By Fair Means or Foul - Supplementary Material

Michael Schlechtinger[1] , **Damaris Kosack**[2] , **Franz Krause**[1] and **Heiko Paulheim**[1]
[1]University of Mannheim - Chair of Data Science
[2]University of Mannheim - Chair of Law
{schlechtinger, damaris.kosack, franz.krause, heiko.paulheim}@uni-mannheim.de

A Ablation Study

In addition to the main scenarios of this study, we aim to understand the importance of certain simulation and algorithm (hyper-)parameters and how they collectively affect the agents' behavior. Similar to the main scenarios, the prices and profits of the episodes are computed as the mean values derived from the corresponding step prices, which have also undergone LOWESS smoothing. Each experimental iteration comprises 10,000 episodes, each encompassing 365 sequential steps. Every plot comprises the price and profit achieved in the episodes and an insight into the individual step pricing of the last episode. We employ PPO during these runs.

A.1 Number of Agents

Initially, we endeavor to scrutinize the influence exerted by both the quantity of agents and their respective market entry timings upon the outcomes derived from the simulation. To execute that, we ran a simulation with 15 agents as well as two simulations where agents join throughout the run.

15 Agents (Figure 1)
We discern a pricing behavior that mirrors the outcomes seen in the scenarios featuring fewer agents. Nevertheless, the visually intricate and seemingly stochastic pricing trajectory contrasts distinctly with the notably coherent profit graph. Adding to that, we observe the same oscillation pattern within the step pricing behavior.

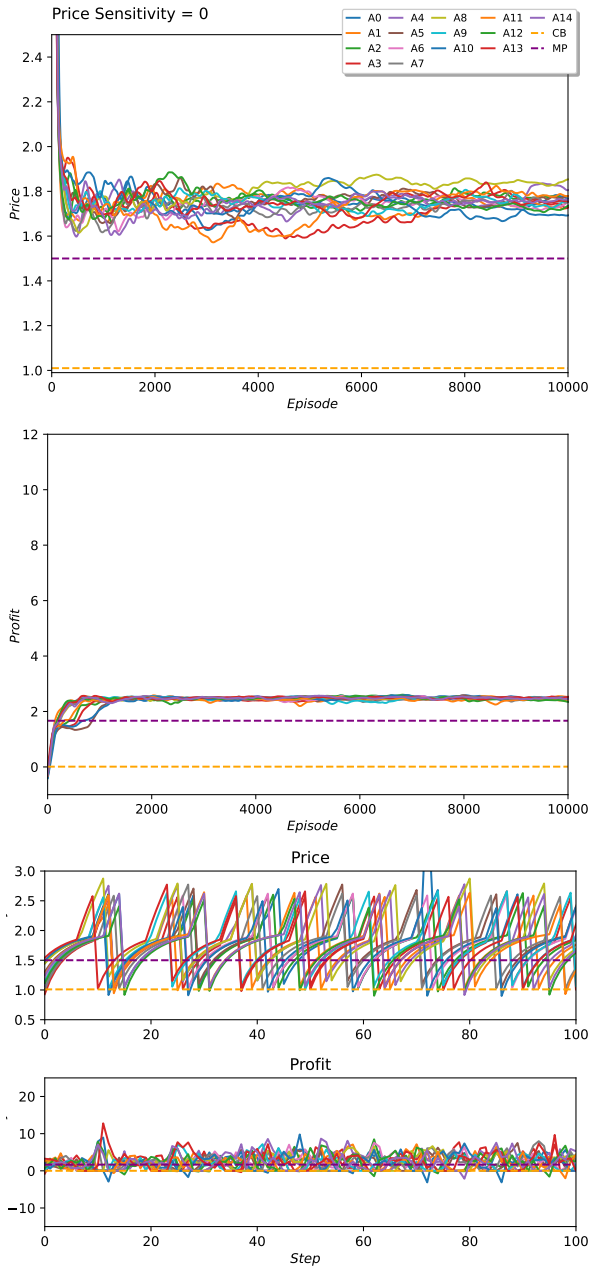

Figure 1: PPO, 15 Agents

Agents Join Later (Figures 2, 3)
To conduct a more comprehensive analysis of the influence wielded by varying the number of agents, we employ a run starting with two agents. In episodes 3000 and 6000 an additional agent is introduced to the market. We aim to investigate the response of pre-trained agents to this alteration. The primary run (cf. Figure 2) adheres to conventional settings akin to scenario A, while the secondary run (cf. Figure 3) modifies the agents' observation space by concealing competitors' pricing information.

