Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

A Phase Segmentation Approach for Applying Reinforcement Learning to Batch Polymerization Process Control

Haeun Yoo*, Boeun Kim, Jay H. Lee***

** Department of Biomolecular and Chemical Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea (e-mail: haeungd, jayhlee@kaist.ac.kr)*
***Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA (e-mail: bkim329@wisc.edu)*

Abstract: Nonlinear model predictive control (NMPC) or economic NMPC (eNMPC) is a widely studied optimal control method for batch processes with strongly nonlinear dynamics, but its performance can degrade severely in the presence of uncertainties in feedstock quality and other process characteristics. Reinforcement learning (RL) can be a good alternative in such cases since it can address stochastic uncertainties in a near-optimal manner using data samples from simulations or real operation. The downside is a large data requirement and unstable learning behavior, especially when the target system exhibits highly time-varying behavior as most batch processes do. To apply an RL algorithm to batch process control in a more stable and effective way, this study suggests a phase segmentation approach to consider the distinct dynamic characteristics of different phases. The approach designs separate reward functions and actor-critic networks. As a case study, optimal control of a polyol batch polymerization process is simulated to demonstrate the improvement in control policy brought by the phase segmentation approach and to compare its control performance with standard eNMPC.

Keywords: Batch polymerization, Reinforcement learning, Optimal control, Actor-Critic.

1. INTRODUCTION

Batch or semi-batch process is widely used for producing low-volume, high-value added products. Batch operation generally comprises three steps: feed charge, processing, and product discharge. Operating condition of a batch process is determined to meet given requirements for end product quality (e.g., composition, size, and shape) in a manner that is optimal with respect to productivity maximization or cost minimization. However, its inherent features, such as 1) nonstationary operation, 2) highly nonlinear dynamics, and 3) existence of path and end constraints, present significant challenges for operation and control. These problems are exacerbated by significant uncertainties in feedstock and other process variabilities (e.g. disturbances, noises, model errors).

For optimal control of batch processes, nonlinear model predictive control (NMPC) has been the most widely studied method. NMPC determines optimal control actions by solving an open-loop optimization problem at each time step after appropriate updates using measurements. Rather than following off-line determined reference trajectories, economic NMPC (eNMPC) tries to optimize a profit function on-line to enhance profitability of the process. However, when the model used has significant uncertainty, its performance can deteriorate and constraints can be violated. This is despite the feedback and re-optimization as it reacts to effects of uncertainty rather than proactively navigate through them. To overcome such limitation of MPC, various robust MPC strategies have been suggested, but the need for an uncertainty model, which often is not available, and significantly increased on-line computational load stand as obstacles to their practical implementations (Morari and H. Lee, 1999).

In this regard, the reinforcement learning (RL) approach, which can address stochastic uncertainty through off-line simulations and sample based learning, can be an effective alternative and its potential was demonstrated through several set point tracking control problems (Lee and Lee, 2005). In that case study, the set point tracking error was used as the (negative) reward in training a RL-based controller. RL-based control using economic reward functions have been studied previously but the studies addressed problems with small and discretized control space, which may be inadequate for many batch process control problems (Wilson and Martinez, 1997; Martinez, 1998).

In this work, we propose a RL based control strategy for a batch process with high dimensional and continuous state and action spaces. Three types of rewards are considered and a phase segmentation approach is suggested to account better address distinct dynamic characteristics of different steps in batch operation. Deep deterministic policy gradient (DDPG) algorithm (Lillicrap *et al.*, 2016) adopted as it is known to be effective in handling high dimensional continuous state and action spaces. Performances of the proposed RL based controller and the eNMPC controller are compared using an example of a polymerization process with uncertainty in its kinetic parameter.

