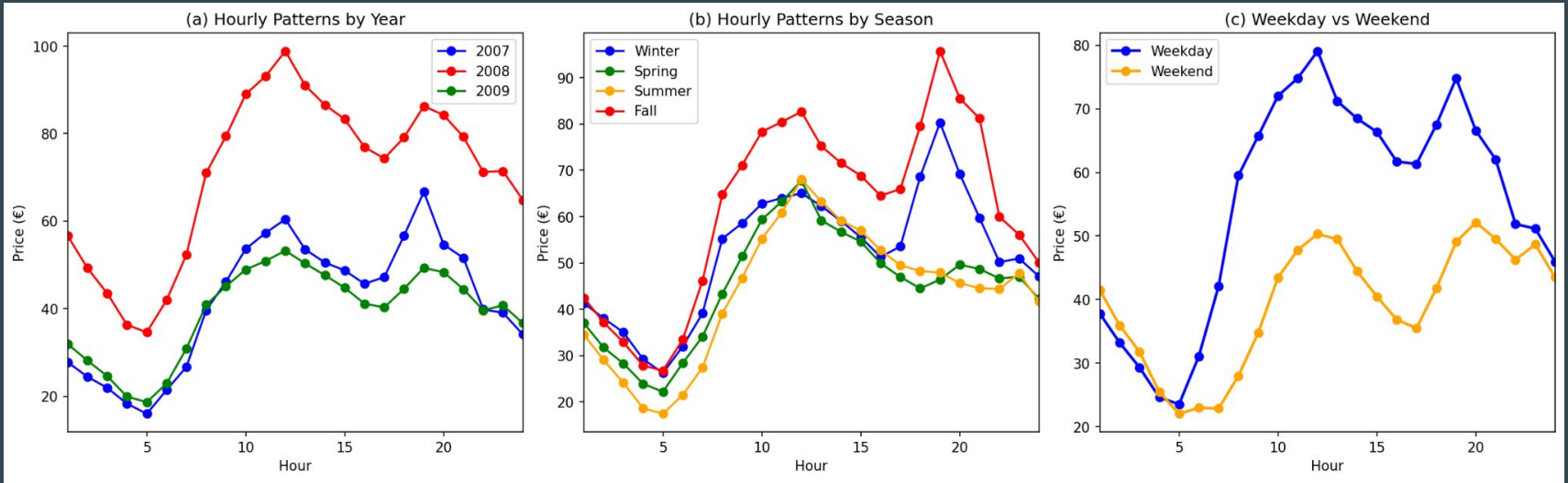


Data analysis: Variability & Patterns

Annual

Seasonal

Daily



State = (storage_level, normalized_price_bin, hour_period, is_weekend, season)

(6 x 6 x 5 x 4 x 2 x 3) = 4.320

Feature Engineering

Feature	Type	Values	Purpose
`storage_bin`	Discretized	0–5 (6 bins)	Current water level
`normalized_price_bin`	Discretized	0–5 (6 bins)	Normalized market price
`hour_period`	Extracted	0–4 (5 periods)	Time-of-day pattern
`is_weekend`	Extracted	0–1 (binary)	Weekday/weekend pattern
`season`	Extracted	0–3 (4 values)	Seasonal pattern

State features for tabular Q-learning.

(6 x 6 x 5 x 4 x 2 x 3) = 4.320

Rewards Shaping

Scale	Mean PnL (€)	Std	Best	Worst	% of Baseline
0.5	29,648	2,530	32,333	23,490	74%
5	31,462	2,114	35,698	28,939	79%
10	34,062	2,712	39,226	30,975	85%
20	38,260	3,501	45,575	33,546	96%
30	44,023	3,233	50,450	40,508	110%
40	49,186	2,906	54,643	44,054	123%
50	**52,590**	**1,646**	**56,464**	**50,503**	**131%**
60	49,416	3,213	54,014	43,426	124%
70	47,053	2,549	50,884	42,240	118%
80	35,294	7,540	43,826	17,861	88%
100	13,911	3,277	19,388	9,573	35%

Scale sweep results ($\gamma = 0.9$, 160 episodes, 10 seeds)