The insertion of a new agent prompts a behavioral recalibration among the existing agents. Following a noticeable disruption within the graphs, this leads to a gradual collective adjustment of the already trained agents towards a reduced profit level, while pricing remains relatively stable. Notably, the main difference between the two scenarios lies in the repercussions of the untrained agent's market entry. In scenario A, discernible disruptions manifest in both pricing strategies and profit dynamics. Conversely, scenario B demonstrates a comparably smoother adaptation process. We attribute this behavior to the reduced amount of information necessary to solve the problem. The step-wise pricing behavior and chosen strategies remain consistent with the main approach of this paper.

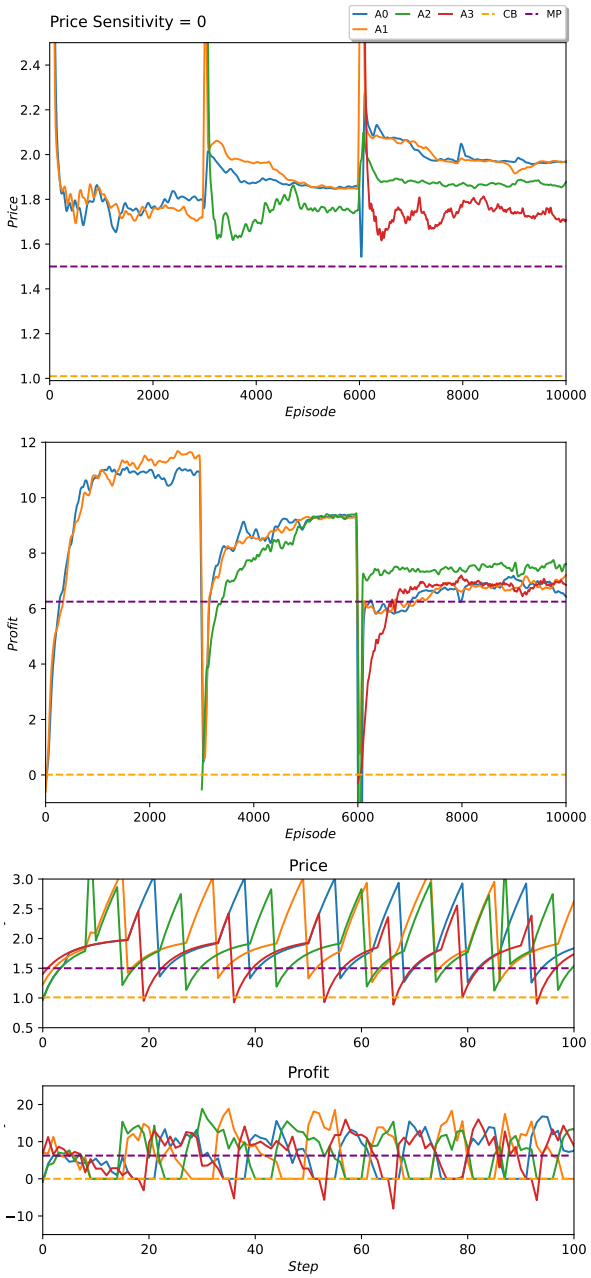

Figure 2: PPO, 3 Agents, One Agent Joining at Episode 3000, One Agent Joining at Episode 6000

