Karlsruhe Institute of Technology

Using Reinforcement Learning for Power Plant Dispatch in Short-term Electricity Markets

Final Presentation - Case Challenge - Business Data Analytics: Application and Tools

Students: Dominik Röhrle, Louis Skowronek, Louis Karsch, Ingo Hartmann, Leandra Fleck

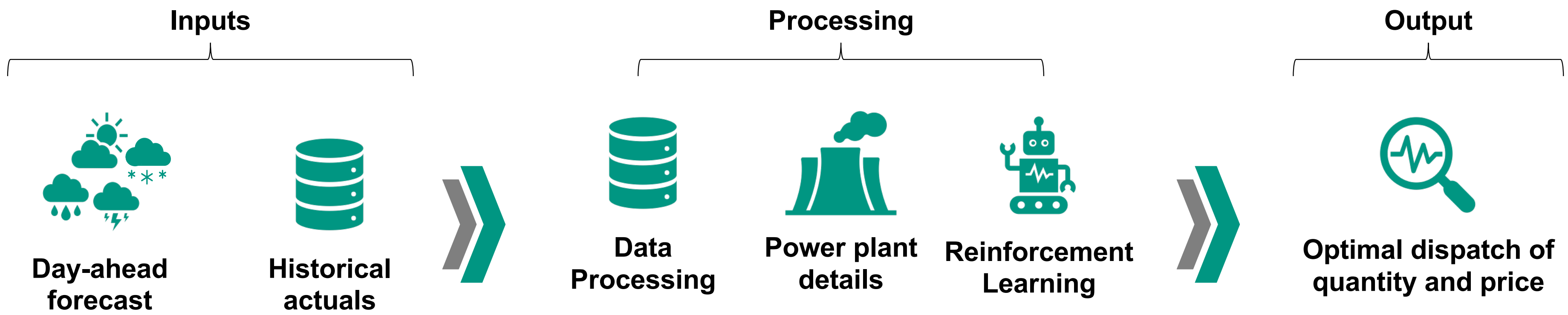
Source: https://www.wallpaperflare.com/electricity-pylon-electrical-grid-communications-tower-dividing-line-wallpaper-aosmj/download/1920x1080

Agenda

- Case Challenge & Inputs
- Algorithm Theory
- Implementation
- Results
- Outlook

Case: Using reinforcement learning we optimize the dispatch of a controllable power plant in the day-ahead market

Energy: Day-Ahead-Market



Reinforcement learning facilitates efficient, real-time decision making for power plant dispatch amid increased market volatility and uncertainty.

As inputs we build upon generation forecasts and production actuals of wind and solar power

Karlsruhe Institute of Technology

Excerpt from the dataset of generation actuals

Input Platform: ENTSO-E

Germany: Actual Generation per Production Type [MW]			
Time [hh:mm]	Wind		Solar
	Onshore	Offshore	
07:00 - 07:15	16,669	4,398	5,464
07:15 - 07:30	15,402	4,360	7,035
07:30 - 07:45	14,036	4,413	8,785
07:45 - 08:00	12,671	4,271	10,630
08:00 - 08:15	11,411	4,155	12,608
08:15 - 08:30	10,161	4,312	14,654
08:30 - 08:45	9,092	4,398	16,591
08:45 - 09:00	8,461	4,425	18,431
09:00 - 09:15	7,955	4,654	20,173
09:15 - 09:30	7,660	4,669	21,876
09:30 - 09:45	7,659	4,614	23,531
09:45 - 10:00	7,884	4,484	25,038
10:00 - 10:15	8,304	4,640	26,430
...	…	...	…