Copyright lies with the authors

2. METHODS

2.1 Reward types

In RL, the agent gets rewards (or penalties) from the environment, which reflect the goal of a decision-making problem. In this study, three types of reward terms (as shown in Table 1) are contemplated to train the agent for high process performance and constraint satisfaction. The first is a reward term for satisfying path constraints (r_{path}). If all the path constraints are satisfied, the agent gets a reward, otherwise a penalty. The second is a reward term for satisfying end-point constraints (r_{end}), expressed as a sum of the reward for each end-point constraint ($r_{end,i}$). If the constraint is violated, the agent receives a penalty value proportional to degree of constraint violation. The last reward term is for process productivity (r_{prod}). This reward is needed to ensure high productivity in addition to on-spec product. For example, total mass in the reactor can be used as v to guarantee a sufficient production quantity as in the case study. α_j for the reward and penalty terms are hyperparameters which should be tuned along with the discount rate γ.

Table 1. Reward types

$\mathbf{r_{path}}$	$\begin{cases} +\alpha_{path}, & \text{if satisfies all path constriants} \\ -\alpha_{path}, & \text{otherwise} \end{cases}$
$\mathbf{r_{end,i}}$	$\begin{cases} +\alpha_{end}, & \text{if satisfies end point constraint i} \\ -\alpha_{end}{}' - (\text{var} - \text{boundary val.})_{scaled}, & \text{otherwise} \end{cases}$
$\mathbf{r_{prod}}$	v_{scaled} (v is a measure of process performance)

2.2 Phase segmentation

As said before, a batch process normally operates in three steps: feed charge, processing, and discharge. Without considering the discharge step, the entire reaction time can be divided into two phases. *Phase I* is the feeding-focused phase, so the reward is the sum of the rewards for satisfying the path constraints (r_{path}) and for achieving high productivity (r_{prod}) without regard to the end constraints. *Phase II* is the reaction-focused phase which must emphasize the satisfaction of end product quality specs, so r_{path} and r_{end} are assigned to non-terminal state and the terminal state, respectively. Even though rewards differ for different phases, the return value which is the cumulative rewards is calculated along the entire batch time and the critic network predicts that value. Table 2 summarizes the choice of rewards for different phases.

Table 2. Reward for each phase

Phase I: Feeding-focused	Phase II: Reaction-focused
$\mathbf{r} = \mathbf{r_{path}} + \mathbf{r_{prod}}$	if not terminal: $\mathbf{r} = \mathbf{r_{path}}$ if terminal: $\mathbf{r} = \mathbf{r_{end}}$

2.3 Monte-Carlo DDPG

DDPG is a method to train the actor and critic when state and action spaces are continuous and the policy trained is deterministic (Lillicrap *et al.*, 2016). This method uses the temporal difference (TD) update along with target networks to promote stable bootstrapping. However, even with the use of target networks, the critic and actor may converge to sub-optimal ones or even diverge to the boundary values. Due to the bootstrapping, the actor can be updated towards a wrong direction based on inaccurately estimated values and this in turn leads to bad samples with low rewards (Tsitsiklis and Roy, 2000; Fujimoto, Van Hoof and Meger, 2018). The Monte-Carlo (MC) update method can be a better choice for batch process problems which often involve irreversible transitions and require a precise prediction of the terminal reward early on. Therefore, we modified the DDPG algorithm to adopt the MC update in place of the TD update. The actors are updated with the same gradient calculation as in DDPG, but target networks and bootstrapping are not used. The return value is calculated once an episode ends ($G_t(s_t, a_t) = r_t + \gamma G_{t+1}$), and this value is used as a target of the critic networks. To initialize the network parameters with reasonable values, a sub-optimal controller such as eNMPC can be used in the simulation for the first few episodes.