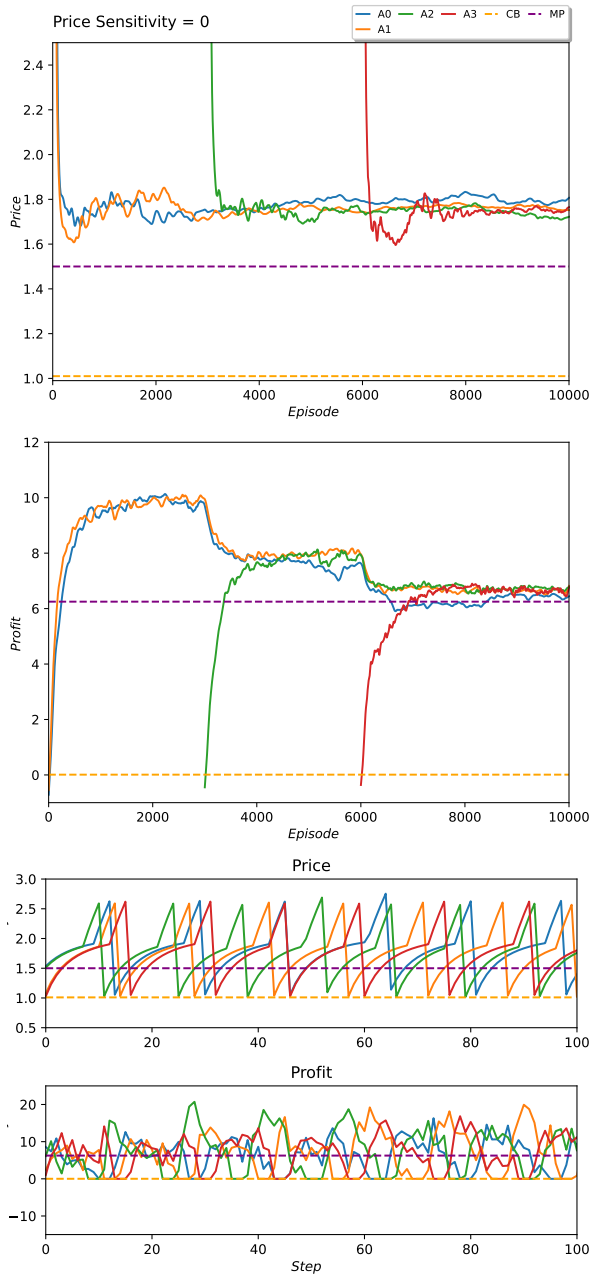

Figure 3: PPO, Blind, 3 Agents, One Agent Joining at Episode 3000, One Agent Joining at Episode 6000

A.2 Learning Rate (cf. Figures 4, 5)

Subsequent to this point, each experimental iteration incorporates 'blind' agents, characterized by a constrained action space, with the intention of potentially heightening collaboration challenges. In this context, we adjusted the agents' learning rates from 1e-5 to 1e-4 and 5e-5. This modification aims to highlight potential variations in learning speed in order to impede the agents' capacity to establish a collusive outcome. The modifications to the learning rates do not result in visible impacts on agent performance, selected strategies, or discernible patterns. Nevertheless, we observe a divergence within the later stages of the run associated with a learning rate of 5e-5. Upon attaining an equilibrium-like state, agent 0 enacts a reduction in its pricing strategy, thereby yielding a diminished profit margin. Given the predominance of runs converging toward equilibrium within a certain episode range, we anticipate this isolated occurrence to be an outlier. Nonetheless, this divergence could be a potential lead toward collusion prevention.

