Reinforcement learning-based control system for biogas plants in laboratory scale

Alberto Meola[1,2], Oliver Kiefner[1,2] , Félix Delory[1] and Sören Weinrich[1,3]

1 - DBFZ, Deutsches Biomasseforschungszentrum gemeinnützige GmbH, Biochemical Conversion Department, Torgauer Straße 116, Leipzig, 04347, Germany

2 - Leipzig University, Faculty of Mathematics and Computer Science, Augustusplatz 10, Leipzig, 04109, Germany

3 - Münster University of Applied Sciences, Faculty of Energy · Building Services · Environmental Engineering, Stegerwaldstraße 39, Steinfurt, 48565, Germany

Abstract. Reinforcement learning techniques can be used to learn effective policies for complex tasks, but they are rarely applied for control of biogas plants. While control of the anaerobic process is necessary for optimal plant operation, process complexity and instability prevent the usage of advanced control mechanisms in industrial settings. In this study, a proximal-policy optimization algorithm has been applied on the feeding schedule of a lab-scale biogas reactor for biomethane conversion to electrical energy depending on dynamic energy prices. The algorithm effectively optimizes feeding and selling strategies, outperforming traditional methods.

1 Introduction

Agricultural biogas plants in Germany are generally used for base load power supply, but their profitability without state subsidies is uncertain, and the need for demand-oriented electricity requires more advanced control techniques. Within biogas plants, organic material is converted into biogas through the Anaerobic Digestion (AD) process. The produced biogas can subsequently be converted into electricity, which is fed into the power grid. Generally, the feeding pattern of biogas plants can be altered to provide demand-oriented power at high electricity prices [1]. However, due to highly dynamic process operation, stability concerns of the AD process arise. Machine Learning (ML) techniques are being applied for process simulations [2], yet their application in control systems remains limited. Reinforcement Learning (RL) algorithms could be able to effectively control biogas reactors for on-demand energy production, but the research in this topic is limited [3]. This study demonstrates the application of the Proximal Policy Optimisation (PPO) algorithm to a laboratory-scale biogas plant for increased profitability. The agent is initially trained on the semi-mechanistic ADM1-R3 model [4], and then tested in laboratory scale in two configurations for four weeks. The models were extensively trained to ensure policy stability and convergence. Additionally, a third configuration is introduced to better mimic industrial conditions. The applied PPO optimizes the feeding schedule and the timing of electricity provision.

2 Materials and methods

Within the developed framework, the agent is first recurrently trained on the ADM1-R3 model, receiving the biogas production data resulting from the RL agent. In a second step, the combinations of actions resulting in the highest reward are extracted, and in a third step they are applied to a lab-scale reactor. In a fourth step, theoretical revenue from the sales of the gas is calculated, and in the fifth and last step all the required data is presented to the agent for further training. A schematic representation of this process is presented in Figure 1.

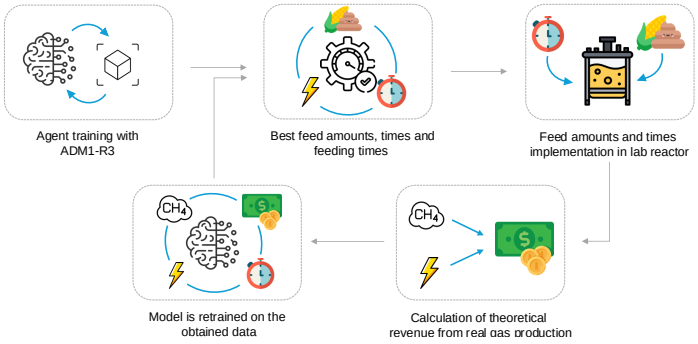

Fig. 1: Control System

2.1 Data availability

The ADM1-R3 model, applied in the first step of the control process, is a simplified version of the ADM1 model, a semi-mechanistic model that describes the AD process [5]. This simplification enables faster simulation times.