Excerpt from the dataset of generation forecasts

Input Platform: ENTSO-E

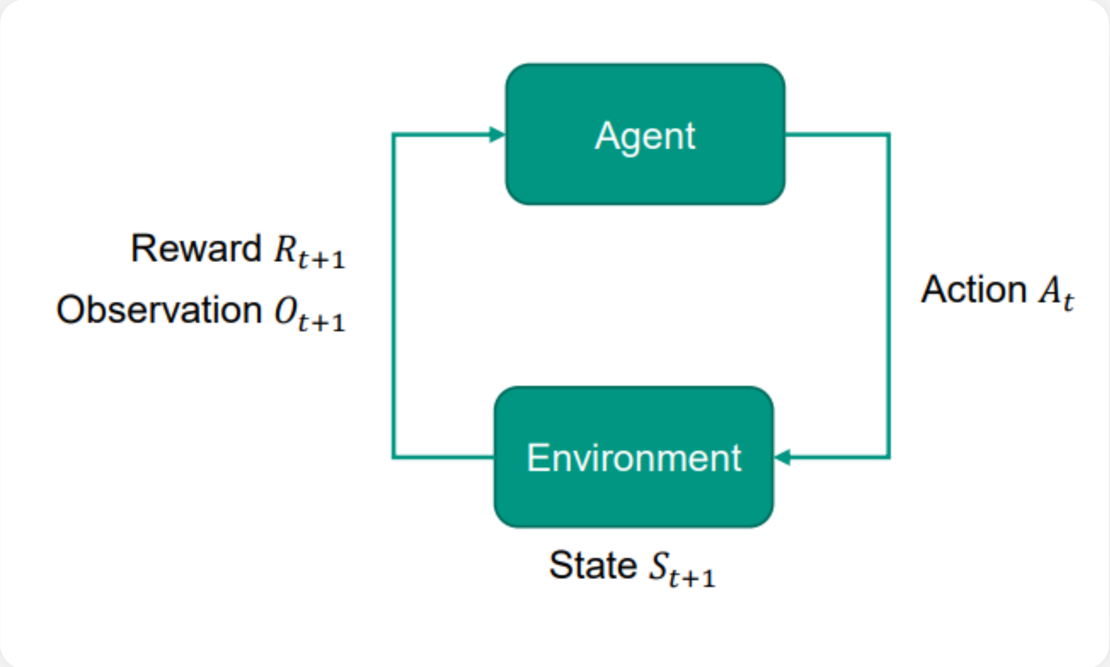
Germany: Day-ahead Generation Forecast [MW]			
Time [hh:mm]	Wind		Solar
	Onshore	Offshore	
07:00 - 07:15	18,590	3,964	5,159
07:15 - 07:30	17,815	3,910	6,750
07:30 - 07:45	17,002	3,853	8,551
07:45 - 08:00	16,191	3,945	10,479
08:00 - 08:15	15,046	3,889	12,534
08:15 - 08:30	13,835	3,846	14,707
08:30 - 08:45	12,610	3,946	16,935
08:45 - 09:00	11,570	3,882	19,095
09:00 - 09:15	10,809	3,756	21,168
09:15 - 09:30	10,324	3,737	23,175
09:30 - 09:45	9,867	3,867	25,043
09:45 - 10:00	9,450	3,834	26,736
10:00 - 10:15	9,169	3,787	28,227
...	…	...	…

