DEEP DETERMINISTIC POLICY GRADIENT ALGORITHM FOR BATCH PROCESS CONTROL

Haeun Yoo, Boeun Kim, Jay H. Lee*
Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea

Abstract Overview

In general, a batch process shows highly nonlinear dynamic behavior and absence of a steady state, leading to difficulties in its operation and control. Nonlinear model predictive control (NMPC) is widely used to address this problem but performance cannot be guaranteed when the plant model is not accurate or significant uncertainties exist. Reinforcement learning is a model-free data-driven decision making strategy which has shown notable success in games and other similar areas. A key advantage over MPC is that it can handle stochastic uncertainty in an active and optimal manner. In this research, we propose a reinforcement learning based control strategy for batch processes under parameter uncertainty. Q function reflects the performance of the process by receiving penalty or reward for violating or satisfying the path and end-point constraints of the process. Deep deterministic policy gradient (DDPG) algorithm is adopted to accommodate high-dimensional continuous state and action space. Performance of the proposed control strategy is evaluated through an example of an anionic propylene oxide polymerization process. Standard NMPC is used as the initial controller, of which the performance is improved by the proposed RL method. Their performances are compared when an uncertain kinetic parameter is perturbed.

Keywords

Batch process control, reinforcement learning, DDPG, Q function, Actor-Critic, Polymerization

Introduction

In industry, batch reactors are commonly used for producing low-volume, high value-added products such as high-end polymers. However, owing to its inherent nonstationary characteristics, nonlinear dynamics, and time-varying operational constraints, its operation and control are particularly challenging. In addition to the nonstationary and nonlinear characteristics, significant feedstock and process variabilities can exist from batch to batch. Therefore, for successful batch operation, controller should effectively compensate for the undesired effects of uncertainties on the process performance (Bonvin et al., 2001).

Nonlinear model predictive control (NMPC) determines optimal control moves on-line by directly solving an optimization problem formulated with a nonlinear model, path and end-point constraints, and reference trajectories sent down from a high-level optimization. Instead of following predetermined optimal trajectories, it can also optimize a profit function directly at the control level, which is known as economic NMPC. NMPC controllers are known to give excellent performance when the model used is accurate and the influence of uncertainty on the process performance is minor. To handle the problems arising from model-plant mismatch, various robust MPC strategies have been suggested as extensions (Morari and Lee, 1999), but they also require prior knowledge about the uncertainty, which often is not available.

Meanwhile, reinforcement learning (RL), a model-free data-driven decision making strategy, has shown some remarkable performance in game playing, such as Alpha Go in the game of "Go" (Silver et al., 2017, 2016), and has been used in various other areas such as robotics and scheduling problems (Kober et al., 2013; Shin et al., 2019;

* jayhlee@kaist.ac.kr

Sutton and Barto, 1998). In the process system engineering area, RL has potential to complement the weaknesses of model predictive control, esp. in handling highly stochastic or more generally uncertain systems.

In this research, we suggest a RL based control strategy for batch processes, with the aim of deriving an optimal control policy or an improved control policy starting with the NMPC controller. Unlike in other applications, satisfying end quality constraints is of utmost importance in batch processes. The proposed RL based controller learns a Q function that reflects the satisfaction of constraints as well as the performance index. The Q function and policy are learned through deep deterministic policy gradient (DDPG), which can handle continuous and high-dimensional state and action space which are common for chemical processes.

DDPG algorithm for batch process control

For high dimensional processes, an optimal policy can be obtained by approximately solving the Bellman's optimality equation using RL algorithms (Sutton and Barto, 1998). In RL, an action-value function, called the Q function, is widely used to express the expectation of rewards with respect to state and action. We propose to use the Q function based learning for batch process control is suggested. Batch process can be considered as a game, i.e., one batch is one round of a game. Agent gets a reward if constraints are satisfied, otherwise gets a penalty. The reward can be related to the performance index to encourage achieving a high performance index value at the end. The Q function evaluation scheme is illustrated in Table 1. $p(s_t)$ denotes the penalty or reward determined by the degree to which the path constraints are satisfied.

Table 1. Q function evaluation criteria

$Q(s_t, a_t)$ evaluation criteria
If a_t is a final action:
If s_{t+1} satisfies all end-point constraints:
$Q(s_t, a_t) = r_t \leftarrow +Performance\ Index\ (reward)$
Else:
$Q(s_t, a_t) = r_t \leftarrow -1\ (penalty)$
Else:
$Q(s_t, a_t) = r_t + \gamma \max_{a_{t+1} \in \Omega} Q(s_{t+1}, a_{t+1})$,
$r_t \leftarrow p(s_t)$

In this study, a DDPG algorithm is used to learn the Q function and the corresponding policy for batch process control. DDPG is a model-free, off-policy, actor-critic algorithm based on a deep deterministic policy (DPG) with deep neural networks. This algorithm are known to be capable of learning policies in high-dimensional, continuous action spaces, which are common for chemical process control problems (Lillicrap et al., 2016). The actor provides a policy for a given state and the critic predicts the Q function value. The actor and critic networks are updated to minimize a TD (temporal-difference) error. Pseudo code of the DDPG algorithm for batch process control is represented in Table 2, which modifies the description of the algorithm given by Lillicrap et al. (2016)

