Self-learning Agents for Recommerce Markets

Jan Groeneveld, Judith Herrmann, Nikkel Mollenhauer, Leonard Dreeßen, Nick Bessin, Johann Schulze Tast, Alexander Kastius, Johannes Huegle, Rainer Schlosser

Business & Information Systems Engineering (2023)

Appendix (available online via http://link.springer.com)

Appendix

A Hyperparameters

Parameter	Value
Learning Rate	$7 \cdot 10^{-4}$
Steps per update	5
Discount factor (γ)	0.99

Table 4: Hyperparameters for Advantage Actor Critic (A2C).

Parameter	Value
Learning Rate	$3 \cdot 10^{-4}$
Steps per update	2048
Minibatch size	64
Epochs per update	10
clip_range (ε)	0.3
Discount factor (γ)	0.99

Table 5: Hyperparameters for Proximal Policy Optimization (PPO).

Parameter	Value
Learning Rate	$3 \cdot 10^{-4}$
Experience Buffer Size	10^6
Steps until learning starts	100
Minibatch size	256
Entropy coefficient (α)	automated
Polyak coefficient (τ)	0.005
Discount factor (γ)	0.99

Table 6: Hyperparameters for Soft Actor Critic (SAC).

B Further Diagrams On Training

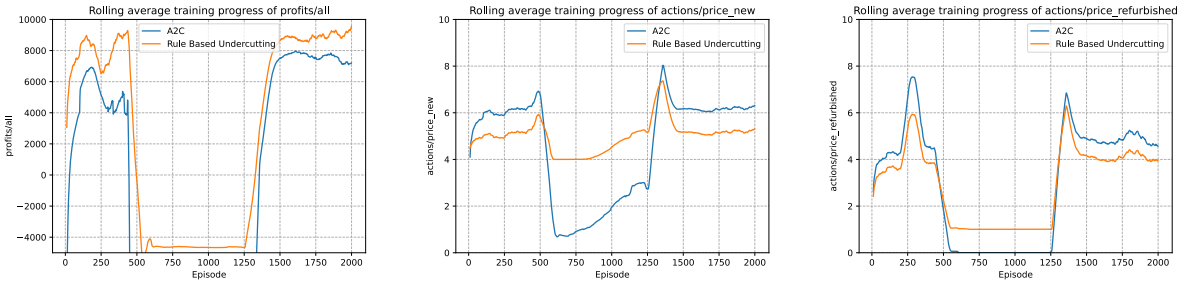

Figure 13: Detailed view of an A2C training run to visualize instability: (left) learning curve with rapid initial rise, then crashes and recoveries; (center) average selection of new prices, showing significant fluctuations; (right) average selection of used prices, also unstable.

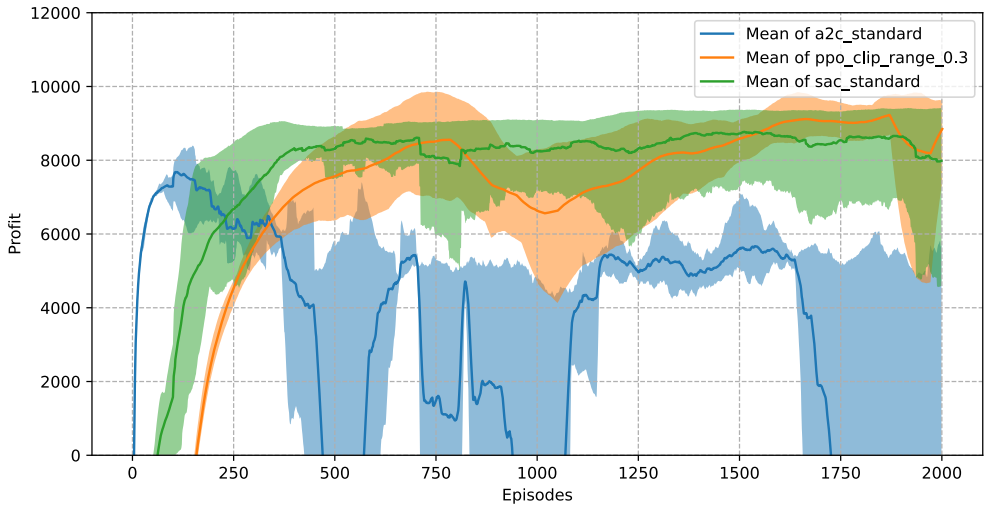

Figure 14: Learning curves of four A2C, PPO, and SAC runs in self-play with mixed reward; algorithms are trained over 2 000 episodes (4 runs).