2.1.1 Experimental setting

The actions set by the RL agent are applied by the laboratory personnel in a 12 l continuous stirred-tank AD reactor equipped with biomethane production sensors. The substrate mix includes corn silage and cow manure with 0.19 g g^{-1} TS and 0.9 g g^{-1} VS. The biomethane produced from the reactor is then applied in the RL environment as a 1 h resolution time serie. For comparison purposes, a parallel reactor is operated in a naïve Scenario, with fixed feeding time (8 AM) and amounts (4.5 kg VS m^{-3} per day) and constant conversion of biogas and sales of electricity.

2.2 Control Mechanism

For the establishment of a control mechanism, a RL algorithm - a constrained PPO - was developed. Within the PPO framework, an agent takes actions to maximize cumulative rewards $L(\theta)$, as parameterized by policy π_θ. Moreover, a

clip function on $r_t(\theta)$ is applied to retain the policy update within a conservative range and avoid destabilizations to the learning process. The overall goal of the agent is to maximize the cumulative reward by interacting with the environment through a parameterized policy [6].

2.3 Environment configurations

Three environment configurations with increasing potential for application in full-scale biogas reactors were tested. The configurations with their characteristics and performances are displayed in Table 1. In all environments, the agent can feed the reactor seven times a week, while the limitations of the selling actions differ depending on the environment. A 1-week actions set consists of seven floats for feed quantity, feed timing and daily selling patterns per week. The agent is limited in the feed amount - an average of 4.5 kg VS m^{-3} per week, 7 kg VS m^{-3} per day - and in the feeding time - from 8 AM to 3 PM during the week and from 9 AM to 12 PM. Feeding time limitations are implemented for allowing laboratory operations. Moreover, it is assumed that the spot market electricity prices are known one week in advance. A schematic visualization of the environment is presented in Fig 2.

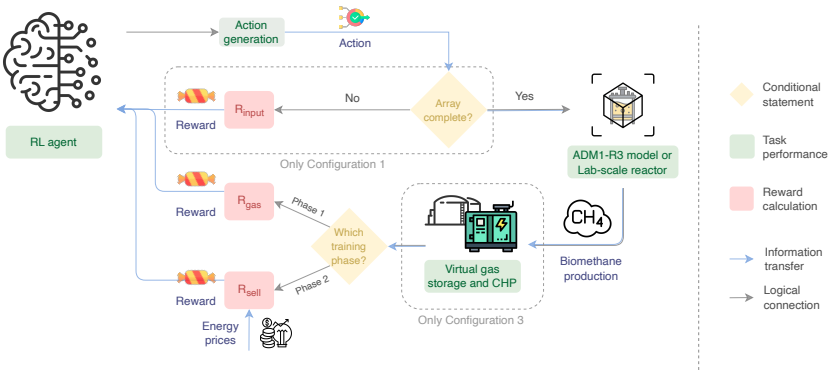

Fig. 2: RL Environment structure

The training process of the agent is divided in two phases. In the first phase, the agent is rewarded with R_{gas}, calculated as the sum of biomethane produced in the present week. During the second phase, the agent is rewarded with R_{sell}, calculated as the revenue difference between the RL agent revenue and the naïve agent revenue.

2.3.1 Configuration A and Configuration B

Configurations A and B assume unlimited gas storage and motor power. Thus, all gas generated is virtually stored, and once the agent defines a selling time, all the available gas is converted instantly into electricity and sold to the market. Within Configuration A, the agent performs 1 action per timestep, indicating

for the first 7 actions the feeding quantities and for actions from 7 to 14 the feeding times. At each timestep, a reward R_{input} of 1 is set until completion of the first 14 actions. From the 15^{th} action, R_{gas} is calculated only for the days of the week where the selling patterns are defined. Instead, within Configuration B, the agent performs 21 actions per timestep, not requiring the reward R_{input}. The daily selling patterns are defined by remapping each agentâs actions from the 14th to the 21st into a list of 24 binary values, representing the operation of the Combined Heat and Power (CHP) unit, which converts biomethane into electricity.