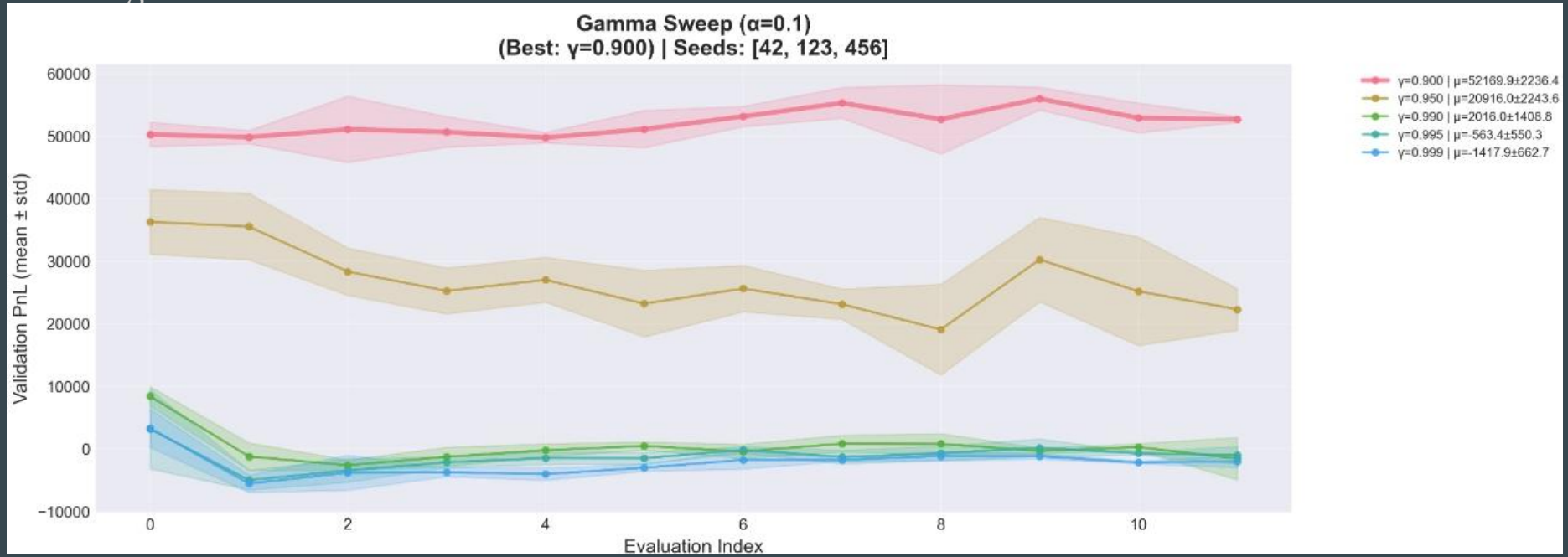
Parameter Tuning

1. 1D Parameter sweeps
   - α, γ, and episodes (one at a time)

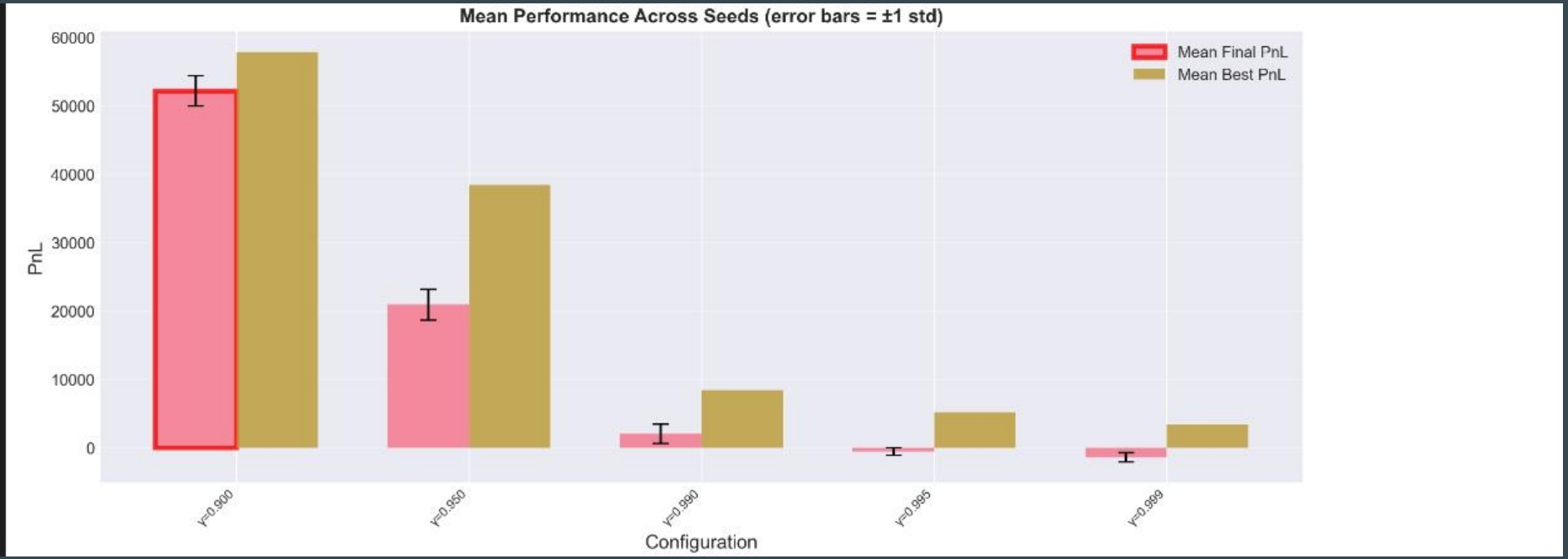
2. 2D Grid search on
   - $\alpha \times \gamma$ combinations to capture interactions
3. Adaptive episodes