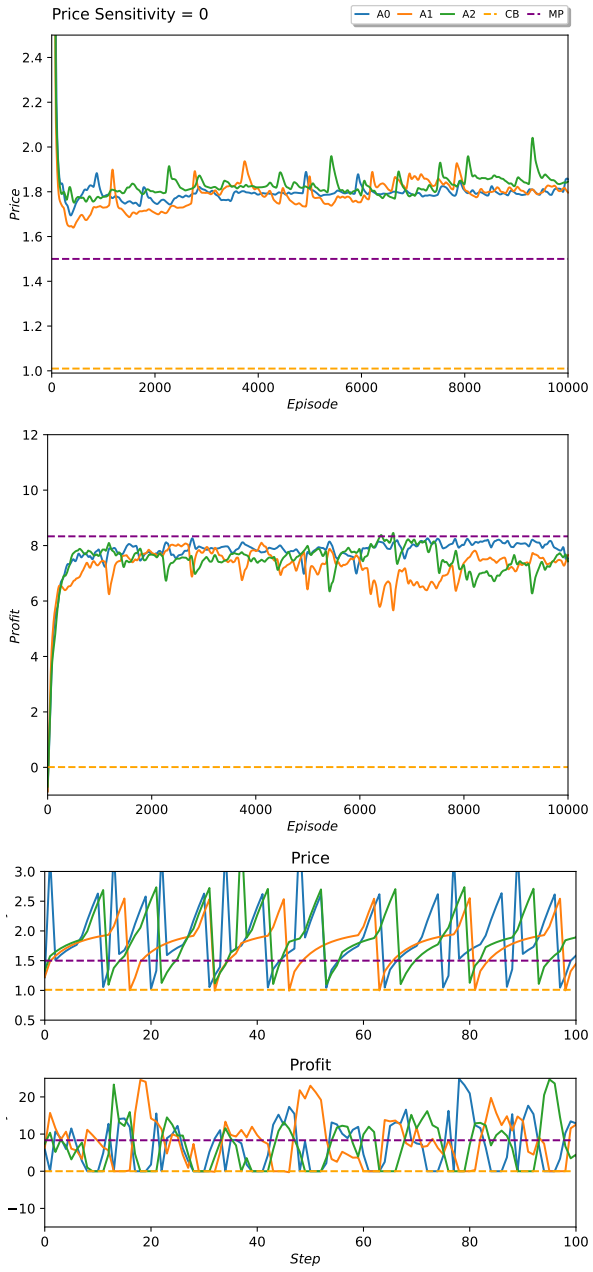

Figure 4: PPO, Blind, 3 Agents, Learning Rate: 1e-4

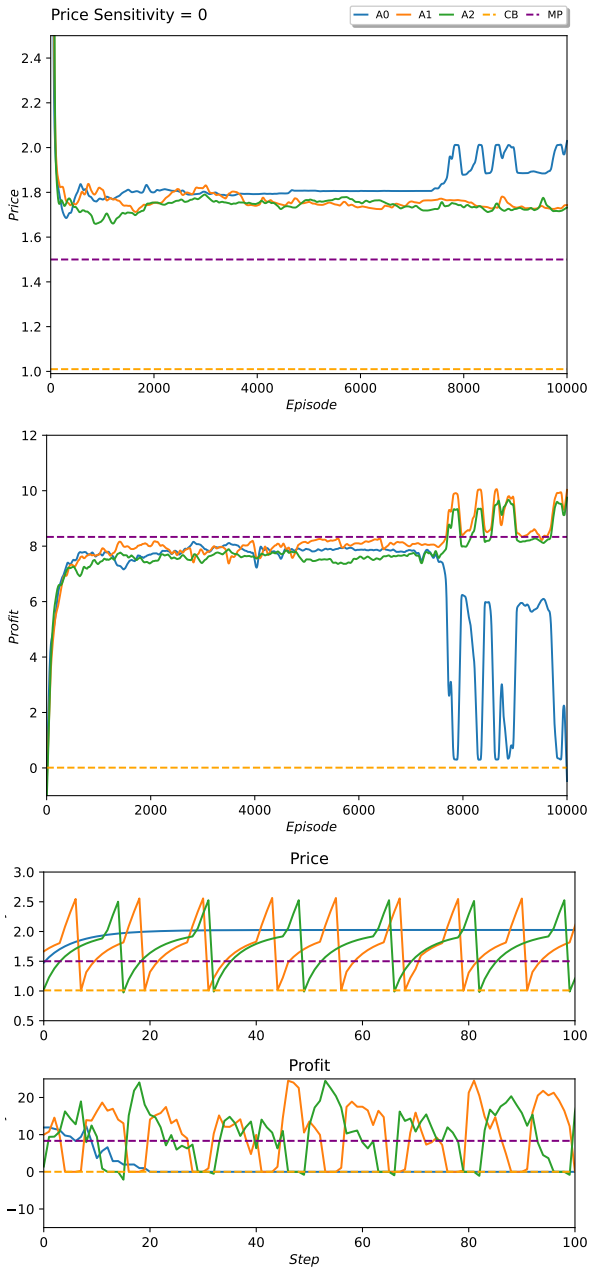

Figure 5: PPO, Blind, 3 Agents, Learning Rate: 5e-5

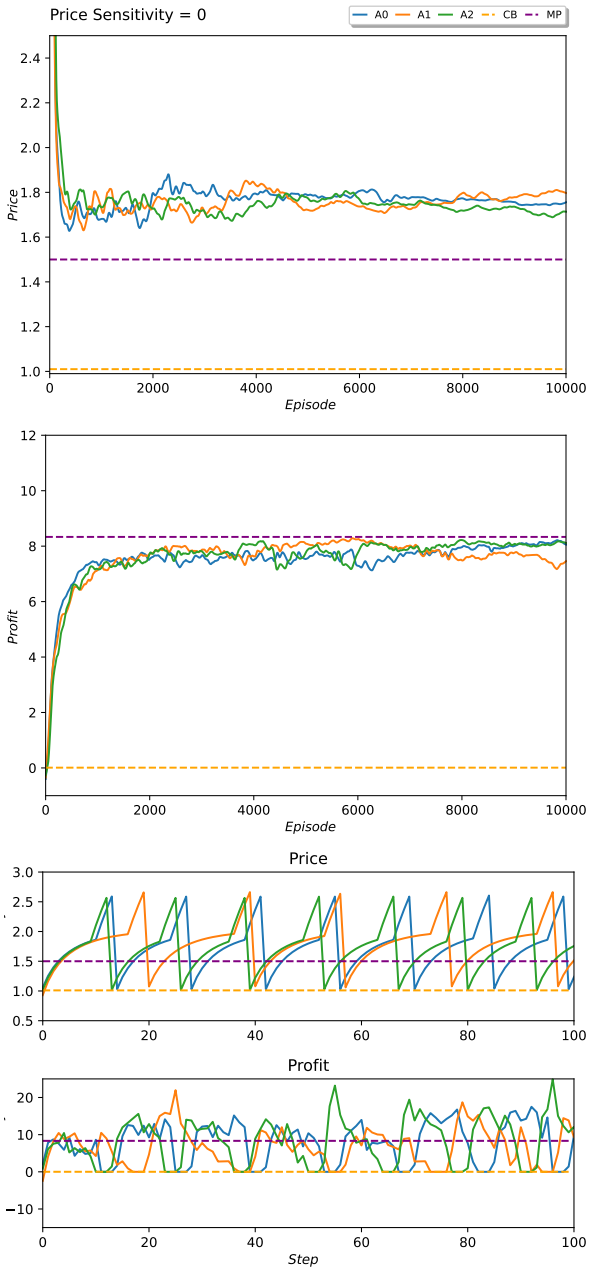

Figure 6: PPO, Blind, 3 Agents, Hidden Layer Neurons: 128x128

A.3 Hidden Layer Neurons (Figures 6, 7)

We reduced the neurons for the two hidden layers from 256x256 to 128x128 and 64x64 respectively. With this simplification of the neural networks, we aim to impede the agents' ability to achieve a collusive outcome. Despite this significant modification, the agents were able to achieve and sustain a collusive outcome.