2.3.2 Configuration C

In Configuration C, a gas storage and motor is simulated for realistic reproduction of industrial conditions. The dimensioning of the gas storage and the motor are extracted from [7]. The gas exceeding the gas storage capacity is virtually vented. An efficiency of 42% is considered for the CHP. The selling patterns are defined as in Configurations A and B, but excluding the combinations where the CHP gets activated more than three times a day, as in industrial scale biogas reactors. Configuration C is the only one not yet applied in lab-scale.

3 Results and discussion

3.1 Simulation and training results

PPO's Learning Rate (LR) and γ were manually tuned - with the LR being 0.0007 and γ being 0.99, while the remaining hyperparameters were left as standard settings. In Fig. 3 it is shown the reward progression for the three scenarios.

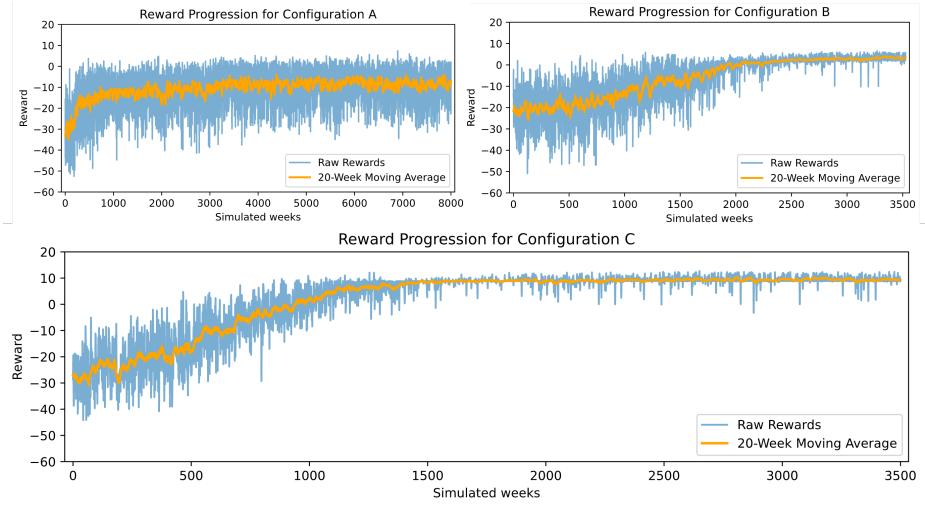

Fig. 3: Average reward and steps per run in Phase 1

Configuration A showed minimal improvement after 2000 weeks, plateauing around -10, whereas Configurations B and C steadily improved to positive rewards. Configuration C converges faster than Configuration B, probably due to the lower amount of the available selling patterns in the latter configuration.

3.2 Lab-scale control results

The applied scenarios in both lab-scale and simulative scenarios with resulting revenue results are shown in Table 1.

Scenario Name	**1**	**2**	**3**	**4**	**5**
Configuration type	A	A	B	B	C
Simulated scenario revenue increase	11,2%	5,79%	9,19%	9,02%	18,72%
Lab-scale scenario revenue increase	1,28%	1,74%	1,58%	7,72%	-
Trained on simulated data	Yes	Yes	Yes	Yes	Yes
Trained on one week lab-scale data	No	Yes	No	Yes	No

Table 1: Summary of scenarios and configurations.

An overview of the revenue obtained in the different scenarios by the RL agent and by the naïve agent are shown in Fig. 4.