   - small lr get 300 episodes (slower convergence)
4. Multi-seed validation: 3 random seeds per config

5. Selection: optimize for performance + reproducibility + stability

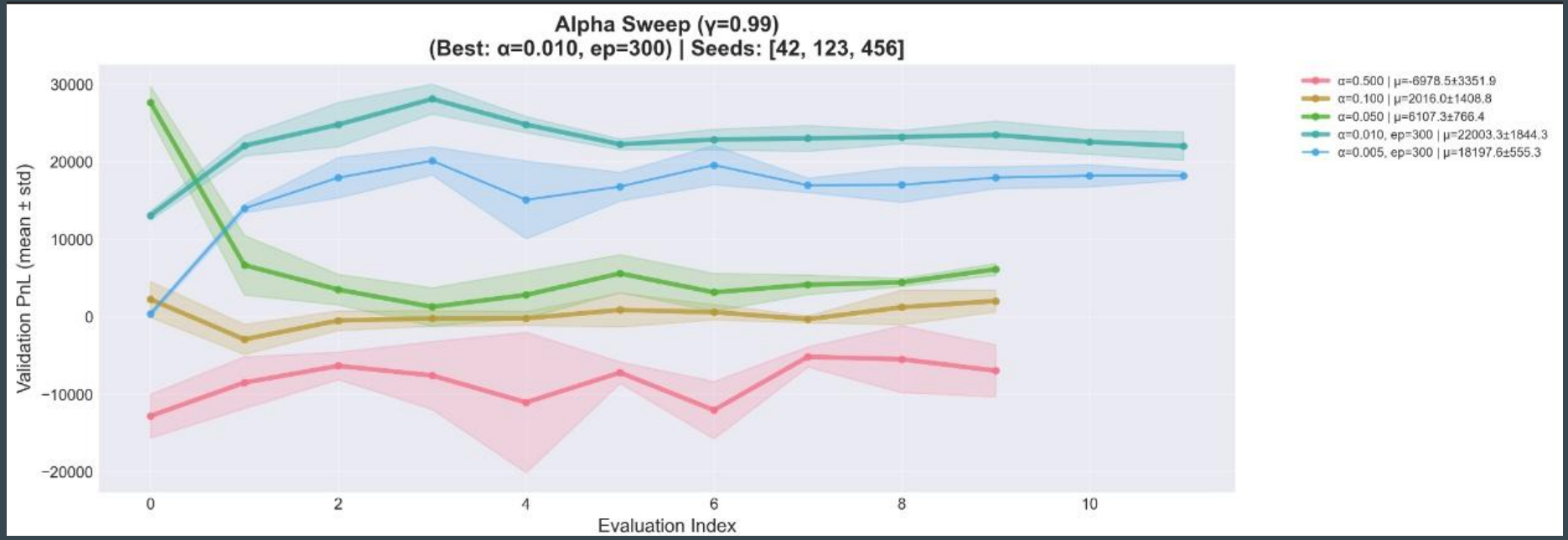
Parameter Tuning - 1D Parameter sweeps - Multi-seed validation: 3 random seeds config