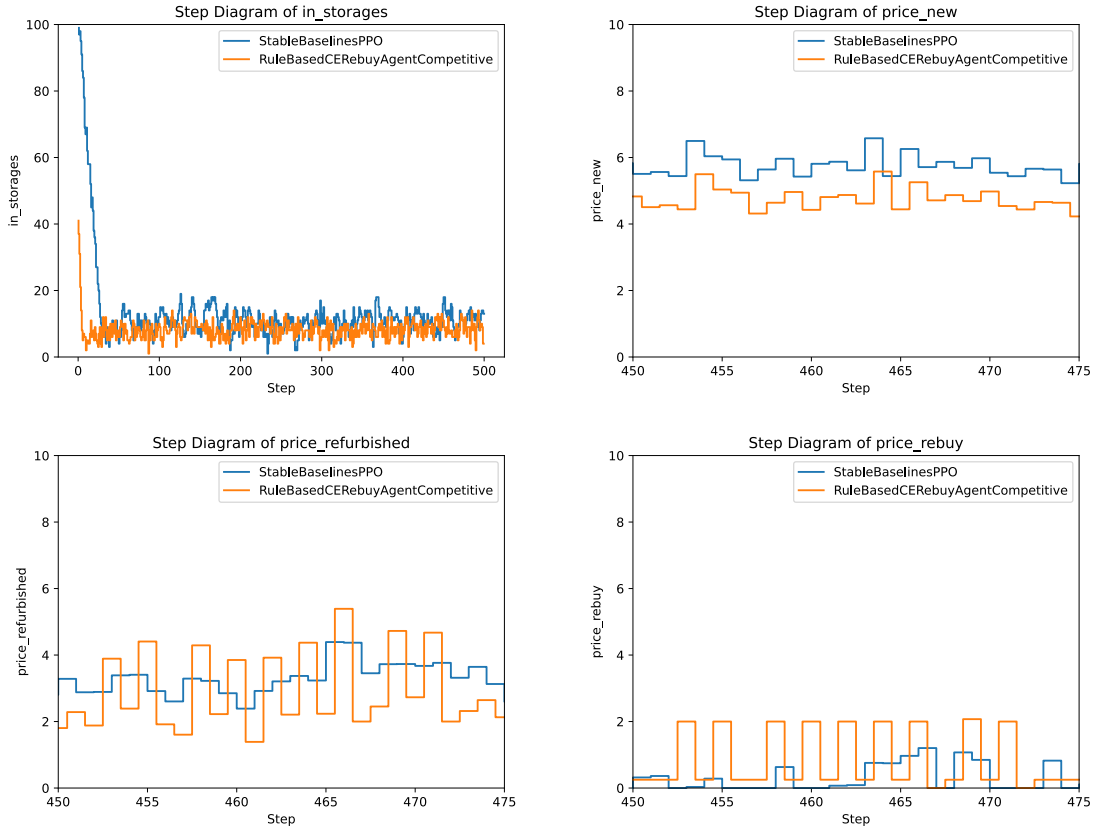

Figure 15: Details of an episode in which a PPO agent (trained with self-play and mixed reward features) competes against the undercutting rule-based competitor.