Table 2. DDPG algorithm for batch process control

Pseudo code
Randomly initialize the critic network $Q(s, a\|\theta^Q)$ and the actor network $\mu(s\|\theta^\mu)$ with weights θ^Q and θ^μ
Initialize the target networks Q' and μ'
Initialize the replay buffer R
For episode = 1, M **do**
Reset the environment and get the initial state s_1
For t =1, T **do**
If episode < Initial policy index:
Get action from the initial controller (ex. MPC)
Else:
Select action $a_t = \mu(s_t\|\theta^\mu) + \mathcal{N}_t(\text{noise})$
Receive r_t and s_{t+1} from the environment with a_t according to the criteria in Table 1
Store (s_t, a_t, r_t, s_{t+1}) in R
N Random sampling (s_i, a_i, r_i, s_{i+1}) from R
If s_{i+1} is the terminal state:
Set $Q_{predict,i} = r_i$
Else:
Set $Q_{predict,i} = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}\|\theta^{\mu'})\|\theta^{Q'})$
Update the critic by minimizing the loss: $L = \frac{1}{N}\sum_i \left(Q_{predict,i} - Q(s_i, a_i\|\theta^Q)\right)^2$
Update the actor policy using the sampled policy gradient: $\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s, a\|\theta^Q)\|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s\|\theta^\mu)\|_{s_i}$
Update the target networks: $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$, $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$
End for
End for

Simulation example: propylene oxide (PO) batch polymerization

To evaluate the performance of the RL based batch process controller, a polyether polyol process for polypropylene glycol production, which encompasses highly nonlinear dynamics and several path and end constraints, is employed. Monomer PO first reacts with alkaline anion and then oxy-propylene anion undertakes propagation followed by cation-exchange and proton-transfer reactions. A first-principles dynamic model including population balance equations of polymer chains

and monomers, and overall mass balances (Nie et al., 2013) were reformulated by using the method-of-moments (Mastan and Zhu, 2015) for the reactor simulation and model predictive control implementation. The process should be operated while satisfying two path constraints for heat removal duty and adiabatic end temperature, and three end-point constraints for final number average molecular weight (NAMW), final unsaturated chains per mass, and final concentration of unreacted monomer. The manipulated variables are reactor temperature T and monomer feed rate F (i.e., action $\in \{T, F\}$).

In this case study, the kinetic parameter of propagation reaction A_p is perturbed by assuming the uniform distribution of $\pm 30\%$ of its nominal value. Total reaction time and sampling interval are set as 480 min and 16 min, respectively, and perfect measurements are assumed as in previous studies (Jang et al., 2016; Jung et al., 2015). The initial controller used in this case is the standard NMPC with shrinking horizon, chosen to effectively deal with end-point constraints. To implement the proposed RL based controller with the DDPG algorithm, we use Tensorflow (Abadi et al., 2016) and OpenAI Gym (Brockman et al., 2016) in Python. The state comprises the reaction time *t* and 11 physical variables of the dynamic model, e.g., mole of PO and moments of polyol product. All of the state and action variables are normalized by their min and max values obtained from the NMPC simulation results with the varying uncertain parameter. $p(s_t)$ is set as the number of path constraints satisfied divided by the total number of path constraints, which is 3, and the performance index is calculated as the normalized value of the final mass of the product indicating the productivity. The results will be presented in the final version of the manuscript.

Conclusion

Control of batch processes is difficult due to their nonlinear and nonstationary dynamic behavior. In this study, a RL based batch process control strategy is proposed that can eventually give an optimal or improved control policy using closed-loop simulation or operation data. It is intended to improve upon the performance of MPC for problems that bear significant under uncertainty, e.g., plant-model mismatch. The controller learns the Q function that reflects the performance index and satisfaction of path and end-point constraints. The DDPG algorithm is used to learn the Q function and the optimal policy. The proposed RL based controller is tested with a PO batch polymerization example considering uncertainty in the model parameter. It is expected that the proposed strategy can handle the plant-model mismatch problem more effectively for batch processes.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Derek G. Murray, Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X. (2016). TensorFlow: A system for large-scale machine learning, *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*. pp. 265–283.

Bonvin, D., Srinivasan, B., Ruppen, D. (2001). Dynamic Optimization in the Batch Chemical Industry. *Chem. Process Control – CPC VI AIChE Symp*, 255–273.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W. (2016). OpenAI Gym.

Jang, H., Lee, J.H., Biegler, L.T. (2016). A robust NMPC scheme for semi-batch polymerization reactors. *IFAC-PapersOnLine* 49, 37–42.

Jung, T.Y., Nie, Y., Lee, J.H., Biegler, L.T. (2015). Model-based on-line optimization framework for semi-batch polymerization reactors. *IFAC-PapersOnLine* 28, 164–169.

Kober, J., Bagnell, J.A., Peters, J. (2013). Reinforcement learning in robotics: A survey. *Int. J. Rob. Res.* 32, 1238–1274.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D. (2016). *Continuous control with deep reinforcement learning. Conf. Pap. ICLR 2016.*

Mastan, E., Zhu, S. (2015). Method of moments: A versatile tool for deterministic modeling of polymerization kinetics. *Eur. Polym. J.* 68, 139–160.

Morari, M., Lee, J., (1999). Model predictive control: past, present and future. *Comput. Chem. Eng.* 23, 667–682.

Nie, Y., Biegler, L.T., Villa, C.M., Wassick, J.M. (2013). Reactor modeling and recipe optimization of polyether polyol processes: Polypropylene glycol. *AIChE J.* 59, 2515–2529.

Shin, J., Badgwell, T.A., Liu, K.-H., Lee, J.H. (2019). Reinforcement Learning – Overview of Recent Progress and Implications for Process Control. *Comput. Chem. Eng.* In Press, https://doi.org/10.1016/j.compchemeng.2019.05.029

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–9.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature* 550, 354–359.

Sutton, R.S., Barto, A.G. (2018). Reinforcement Learning. The *MIT Press.*