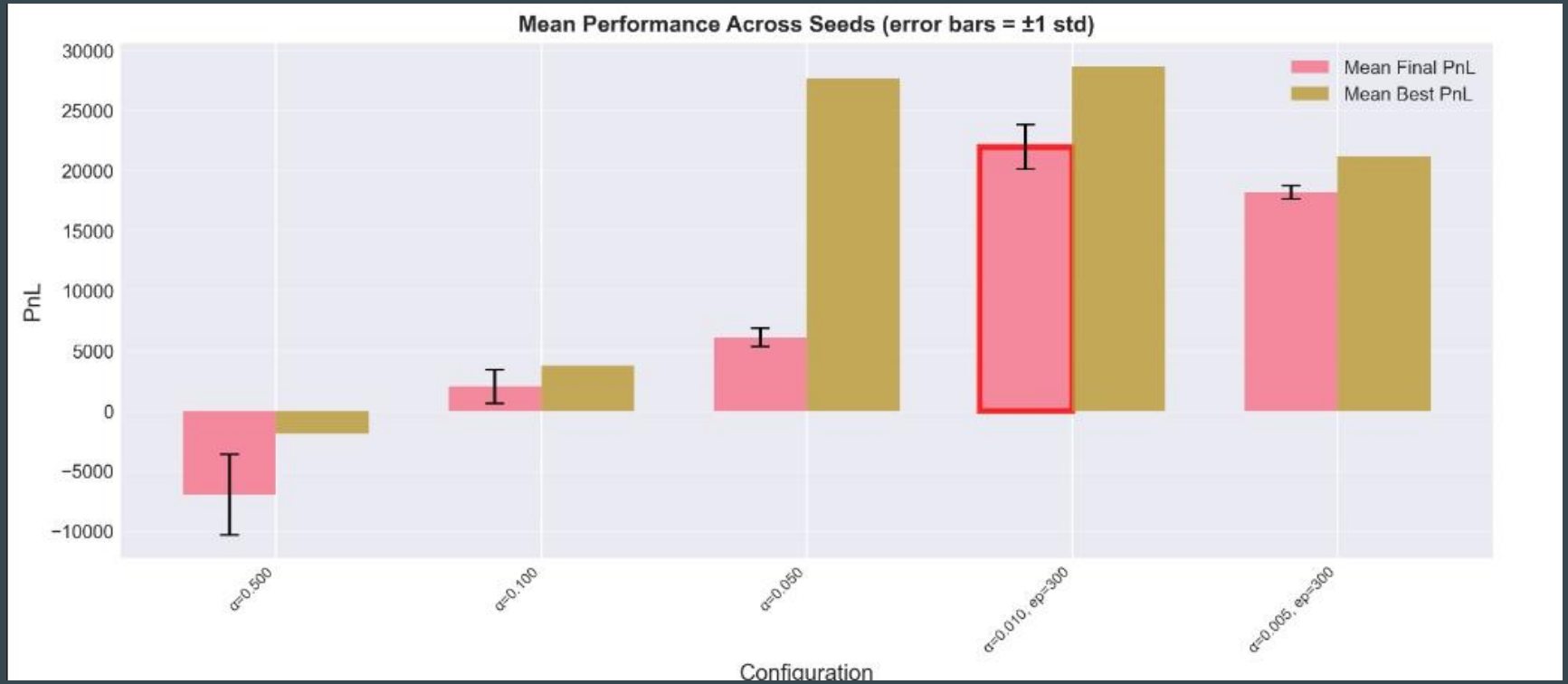
Parameter Tuning - 1D Parameter sweeps



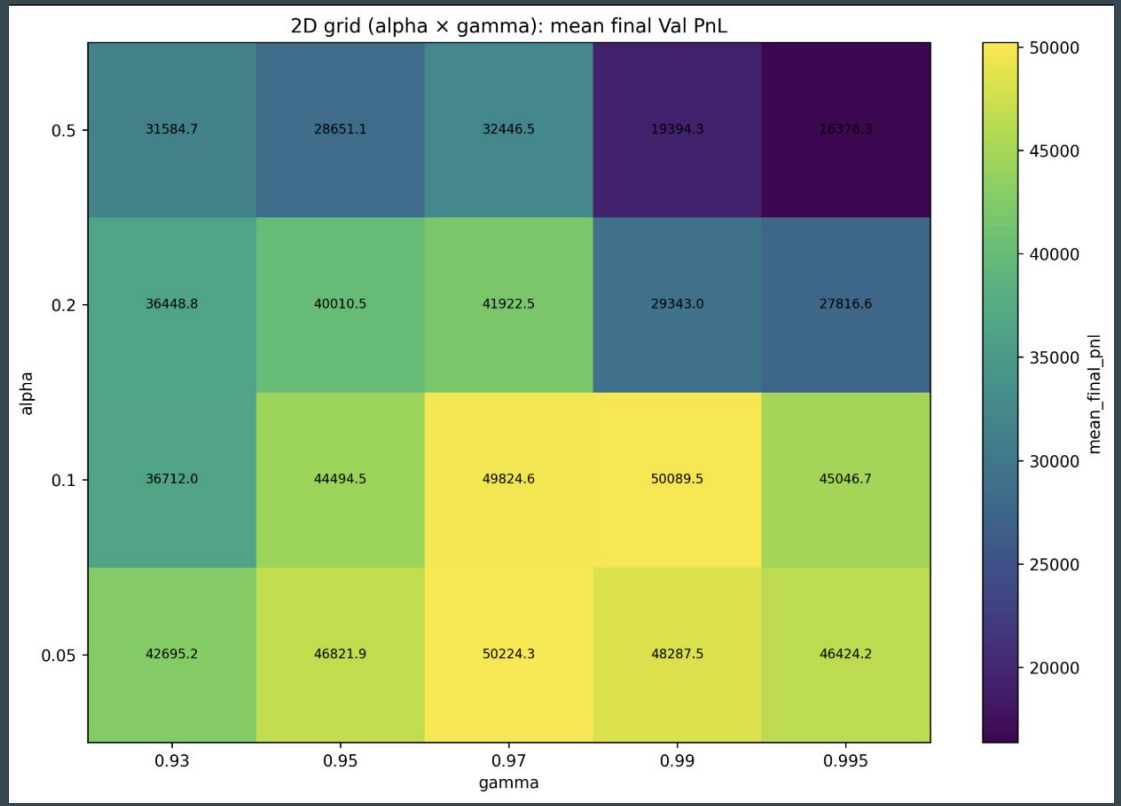
Parameter Tuning - 1D Parameter sweeps, Adaptive episodes