Intended output variables

Dispatch quantity in MW

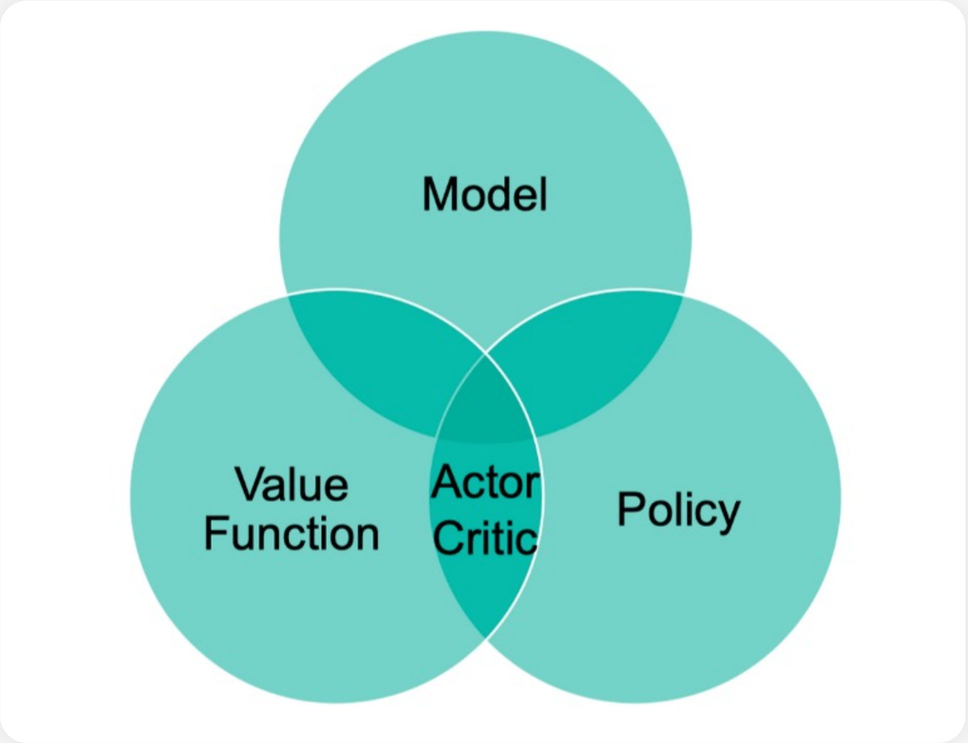
Selling price per MW

Theory: Basics Reinforcement Learning



Source: Prof. Dr. J. M. Zöllner – Maschinelles Lernen I – Grundverfahren

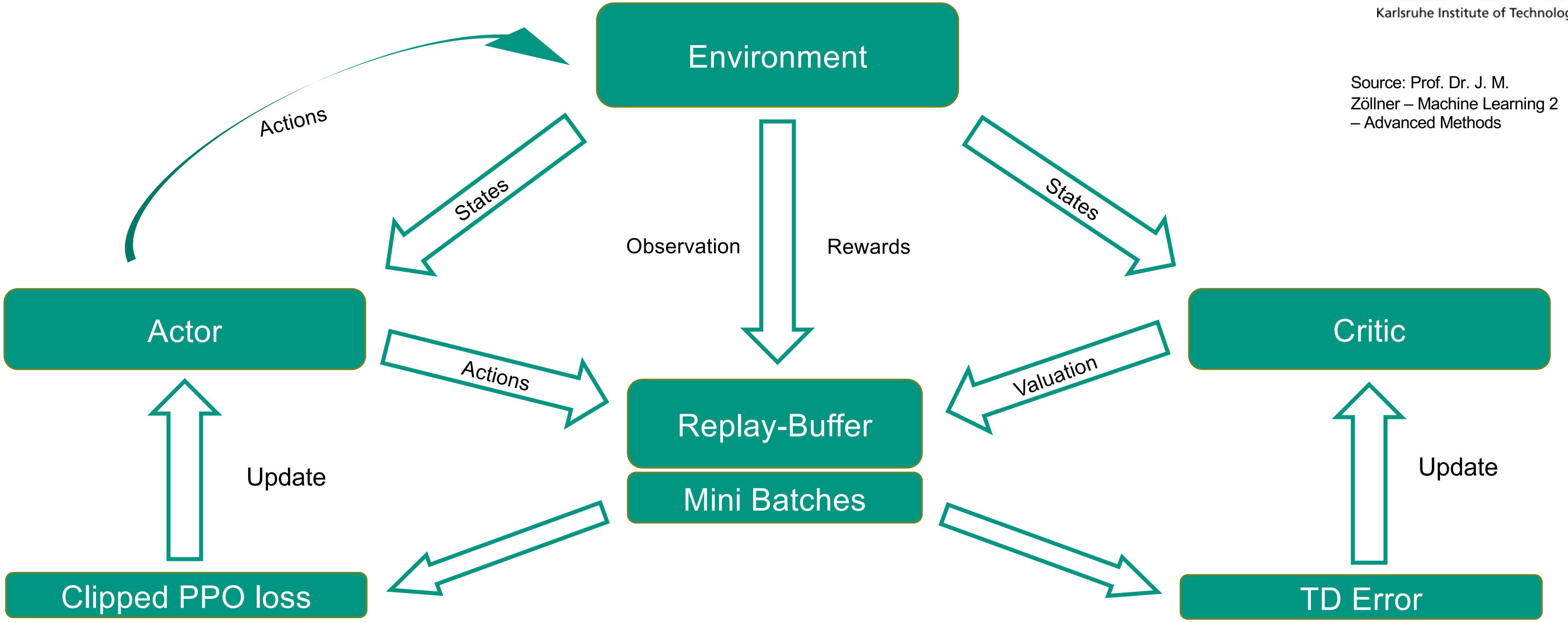
Theory: Taxonomy of Reinforcement Learning



Source: Prof. Dr. J. M. Zöllner – Maschinelles Lernen I – Grundverfahren

PPO Actor-Critic Implementation

Source: Prof. Dr. J. M. Zöllner – Machine Learning 2 – Advanced Methods



Update Critic ϕ

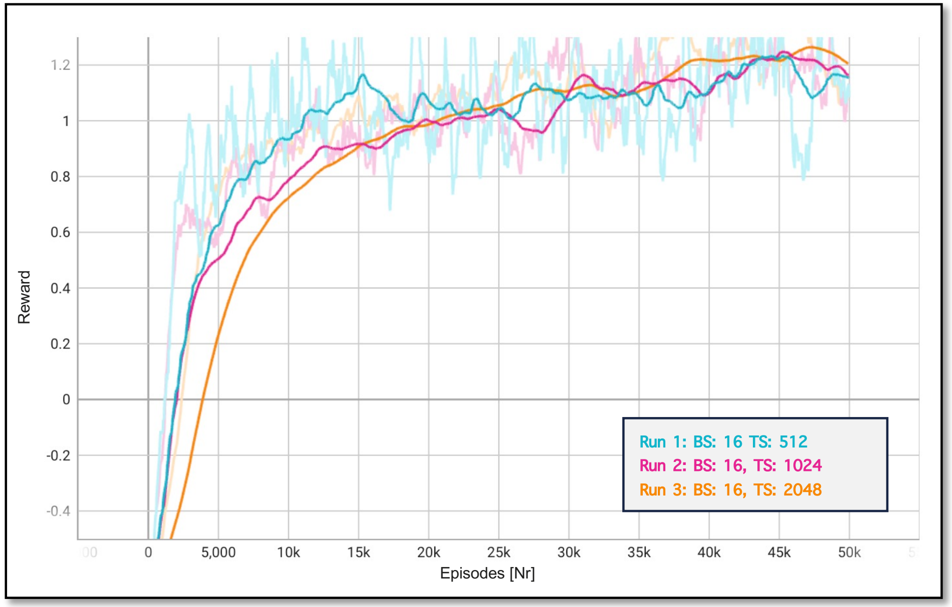
$$\phi \leftarrow \phi - \alpha_c \frac{1}{|D|} \sum_{\tau \in D} \sum_{t=0}^{T} \frac{1}{2} \underbrace{\left\| V_\phi(s_t) - \left(\underbrace{R(s_t, a_t) + V_{\phi'}(s_{t+1})}_{\text{Target}} \right) \right\|^2}_{\text{TD-Error}}$$