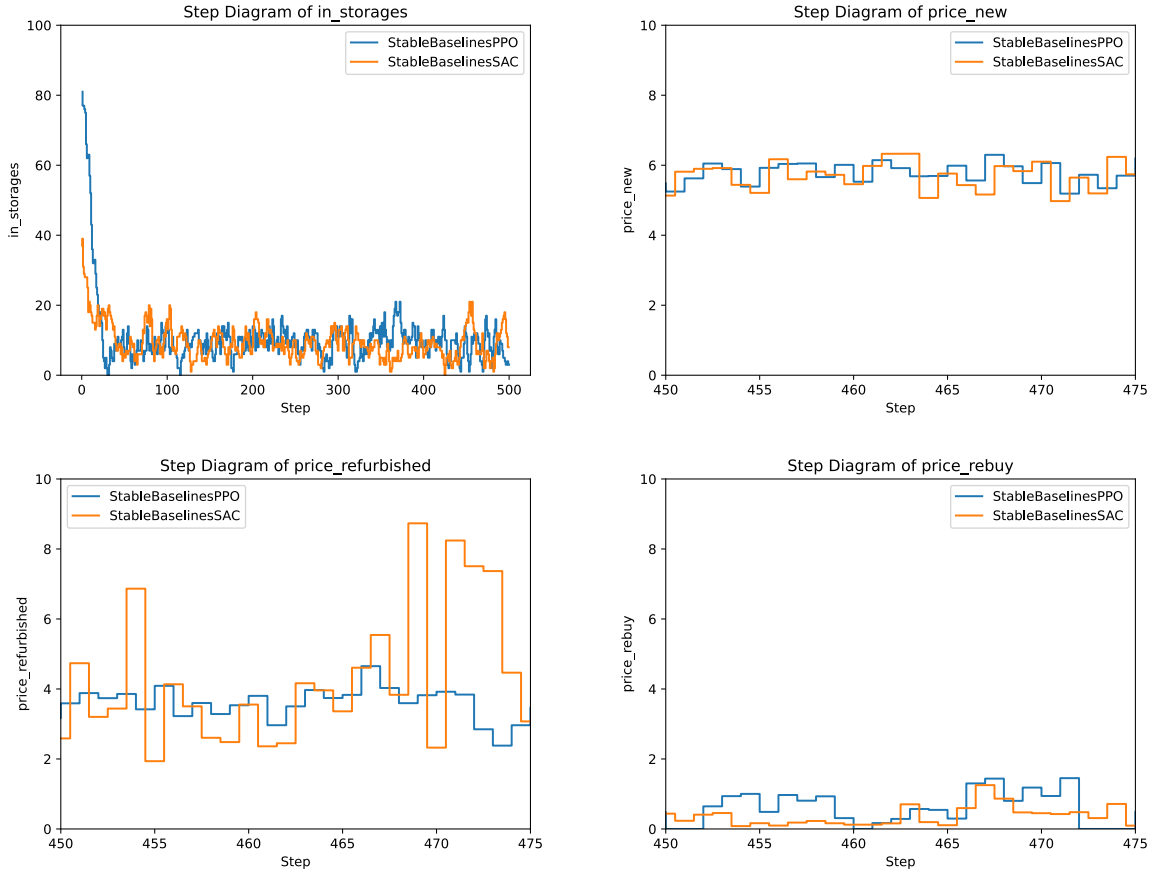

Figure 16: Details of an episode pitting a PPO agent (trained with self-play and mixed-reward capabilities) against an equally trained SAC agent.