3. CASE STUDY

3.1 Propylene oxide (PO) batch polymerization

To evaluate the performance of the proposed RL based control strategy, a polyether polyol process for polypropylene glycol production is used as it involves both path and end constraints. The monomer PO first reacts with the alkaline anion and then the oxy-propylene anion undertakes the propagation, which is followed by the cation-exchange and proton-transfer reactions. A first-principles dynamic model including the population balance equations of polymer chains and monomers and overall mass balance was reformulated with the method-of-moments for the reactor simulation and eNMPC implementation (Nie *et al.*, 2013; Mastan and Zhu, 2015). There are two path constraints, one on the heat removal duty and the other on the adiabatic temperature, and three end-point constraints, which represent the specs on the final number average molecular weight (NAMW), final unsaturated chains per mass (USV), and final concentration of unreacted monomer (Unrct). The manipulated variables are the reactor temperature T and the monomer feeding rate F (i.e., $action \in \{T, F\}$). In this case study, the kinetic parameter of the propagation reaction A_p is perturbed by assuming a uniform distribution in the range of ±10 % of its nominal value. Total reaction time and sampling interval are set as 480 min and 20 min, respectively, as in the previous studies (Jung *et al.*, 2015; Jang, Lee and Biegler, 2016) and perfect measurements of the state are assumed.

3.2 Actor-Critic networks training

To implement the proposed algorithm, we employed PyTorch (Paszke *et al.*, 2017) in Python. The state comprises reaction time t and 11 physical variables of the dynamic model including the number of moles of PO and moments of polyol product. Phase I and Phase II are divided at 400 min according

Table 3. Rewards for the path and end-point constraints (Jung *et al.*, 2015)

r_{path}	**Heat Duty**	Heat removal duty ≤ 430 [J/s]	if (HeatDuty ≤ 430)&(Tads ≤ 192): $r_{path} = 0.5$ else: $r_{path} = -0.5$
	Tads	Adiabatic temperature rise ≤ 192 [°C]	
r_{end}	**NAMW**	Final NAMW ≥ 3027.74 [g/mol]	if (NAMW ≥ 3027.74): $r_{end,1} = 1$ else: $r_{end,1} = -0.5 - (3027.74 - NAMW)/1500$
	USV	Final Unsat. Value ≤ 0.02 [mmol/g polyol]	if (USV ≤ 0.02): $r_{end,2} = 1$ else: $r_{end,2} = -0.5 - 3100 * (USV - 0.02)$
	Unrct	Final unreacted PO ≤ 2000 [ppm]	if (Unrct ≤ 2000): $r_{end,3} = 1$ else: $r_{end,3} = -0.5 - (Unrct - 2000)/3900$

to the solutions from the initial sub-optimal controller, which is the eNMPC controller with a shrinking horizon, chosen to deal with the end-point constraints. v in r_{prod} is defined as the total mass in the reactor and other reward terms are chosen as reported in Table 3. We initially trained the networks using samples from 21 episodes with eNMPC and then trained the resulting networks further with the MC-DDPG algorithm using data from 2500 more episodes.

3.3 Results – Network training with phase segmentation

To demonstrate the benefit of using the phase segmentation for training the RL based controller, two cases are compared: Case 1) One actor-critic network for the entire batch, Case 2) Two actor-critic networks, one for each of the two phases of batch.

In the first case, one actor and critic network are trained without using the phase segmentation approach and the sum of r_{path} and r_{prod} are always used as the reward. The second case trains separate actor and critic networks for each of the two phases. Their performances are evaluated in the cases of -10, 0 and +10% perturbations to A_p.

As shown in Fig. 1, the actor trained without phase segmentation converged to the unreasonable value which also incurs less amount of product with unsatisfied quality (see Table 4). Note that, in this case study, the action policy should be given by a high feeding rate with a low temperature during Phase I and a low feeding rate with a high temperature during Phase II. In Case 2, the correct phase-dependent trends in the feeding rate and temperature profiles are observed as shown in Fig. 3, and also the path constraints are almost satisfied showing that in -10% of A_p the constraint for adiabatic temperature is slightly violated. From this result, we can see that the RL based controller gives a converged profile which satisfies all the constraints, similar to the result of robust MPC. In terms of end-point constraints, one constraint is violated by a small bit in the case of -10% of A_p as shown in Table 6.