- **PPO v2:**
  - Optimize the problem:

$$\max_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_{old}}} \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_{old}(s_t, a_t), clip\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) \hat{A}_{old}(s_t, a_t) \right) \right]$$

Results: Fine-tuning PPO via Grid Search

Average Reward After 50 000 Episodes for 3 Example Runs



Considered Hyperparameters

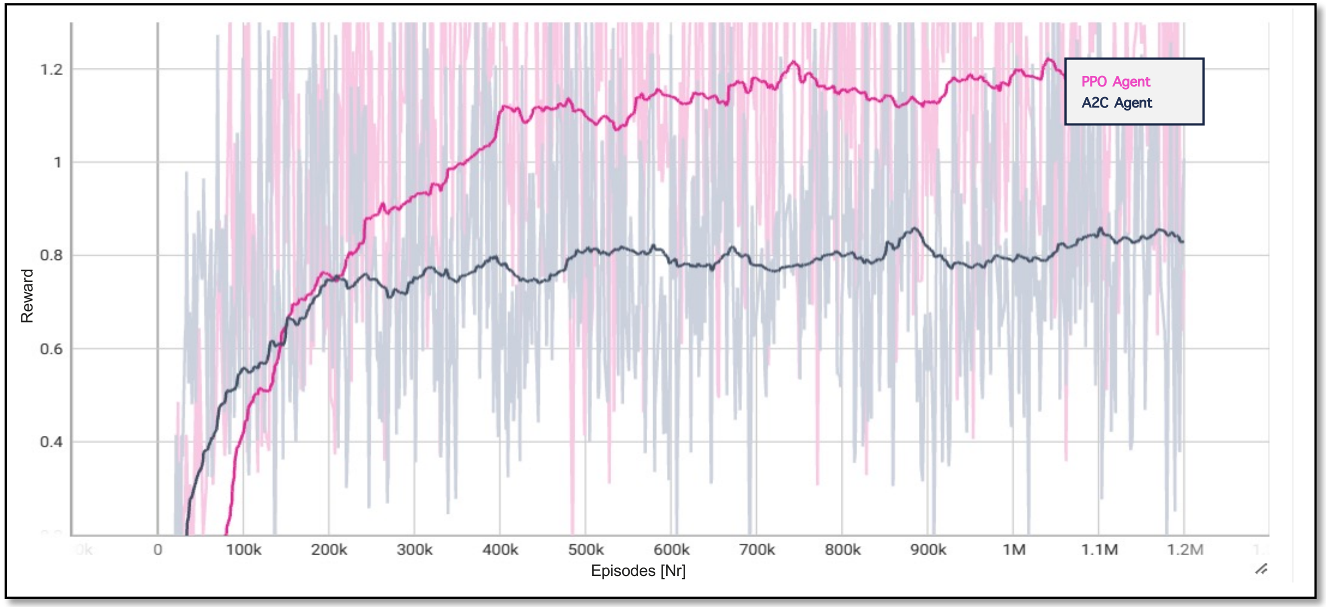
- Batch Size: <u>16</u>, 32, 64, 128
- Update Time steps: 512, 1024, <u>2048</u>

→ could be expanded to test more hyperparameters, but costly

Hyperparameters affect the convergence speed and stability of training. In our case higher time steps lead to lower convergence speed but higher stability of training.

Results: Comparison of PPO vs A2C (stable baselines)