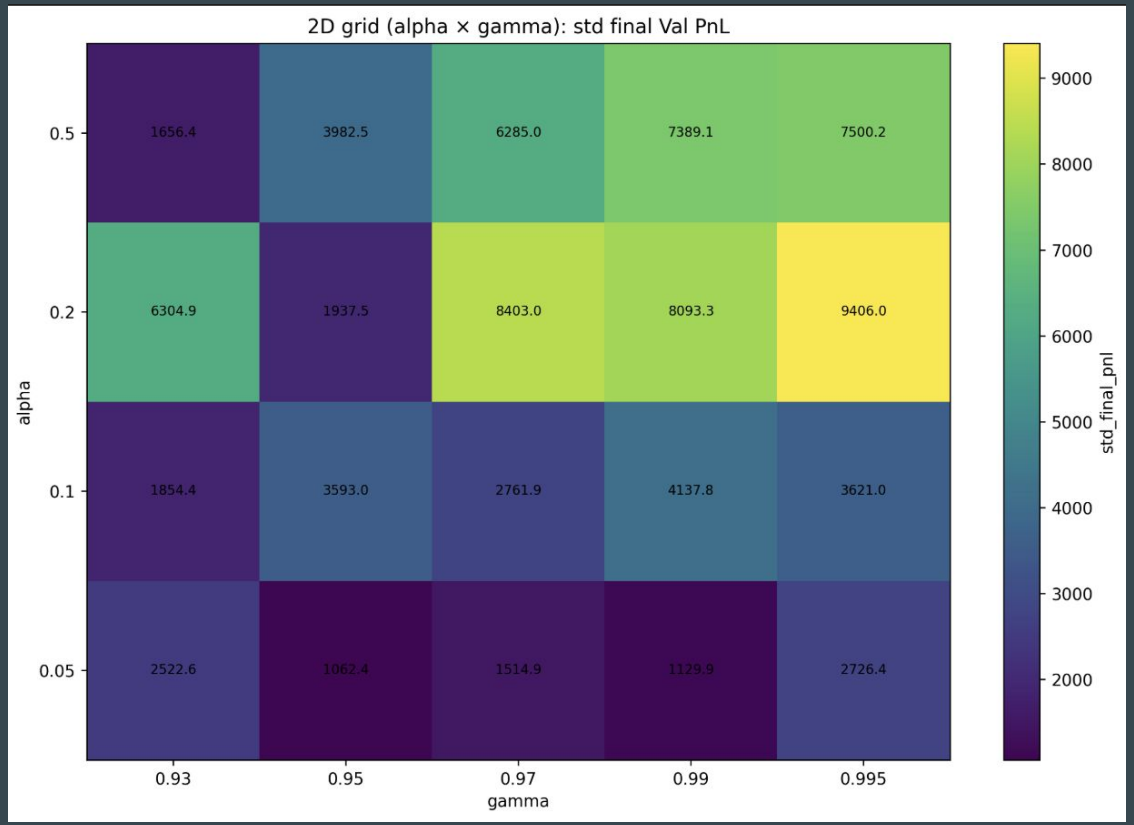
Parameter Tuning - 1D Parameter sweeps



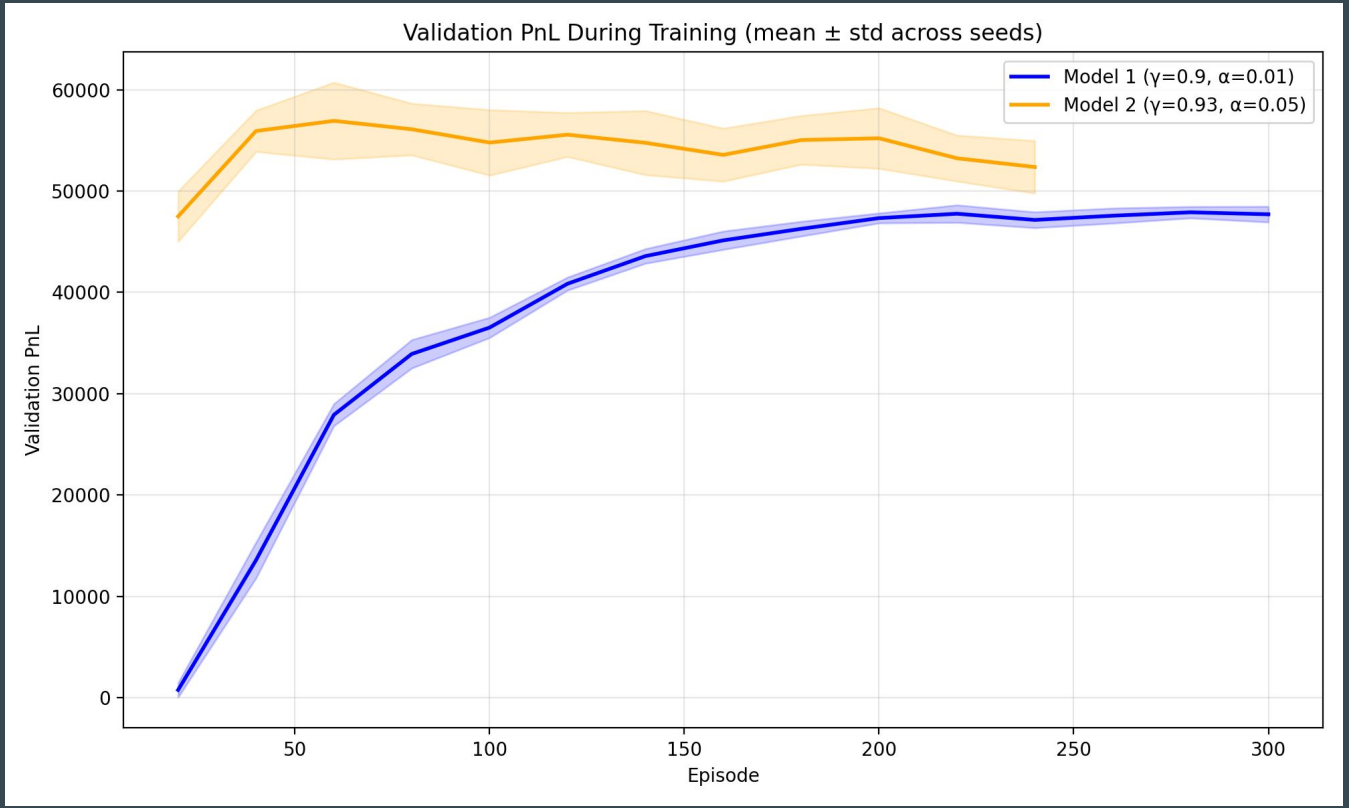
2D Grid Search:



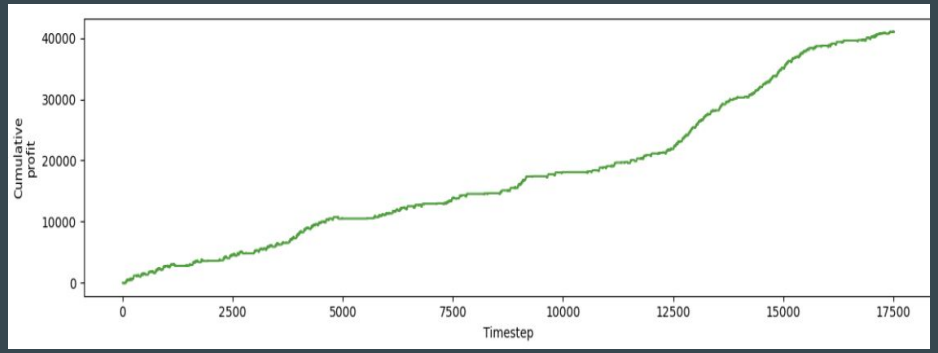
2D Grid Search:

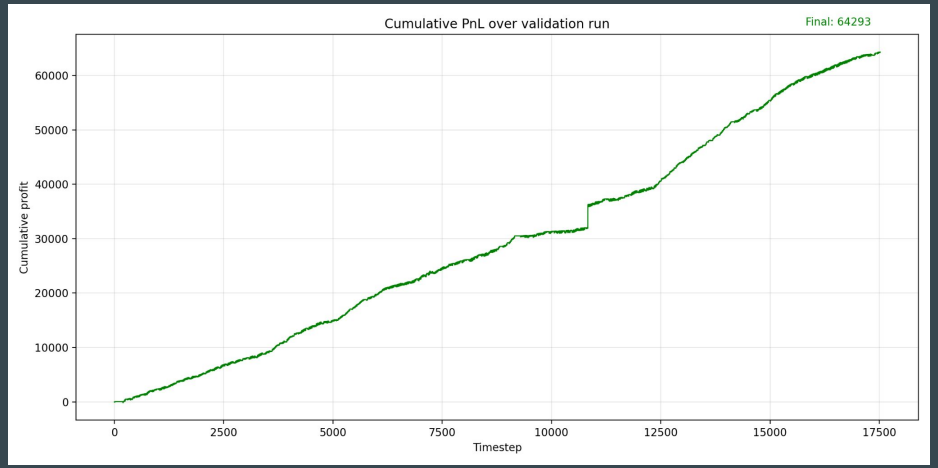


Model with best parameters, 10 seeds, select best num. of episodes

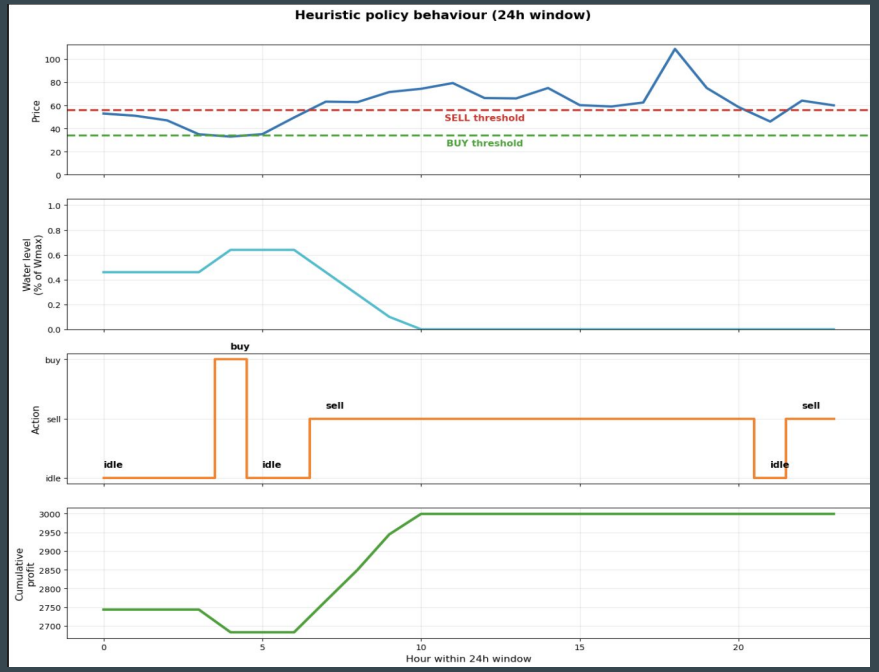


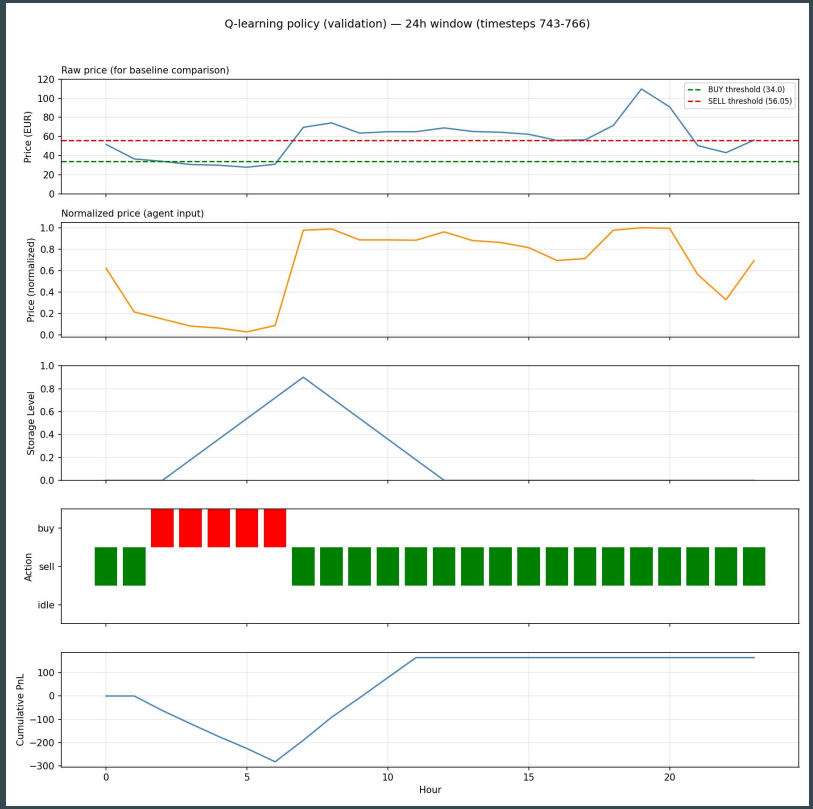
Comparing Agents





Results / Limitations





Observations & Limitations

- RL achieves higher cumulative profit than baseline
- Baseline shows stable, conservative behavior