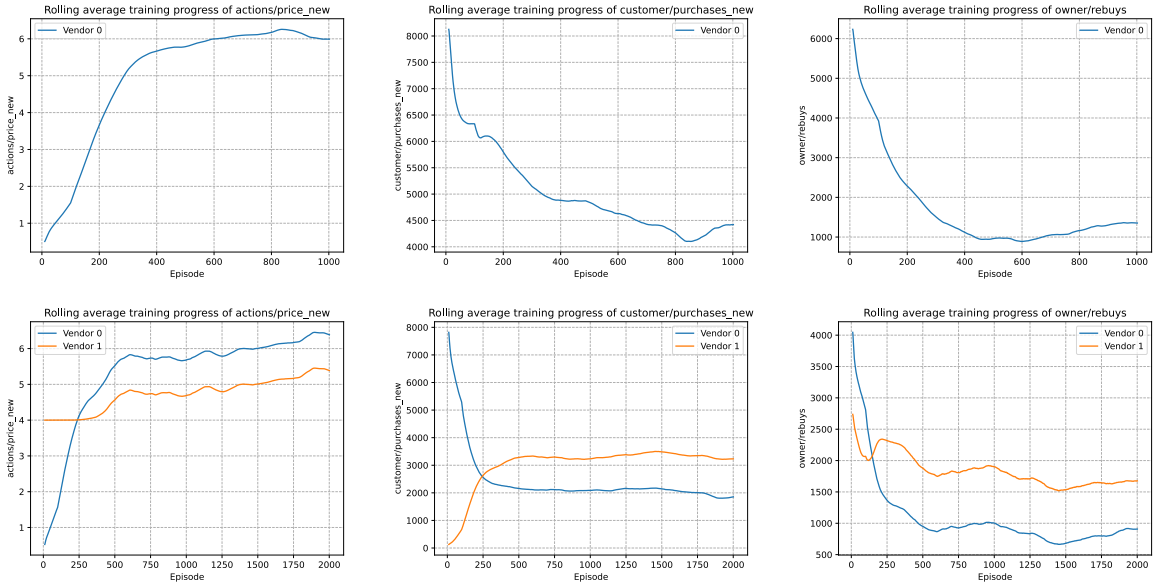

Figure 17: Measurements in PPO training runs on a monopoly (top) and a duopoly (bottom): The reason that the new prices set are not higher despite the monopoly is that customers' willingness to buy decreases as prices increase. At the same price level, the PPO agent in the monopoly sells twice as many new products, and activity in the rebuy market is also about twice as high.

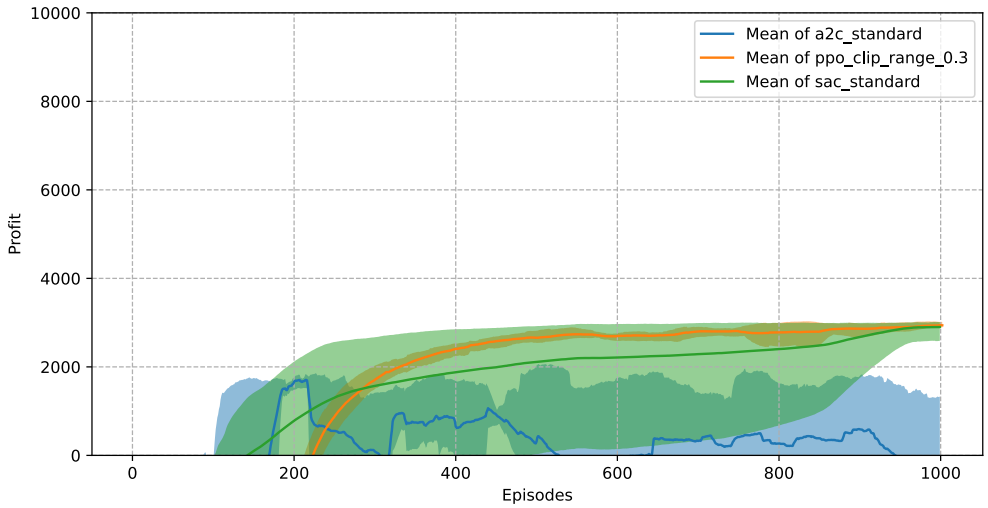

Figure 18: Learning curves of four A2C, PPO, and SAC agents in an oligopoly scenario over 1 000 episodes with mixed reward function and full observation.

Table 7: Ablation Study II: Steady state results for variations of other models parameter, cf. Table 2, with respect to our Base Case (given by $\gamma = 0.99$, $c_{virgin} = 3$, $c_{inv} = 0.1$, $B = 20$, $w = 0.05$, $\theta_{new} = 0.8$, $\theta_{used} = 0.5$, $\kappa_{used} = 0.55$, in a duopoly ($K = 2$) against RBB with $h = 1$, $M = 100$, cf. Section 4.1.3).