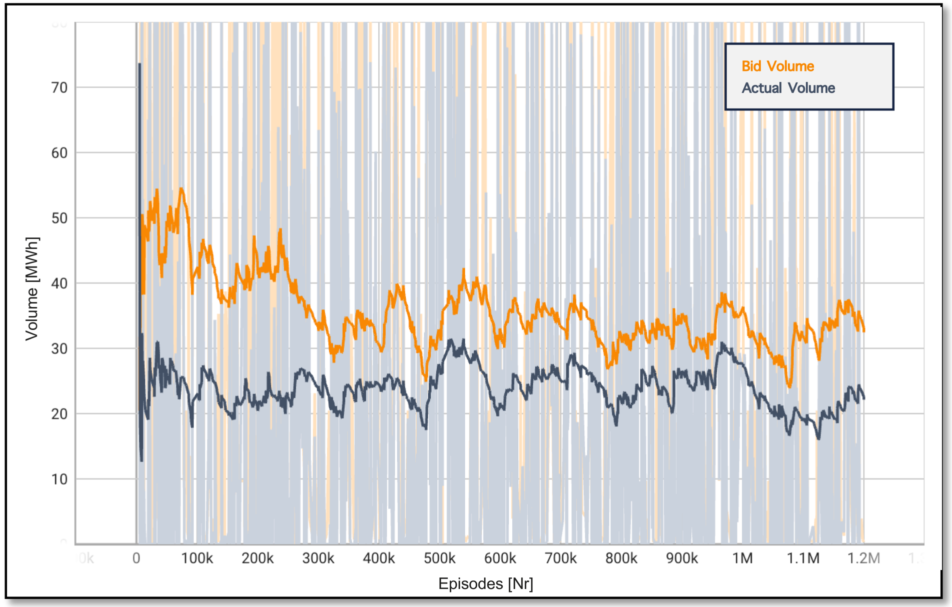
Average Reward after 50 000 Episodes



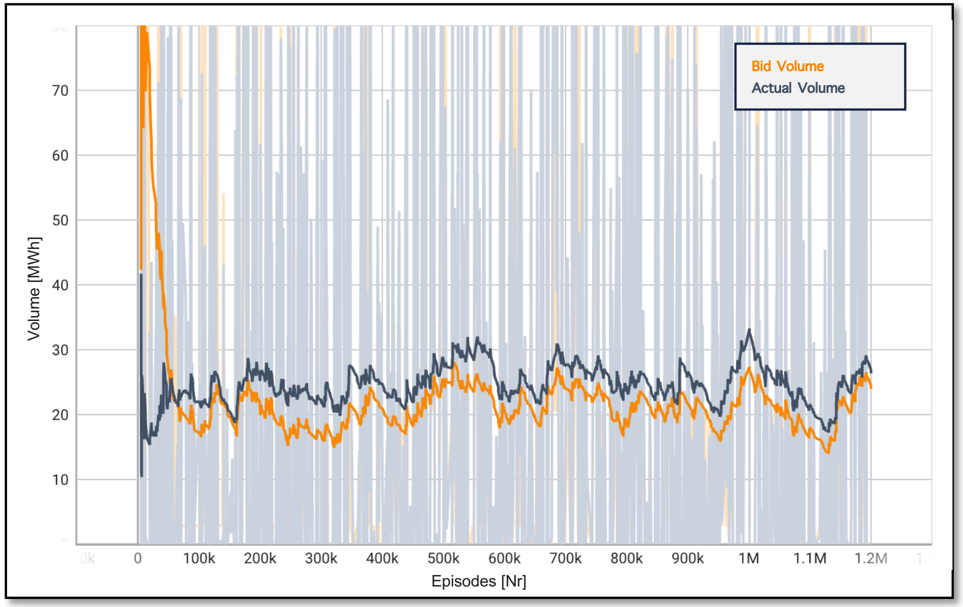
PPO achieves a significantly higher average reward after learning more slowly as a more zurückhaltender algorithm.

Results: Reward functions using PPO - The reward function strongly impacts the RL-Agent's bidding behaviour

Bid Volume – Linear Scaled Reward Function



Bid Volume – Logarithmic Scaled Reward Function



Therefore, reward engineering is a crucial task when setting up optimal agents.

Further improvements of the agent can be achieved through enhancement of data inputs, the environment and the agent itself

Karlsruhe Institute of Technology

1

Data Inputs
- ➢ Incorporate intra-day prices
- ➢ Additional features (e.g., generation failures)

2

Environment
- ➢ Variable marginal costs (currently fixed)
- ➢ Powerplant-related restrictions (e.g., energy storage, maintenance)