Table 4. Performance of Case 1 for handling the end-point constraints

	-10%	0%	+10%	
NAMW	3247.17	3599.38	3953.41	≥ 3027.74
USV	0.0297	0.0279	0.0264	≤ 0.02
Unrct	3	3	2	≤ 2000
Total mass	797.67	797.64	797.63	

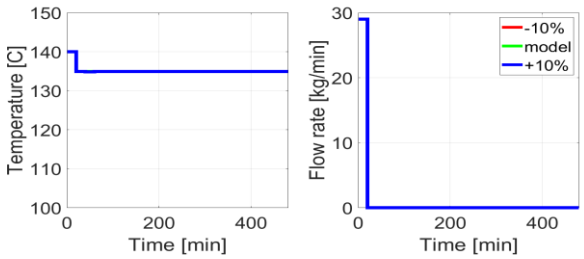

Fig. 1. Action profile of Case 1) One actor-critic network for the entire batch

3.4 Results – Control performance

Performances of the RL based controller with phase segmentation and the eNMPC controller are compared for the same three parameter perturbation cases tested in Section 3.3. Fig. 2 shows the simulation results using the eNMPC controller. It gives very different profiles of the feeding rate with the perturbations in A_p and violations in both of the path constraints occur, especially the upper limit of adiabatic temperature. The end-point constraints are also violated with the -10% perturbation in A_p as shown in Table 5.

On the other hand, the RL based controller with phase segmentation and two actor and critic networks shows good performance, giving consistent action profiles despite the perturbations in the parameter (see Fig. 3). The resulting action policy satisfies the path constraints except for the near boundary in the case of -10% of A_p . The end-point constraints are satisfied except that for USV in the case of -10% of A_p as reported in Table 6. Compared with the eNMPC results, the RL based controller shows better performance in satisfying the constraints using a smoother action profile. The one slight constraint violation that remained will be addressed in our future work by relaxing the reaction time until all the constraints are fully satisfied.

Table 5. Performance of eNMPC for handling the end-point constraints

	-10%	0%	+10%	
NAMW	3158.95	3866.70	4680.56	≥ 3027.74
USV	0.0204	0.0200	0.0198	≤ 0.02
Unrct	2444.77	1974.27	1612.73	≤ 2000
Total mass	5555.16	6955.77	7839.42	

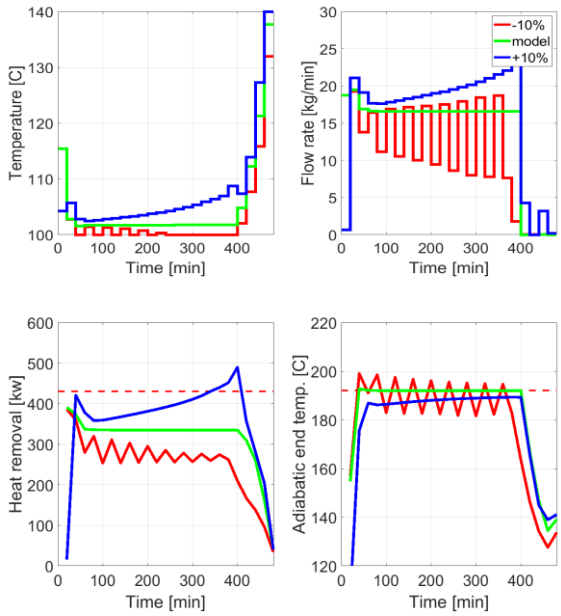

Fig. 2. Performance of eNMPC with manipulated variables and path constraints for the three perturbation cases

Table 6. Performance of the RL based controller for handling the end-point constraints

	-10%	**0%**	**+10%**	
NAMW	3044.69	3059.14	3071.22	$\geq$ 3027.74
USV	0.0214	0.0192	0.0174	$\leq$ 0.02
Unrct	1898	1094	635	$\leq$ 2000
Total mass	5798.88	5808.77	5817.14	