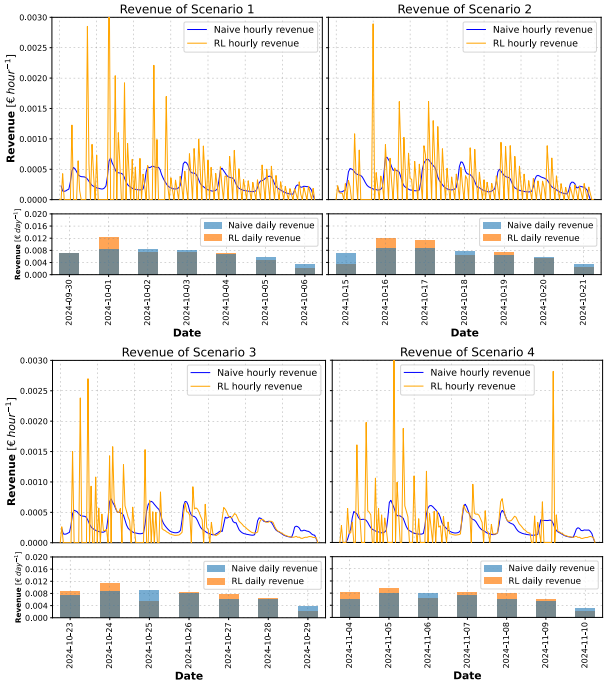

Fig. 4: Hourly and daily revenue of naïve and RL agent.

The RL agent overperforms the naïve agent over the applied scenarios, and it is able to identify different selling patterns depending on the scenario, confirming the efficacy of the learned policies. Moreover, the impact of the selling time - especially in Scenarios 3 and 4 - is lower than the impact of the feeding process, since the agent sells electricity at a constant rate in several periods of the week. In general, the RL agent adapts dynamically to price fluctuations, potentially allowing demand-based control in real time. Although immediate revenue improvements vary, the adaptability of the RL agent could offer long-term advantages under changing market conditions.

4 Conclusion

A constrained PPO algorithm applied to biogas production and electricity generation was successfully tested and applied on a lab-scale reactor. Both applied configurations demonstrated resiliency to real-data training, but the configurations using multidimensional actions - Configurations B and C - shows potential for long-term training and control on biogas reactors due to the more stable training process. While the improvement in performances varied between 1.28% and 7.72%, the more realistic the environment got, the higher the performances were, with potential up to 18.72% based on more restrictive gas storage criteria. To assess the efficacy of such methods in industrial applications, further research should consider longer control operations in full-scale.

References

[1] Ohnmacht, B. Lemmer, A. Oechsner, H. Kress, P. Demand-oriented biogas production and biogas storage in digestate by flexibly feeding a full-scale biogas plant, Bioresource Technology, 332, 125099, 2021.

[2] Ling, J.Y.X. Chan, Y.J. Chen, J.W. Chong, D.J.S. Tan, A.L.L. Arumugasamy, S.K. Lau, P.L. Machine learning methods for the modelling and optimisation of biogas production from anaerobic digestion: a review. Environ. Sci. Pollut. Res. 2024

[3] Pettigrew, L. Delgado, A. Neural Network Based Reinforcement Learning Control for Increased Methane Production in an Anaerobic Digestion System, 3rd IWA Specialized International Conference "Ecotechnologies for Wastewater Treatment". 2016.

[4] Weinrich, S. Nelles, M. Systematic simplification of the Anaerobic Digestion Model No. 1 (ADM1) - Model development and stoichiometric analysis, Bioresource Technology, Volume 333, 2021, 125124. 2021.

[5] Batstone, D. Keller, J. Angelidaki, I. Kalyuzhnyi, S. Pavlostathis, S. Rozzi, A. Sanders, W. Siegrist, H. Vavilin, V. Anaerobic digestion model No 1 (ADM1). Water science and technology. 45. 65-73. 2002.

[6] Schulman, J. Wolski, F. Dhariwal, P. Radford, A. Klimov, O. Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347. 2017.

[7] Mauky, E. Weinrich, S. Jacobi, H. Nägele, H. Liebetrau, J. Nelles, M. Demand-driven biogas production by flexible feeding in full-scale - Process stability and flexibility potentials, Anaerobe. 46. 86-95. 2017.