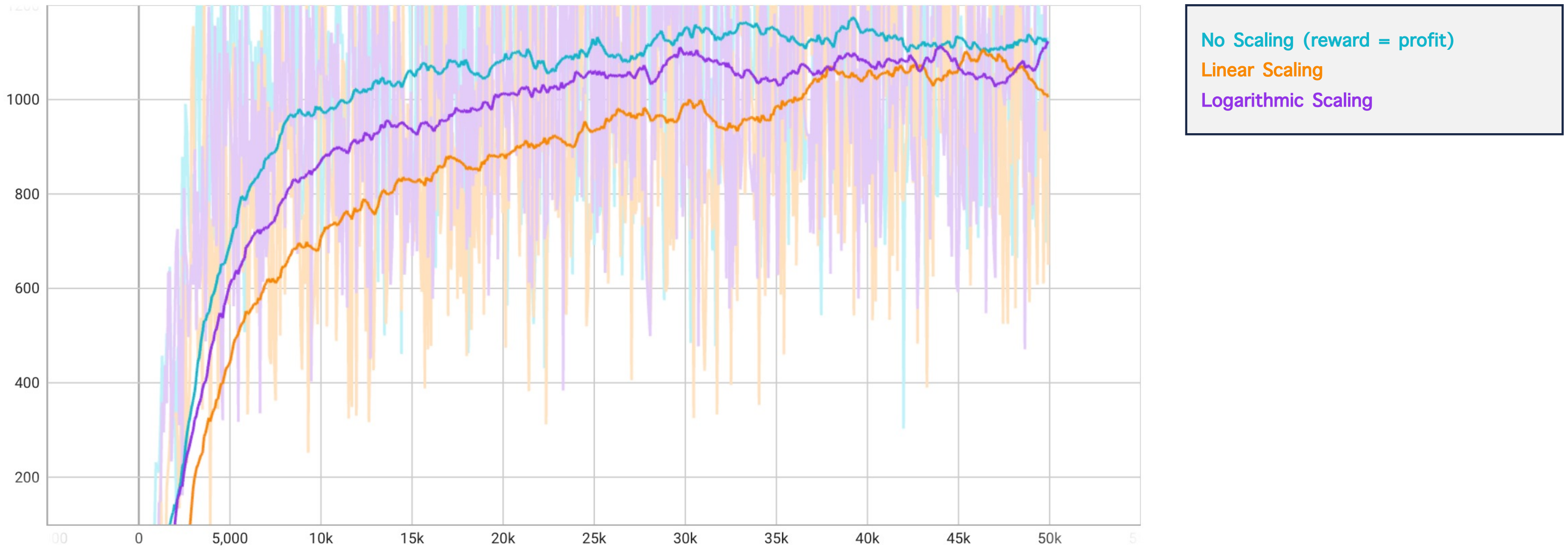
3

Agent
- ➢ Expand hyperparameter tuning
- ➢ Higher dimensional discrete action space (e.g., more than 50x50)
- ➢ Transition to continous action space