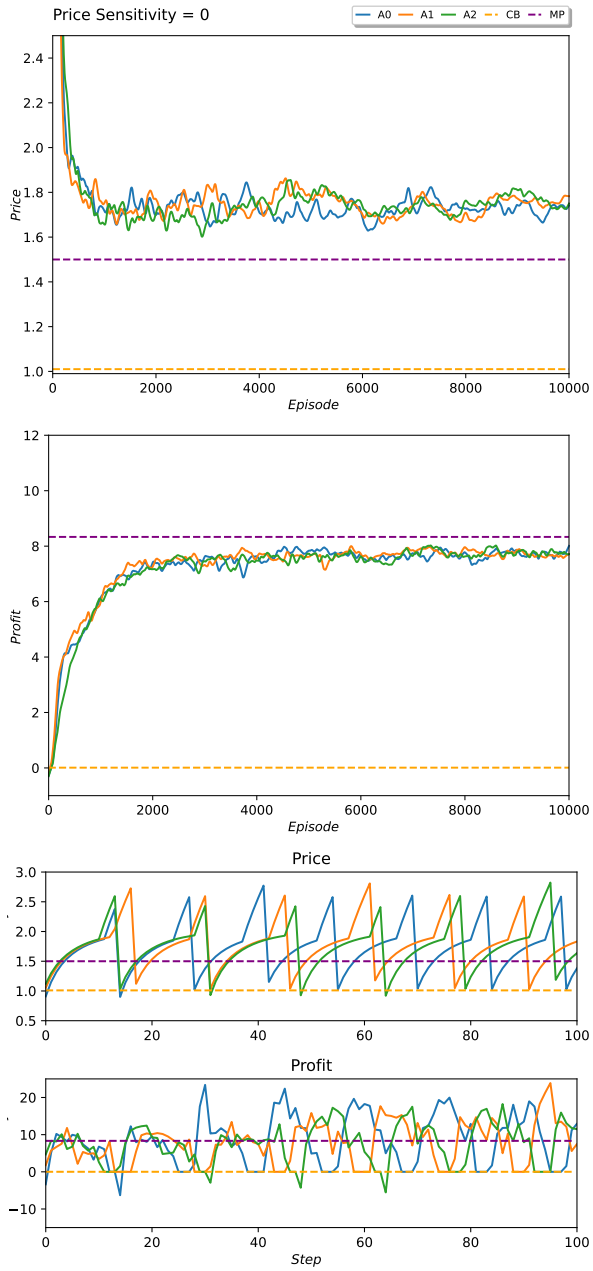

Figure 7: PPO, Blind, 3 Agents, Hidden Layer Neurons: 64x64

A.4 Action Gradients (Figure 8, 9, 10)

In order to prompt different strategies, we increased the action gradients from 7 to 21, 51, and 71 respectively. This adjustment affords the agents the opportunity to finely calibrate pricing, albeit at the expense of heightened computational intricacy. This adjustment generally leads to an increased profit with similar prices, a more severe price stagnation, as well as a quicker time to converge. However, the strategies applied by the agents significantly diverge from the ones applied in the main scenarios of this study.

Employing 21 gradient steps yields the most substantial profit accumulation observed across the experimental sessions, surpassing even the profits attained within scenario A. The agents accomplish this feat through an approach that exhibits distinct characteristics while maintaining a fundamental similarity. Each agent follows its own unique oscillation pattern, yet collectively, they converge toward comparable long-term profit levels. Analogous to the main scenarios of this paper, the agents adeptly capitalize on the simulation dynamics by designating a single agent to sell at a reduced price, thereby enabling the remaining agents to price above the monopolistic threshold. We witness analogous behavior when applying 51 and 71 gradients. Despite the heightened array of action choices resulting from the increased gradients, the agents converge toward a state of market stasis characterized by recurrent price repetition. Nevertheless, within this state of repetition, one of the agents persists in executing an oscillation pattern, thereby propelling the collective profit beyond the monopolistic benchmark.

This adjustment not only shows that there seems to be a sweet spot for the agents (ca. 21 gradients), but it also highlights that, with proper parametrization, sellers could potentially achieve higher profits as well as an equilibrium state that supersedes the ”unaltered” scenarios despite the restricted observation space.