→ RL exploits high-opportunity periods more aggressively

→ Action imbalance: buy $\gg$ sell despite reward shaping

→ Enters peak-periods for selling with an empty storage

→ Coarse discretization near storage limits

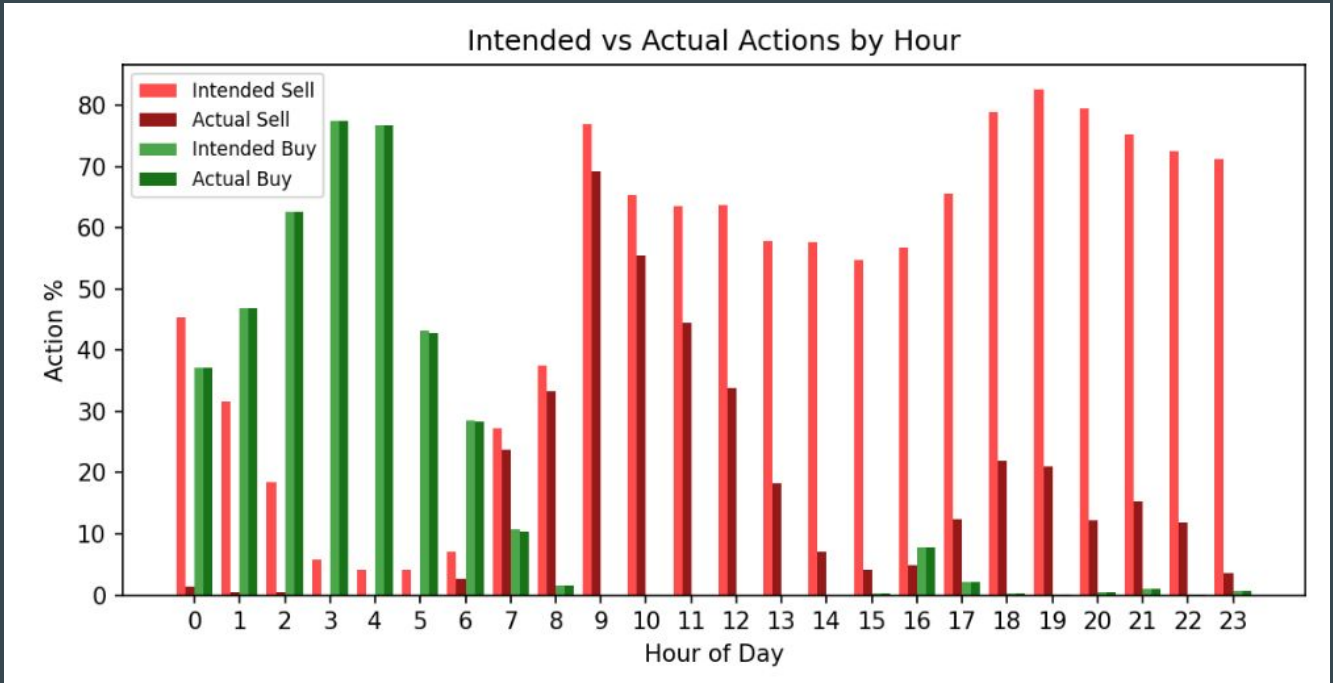
Future work

- Finer / adaptive state representation
- Deep RL with continuous states
- Variable buy / sell volumes

Thank you for listening

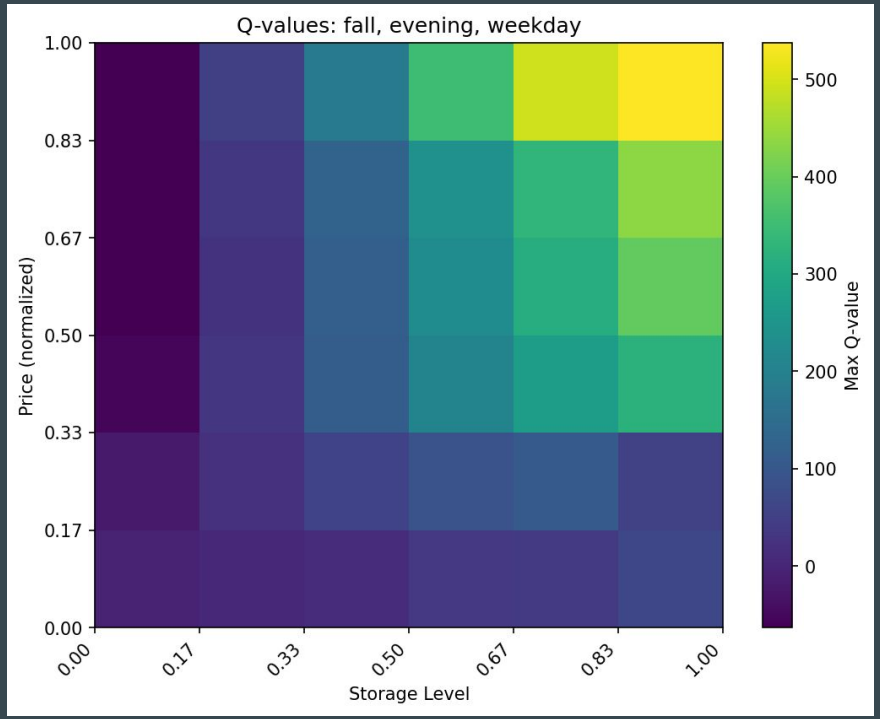
...

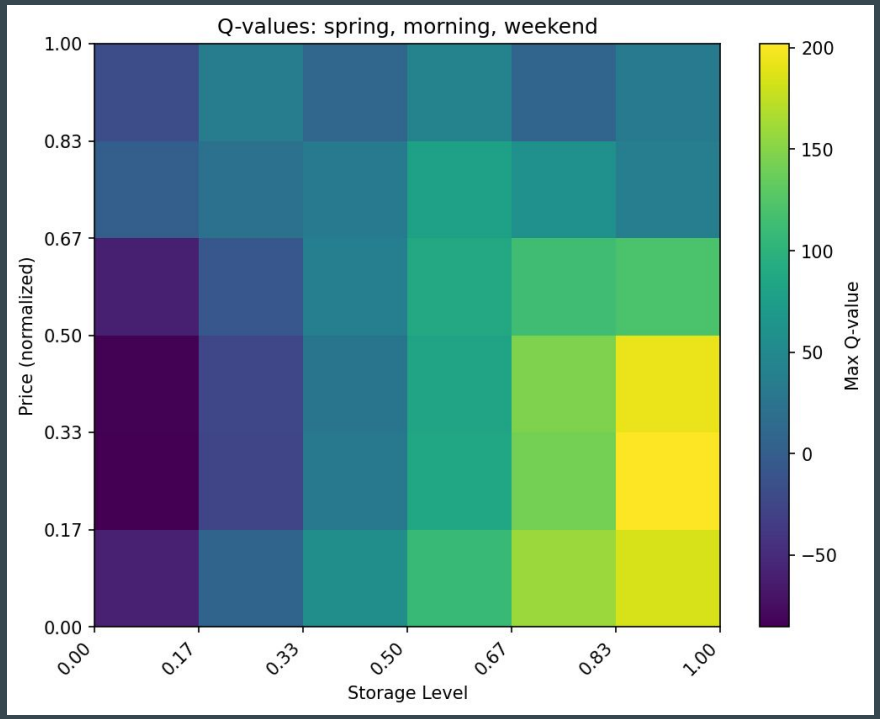
Clàudia Domènech
Alexia Spinei
Satiga Godrie