		Base	h		M		$M/8$=12.5		$M/15$=6.67		θ_{new}		θ_{used}		κ_{used}		c_{inv}
		Case	0.25	0.5	50	200	9.5	15.5	4.5	8.5	0.7	0.9	0.4	0.6	0.45	0.65	0.01
Offer Prices	$\bar{p}^{RL}_{new,offer}$	6.12	6.02	6.40	5.69	6.36	6.12	5.34	5.89	6.26	5.63	6.64	6.17	6.15	6.44	5.66	5.75
	$\bar{p}^{C}_{new,offer}$	5.12	5.05	5.39	4.68	5.35	5.12	4.39	4.89	5.27	4.65	5.67	5.16	5.16	5.42	4.66	4.76
	$\bar{p}^{RL}_{used,offer}$	3.92	4.11	4.06	3.81	4.40	3.67	4.16	4.40	3.73	3.80	3.52	3.21	4.46	3.76	3.57	3.70
	$\bar{p}^{C}_{used,offer}$	3.34	3.45	3.71	3.16	3.76	3.16	3.41	3.75	3.19	3.34	3.04	2.58	4.06	2.88	2.95	3.13
	$\bar{p}^{RL}_{rebuy,offer}$	0.23	0.11	0.15	0.14	0.20	0.06	0.06	0.01	0.06	0.42	0.29	0.00	0.52	0.05	0.37	0.30
	$\bar{p}^{C}_{rebuy,offer}$	0.72	0.66	0.83	0.64	0.64	0.88	0.49	0.58	0.85	0.78	0.76	0.67	0.85	0.44	0.71	0.68
Sales Prices	$\bar{p}^{RL}_{new,sold}$	6.01	5.94	6.27	5.63	6.31	6.05	5.26	5.85	6.18	5.50	6.54	6.10	6.13	6.35	5.60	5.68
	$\bar{p}^{C}_{new,sold}$	5.09	4.98	5.34	4.67	5.32	5.10	4.38	4.85	5.26	4.60	5.68	5.13	5.15	5.41	4.65	4.72
	$\bar{p}^{RL}_{used,sold}$	3.61	3.61	3.80	3.66	4.16	3.52	3.53	4.05	3.51	3.67	3.38	2.96	4.20	3.57	3.39	3.46
	$\bar{p}^{C}_{used,sold}$	3.07	3.24	3.51	2.93	3.38	2.90	2.97	3.58	2.94	3.16	2.94	2.30	3.79	2.73	2.82	2.91
	$\bar{p}^{RL}_{rebuy,sold}$	0.39	0.25	0.32	0.17	0.29	0.25	0.06	0.01	0.06	0.67	0.39	0.00	0.67	0.05	0.49	0.48
	$\bar{p}^{C}_{rebuy,sold}$	0.92	0.84	0.96	0.75	0.95	0.98	0.61	0.80	1.09	0.89	0.96	0.90	1.09	0.67	0.81	0.86
Sales	$\bar{X}^{RL}_{new}$	3.96	3.94	3.78	4.12	4.10	3.44	5.18	5.02	3.58	3.54	3.72	4.12	4.00	3.88	4.50	3.94
	$\bar{X}^{C}_{new}$	6.52	6.56	6.60	6.62	6.68	6.92	7.26	7.22	6.02	6.76	6.02	6.04	6.46	6.66	5.98	6.84
	$\bar{X}^{RL}_{used}$	1.72	1.60	1.62	1.98	1.22	1.96	1.30	1.34	2.14	2.24	2.36	1.62	2.10	1.68	2.58	2.06
	$\bar{X}^{C}_{used}$	3.63	3.50	3.22	3.62	2.78	3.84	2.60	2.96	4.30	3.20	3.48	4.12	3.18	3.12	3.90	3.74
	$\bar{X}^{RL}_{rebuy}$	1.74	1.56	1.42	1.86	1.46	1.90	1.50	1.28	1.88	2.14	2.32	1.94	2.10	1.84	2.48	2.04
	$\bar{X}^{C}_{rebuy}$	3.62	3.46	3.22	3.62	2.66	3.92	2.66	2.86	4.32	3.26	3.58	4.08	3.20	3.20	3.94	3.74
Resource Flows, Stocks & Rewards	$\bar{N}_{inuse}$	258	255	238	267	247	238	284	275	249	243	235	246	256	263	244	266
	$\bar{N}_{garbage}$	5.11	5.64	5.94	5.34	6.74	4.36	8.06	8.14	3.84	4.54	3.86	4.16	5.42	5.50	3.82	4.76
	$\bar{N}_{virgin}$	10.50	10.50	10.38	10.74	10.78	10.36	12.44	12.24	9.60	10.30	9.74	10.16	10.46	10.54	10.48	10.78
	$\bar{N}^{RL}_{stock}$	8.77	13.74	12.02	8.98	28.58	16.24	9.66	7.20	21.06	9.70	8.40	7.10	18.26	12.58	7.54	12.84
	$\bar{N}^{C}_{stock}$	8.35	8.90	7.02	9.22	8.46	7.56	10.14	7.48	9.98	7.68	7.98	8.86	7.70	9.40	8.78	8.24
	$\bar{G}^{RL}_{reward}$	15.60	15.59	16.88	16.50	15.32	15.30	15.21	18.19	16.71	14.27	17.41	16.81	18.12	17.64	17.69	15.43
	$\bar{G}^{C}_{reward}$	16.91	16.80	19.92	14.97	19.02	14.28	13.60	18.11	15.03	14.80	14.71	13.80	19.25	19.87	13.06	15.84