Backup

Profit of PPO using different Reward Functions

Average Profit after 50 000 Episodes

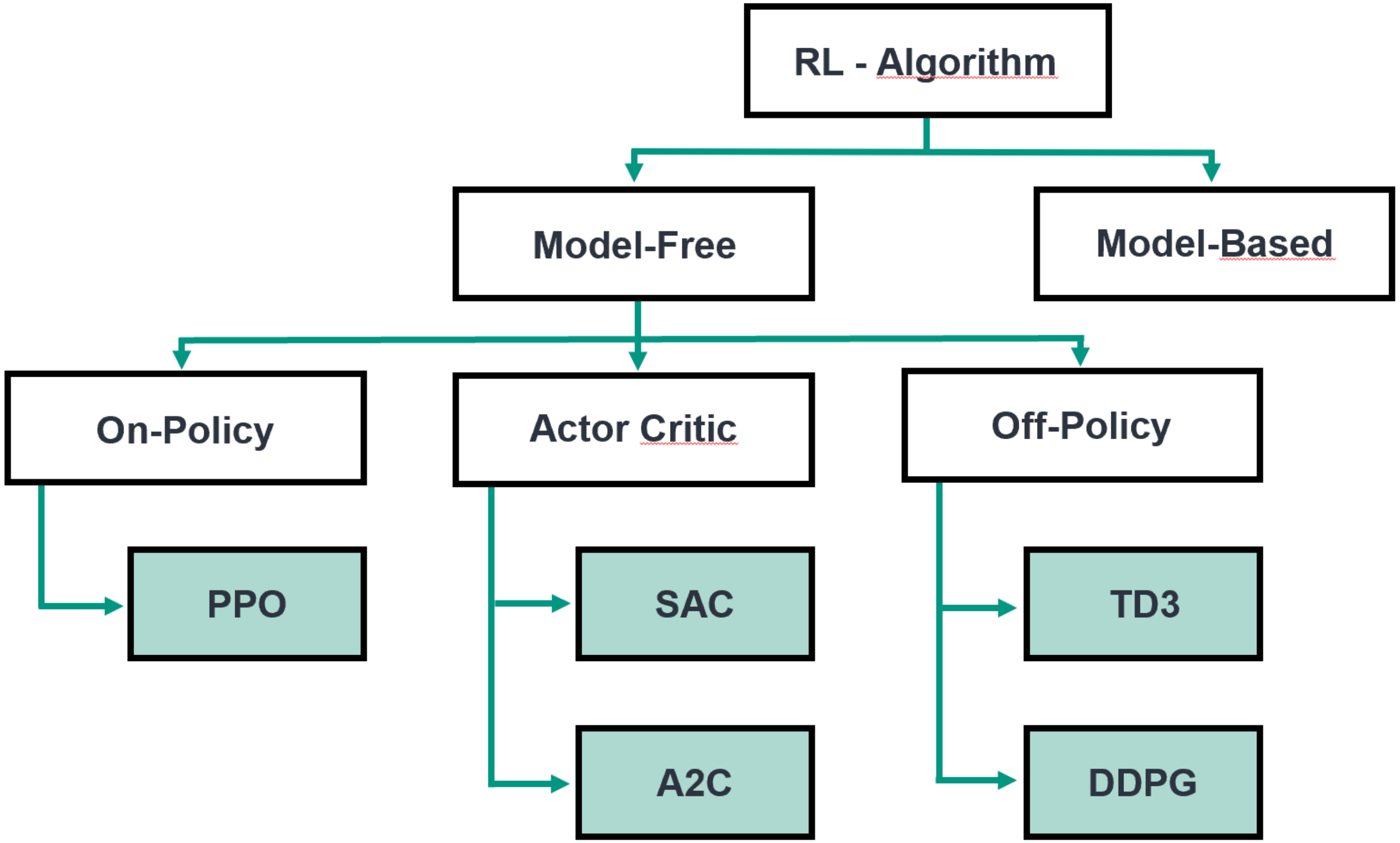


Bid Price converges to the marginal costs of the simulated Power Plant (mc=50)

Hyperparameter

- lower_bound: -20000
- upper_bound: 20000
- batch_size: [16, 32, 64, 128]
- n_episodes: 50000
- update_timestep: [512, 1024, 2048]
- n_epochs: 10
- eps_clip: 0.22
- gamma: 0.99
- lr_actor: 0.0002
- lr_critic: 0.0008

Theory: Overview RL-Algorithm



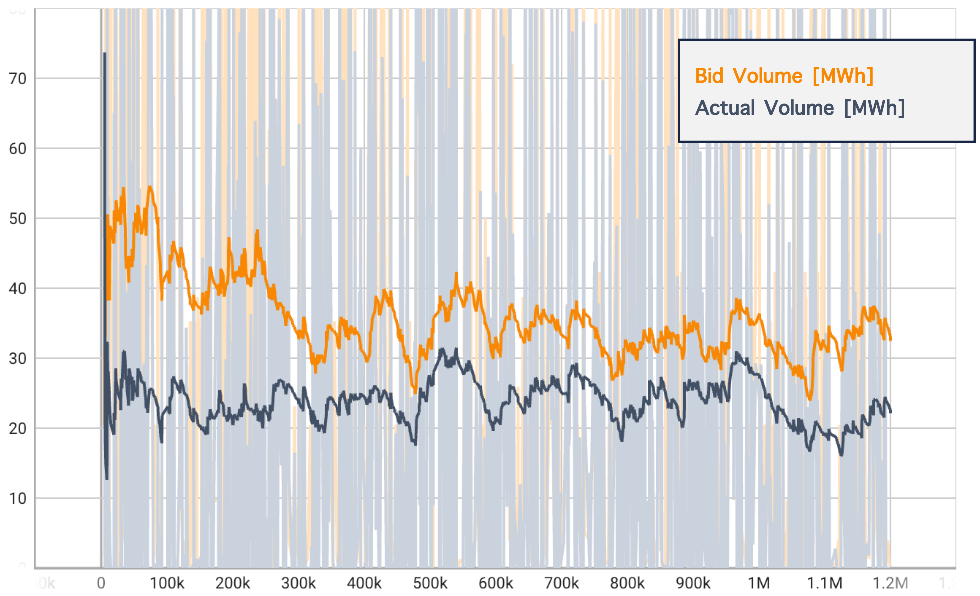
Theory: Value-based and Policy-based

$$\theta \leftarrow \theta + \alpha * \nabla\theta \log \pi(a|s) * A(s, a)$$

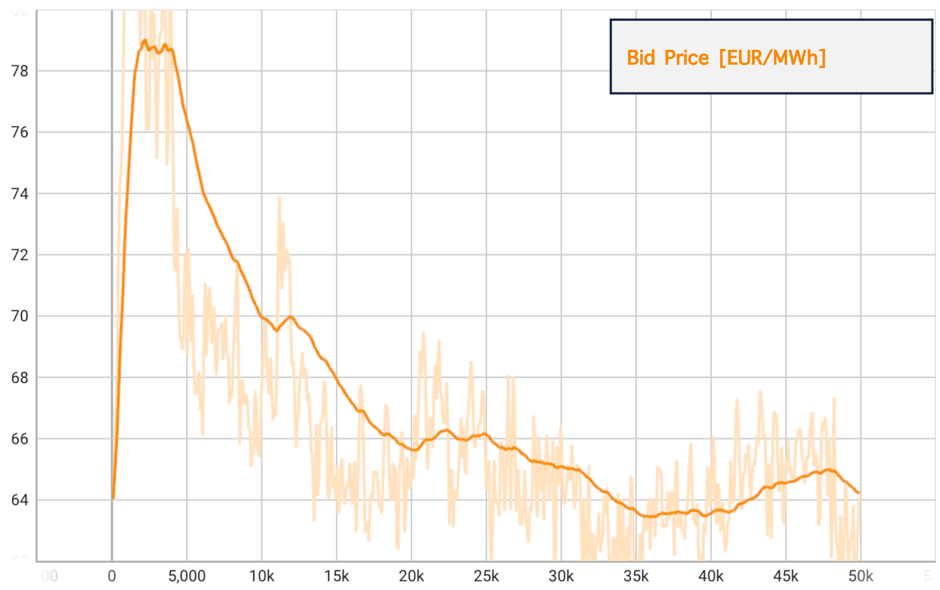
-θ represents the parameters of the policy π(a|s) that are updated
- α is the learning rate that determines the step size of the parameter updates
- ∇θ log
π(a|s) is the gradient of the logarithm of the policy with respect to the parameters
- A(s,a) is the advantage function that estimates the advantage of taking action a in state s compared to the expected value