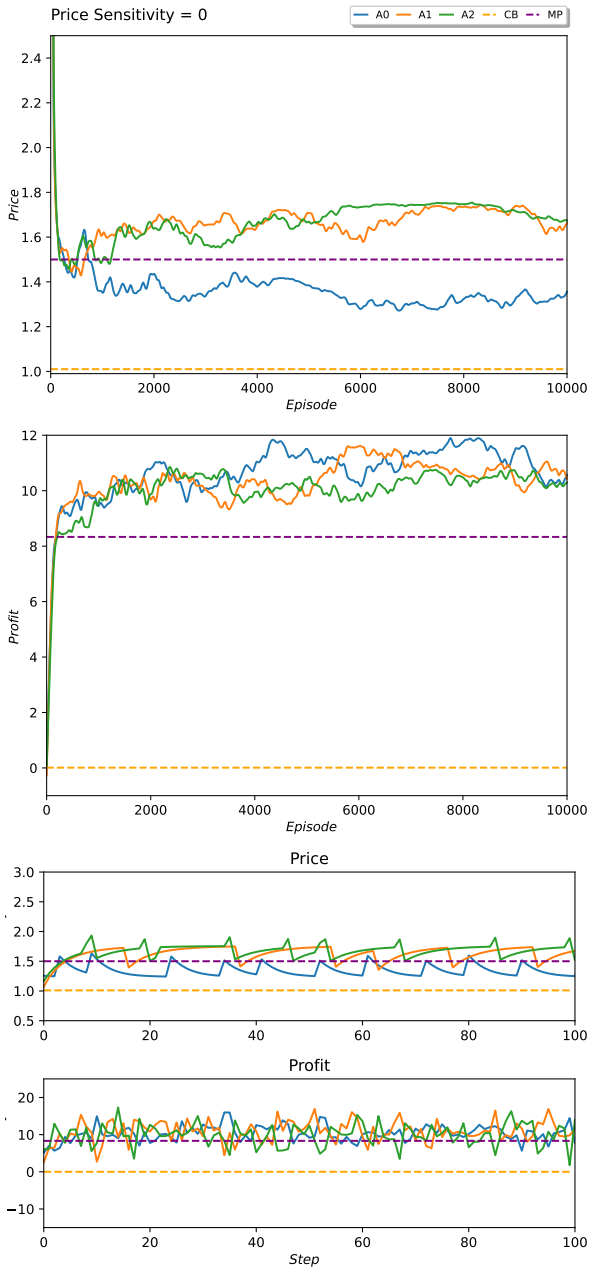

Figure 8: PPO, Blind, 3 Agents, 21 Action Gradients

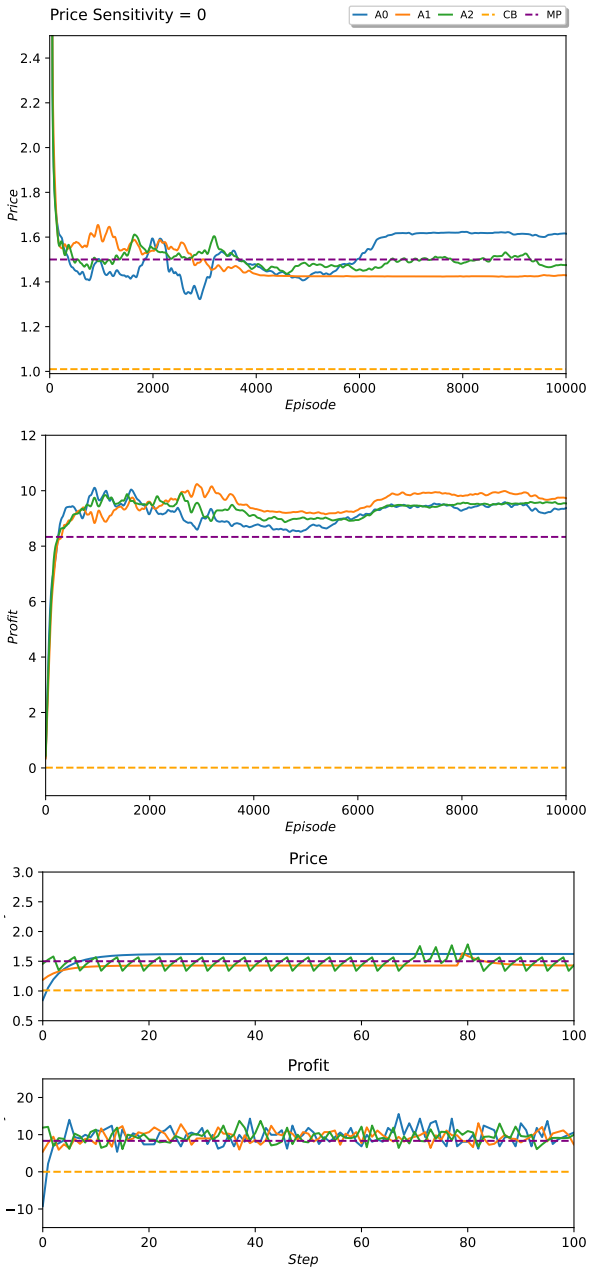

Figure 9: PPO, Blind, 3 Agents, 51 Action Gradients

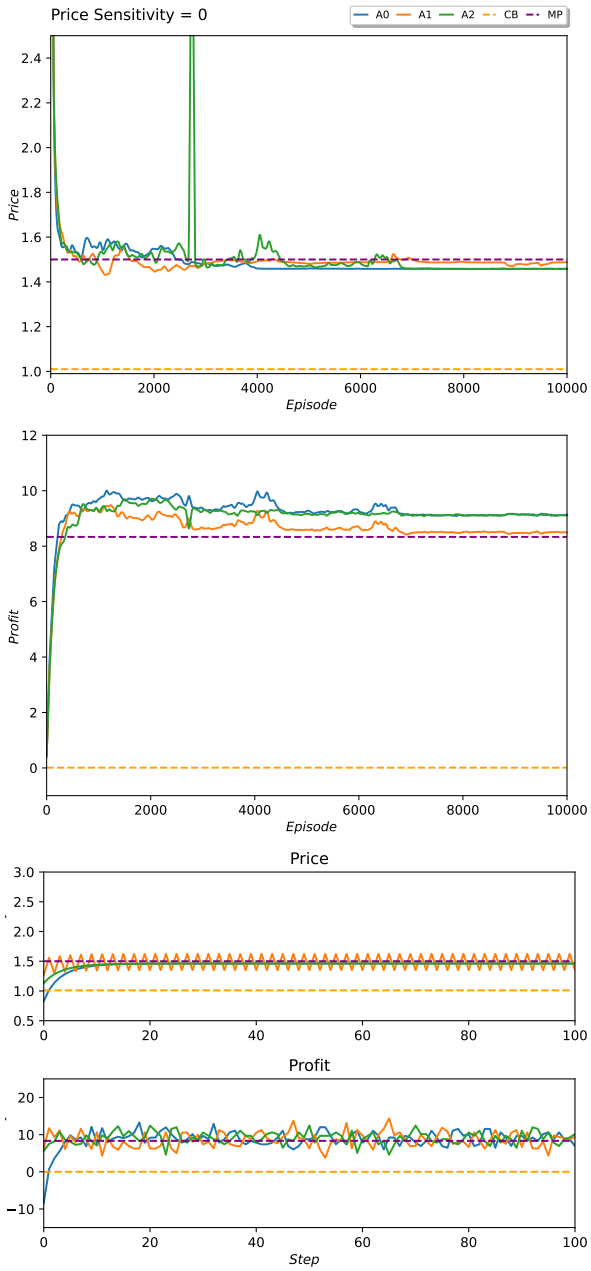

Figure 10: PPO, Blind, 3 Agents, 71 Action Gradients

Table 1: Descriptive Statistics of the runs in each scenario

Scenario	n	μ	Algo	$\overline{P(10000)}$	$\overline{\epsilon(10000)}$	$\sigma\big(P(1000)\big)$	t_{EUC}	Δ
A	3	0	PPO	1.77	9.05	0.01629	1060	1.0749
			DQN	1.75	8.91	0.01799	5247	1.07001
		0.5	PPO	1.87	7.48	0.0644	2532	0.88613
			DQN	1.78	6.03	0.00337	/	/
		1	PPO	2.0	5.57	0.12697	/	/
			DQN	2.0	4.85	0.13228	/	/
	5	0	PPO	1.75	6.15	0.03123	1282	1.22834
			DQN	1.69	5.61	0.02267	6223	1.12777
		0.5	PPO	1.83	4.34	0.03149	797	0.84309
			DQN	1.91	4.13	0.06675	/	/
		1	PPO	1.95	3.11	0.0641	8004	0.63864
			DQN	2.17	2.9	0.0947	/	/
B	3	0	PPO	1.76	7.8	0.01961	958	0.94034
		0.5	PPO	1.73	6.54	0.02549	1236	0.80212
		1	PPO	1.75	5.1	0.03248	1049	0.61831
	5	0	PPO	1.74	5.64	0.02323	907	1.12276
		0.5	PPO	1.75	4.15	0.02231	530	0.84312
		1	PPO	1.82	2.92	0.02335	722	0.57719

Table 2: RL Settings and hyperparametrization

Steps per episode	365
Number of training iterations	10000
Hidden Layers	256 x 256
Learning Rate	1e-5
Gradient Clipping (PPO)	1
Batch size (DQN)	250
Batch size (PPO)	4000
Discount Factor (Gamma)	1
Replay Buffer Size (DQN)	50000
PPO Clip Param	0.3
Value Function Clipping Param (PPO)	10