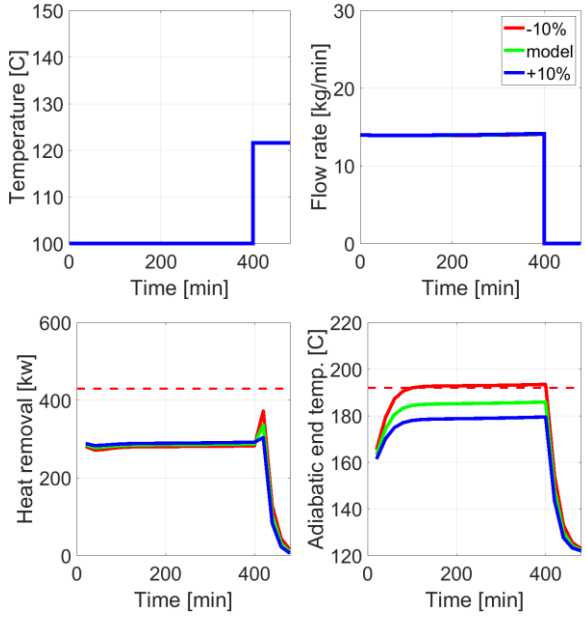

Fig. 3. Performance of the RL based controller with manipulated variables and path constraints

4. CONCLUSIONS

A RL based batch process control strategy was proposed along with the phase segmentation approach (i.e., feeding and reaction phases). The suggested strategy was tested on a batch polymerization example and the beneficial effect of the phase segmentation on the training and control performance was observed. Comparing with the eNMPC, the RL-based controller showed enhanced ability to satisfy the path and end-point constraints in the presence of parameter errors. For future work, the RL based control formulation will be extended to minimize reaction time while satisfying all path and end-point constraints.

REFERENCES

Fujimoto, S., Van Hoof, H. and Meger, D. (2018) 'Addressing Function Approximation Error in Actor-Critic Methods', *35th International Conference on Machine Learning, ICML 2018*, 4, pp. 2587–2601.

Jang, H., Lee, J. H. and Biegler, L. T. (2016) 'A robust NMPC scheme for semi-batch polymerization reactors', *IFAC-PapersOnLine*. Elsevier B.V., 49(7), pp. 37–42.

Jung, T. Y. *et al.* (2015) 'Model-based on-line optimization framework for semi-batch polymerization reactors', *IFAC-PapersOnLine*. Elsevier Ltd., 28(8), pp. 164–169.

Lee, J. M. and Lee, J. H. (2005) 'Approximate dynamic programming-based approaches for input-output data-driven control of nonlinear processes', *Automatica*, 41(7), pp. 1281–1288.

Lillicrap, T. P. *et al.* (2016) 'Continuous control with deep reinforcement learning', *conference paper at ICLR 2016*. Available at: http://arxiv.org/abs/1509.02971.

Martinez, E. C. (1998) 'Learning to control the performance of batch processes', *Chemical Engineering Research and Design*, 76(6 A6), pp. 711–722.

Mastan, E. and Zhu, S. (2015) 'Method of moments: A versatile tool for deterministic modeling of polymerization kinetics', *European Polymer Journal*. Elsevier Ltd, 68, pp. 139–160.

Morari, M. and H. Lee, J. (1999) 'Model predictive control: past, present and future', *Computers & Chemical Engineering*, 23(4–5), pp. 667–682.

Nie, Y. *et al.* (2013) 'Reactor modeling and recipe optimization of polyether polyol processes: Polypropylene glycol', *AIChE Journal*, 59(7), pp. 2515–2529.

Paszke, A. *et al.* (2017) 'Automatic differentiation in PyTorch', *NIPS 2017 Workshop Autodiff Submission*.

Tsitsiklis, J. N. and Roy, B. Van (2000) 'Analysis of Temporal-Difference Learning with Function Approximation', *Journal of the Advances in neural information processing systems*, (1988).

Wilson, J. A. and Martinez, E. C. (1997) 'Neuro-fuzzy modeling and control of a batch process involving simultaneous reaction and distillation', *Computers and Chemical Engineering*, 21(SUPPL.1).