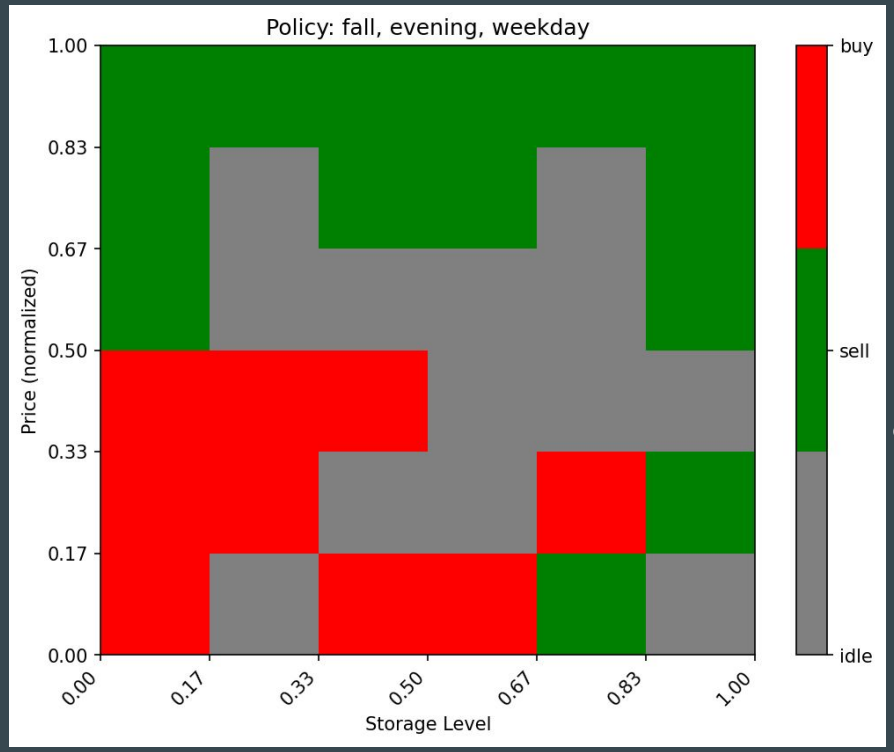
Learned buying and selling pattern, disregarding storage state

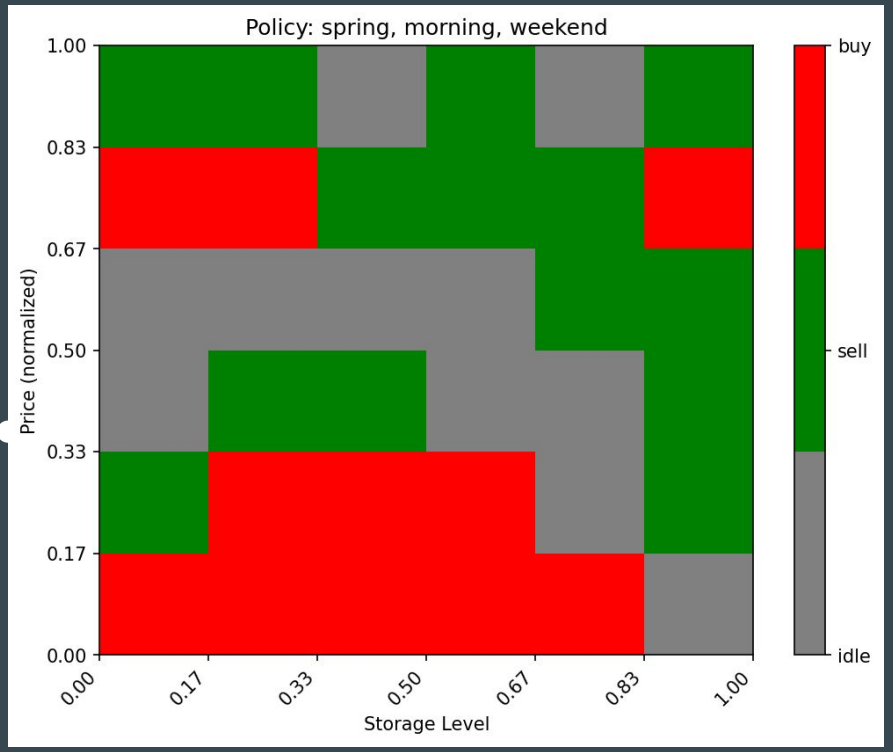
Q Values


Q-values: fall, evening, weekday
Price (normalized)
1.00
0.83
0.67
0.50
0.33
0.17
0.00
Storage Level
0.00
0.17
0.33
0.50
0.67
0.83
1.00
Max Q-value
500
400
300
200
100
0


Q-values: spring, morning, weekend
Price (normalized)
1.00
0.83
0.67
0.50
0.33
0.17
0.00
Storage Level
0.00
0.17
0.33
0.50
0.67
0.83
1.00
Max Q-value
200
150
100
50
0
−50

Policy


Policy: fall, evening, weekday
Price (normalized)
1.00
0.83
0.67
0.50
0.33
0.17
0.00
0.00
0.17
0.33
0.50
0.67
0.83
1.00
Storage Level
buy
sell
idle


Policy: spring, morning, weekend
Price (normalized)
1.00
0.83
0.67
0.50
0.33
0.17
0.00
0.00
0.17
0.33
0.50
0.67
0.83
1.00
Storage Level
buy
sell
idle