Bid Volume and Bid Price of PPO with Linear Scaled Reward Function

Bid Volume



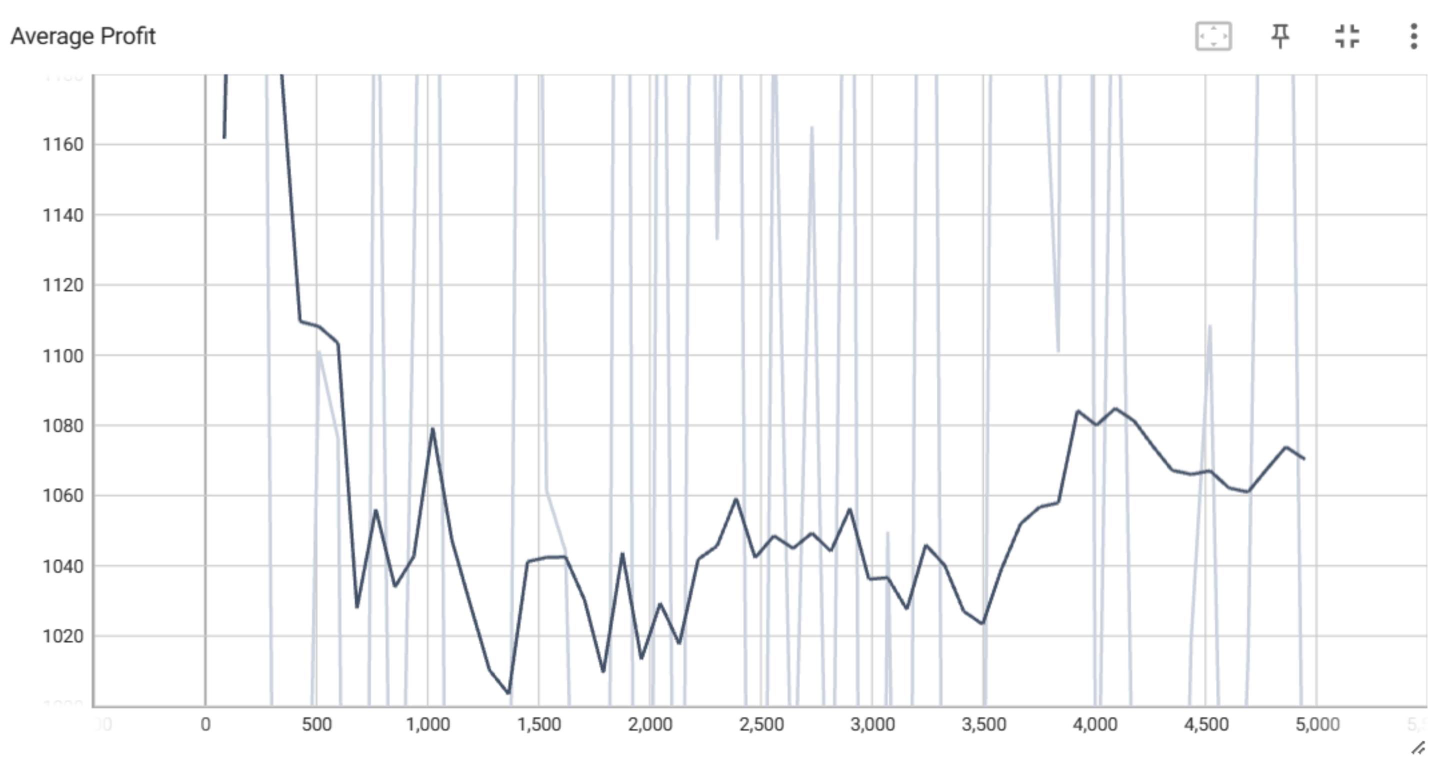
Bid Price



Bid Price converges to the marginal costs of the simulated Power Plant (mc=50)

Backup Folie

- Plotting the Average Profit of the trained model on unseen data



Backup Folie (PPO vs A2C) stable baselines