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

Model-Free Control Design for Loop Heat Pipes Using Deep Deterministic Policy Gradient

Thomas Gellrich * **Yi Min** * **Stefan Schwab** *
Soeren Hohmann **

* *FZI Research Center for Information Technology, Karlsruhe, Germany (e-mail: gellrich@fzi.de)*
** *Institute of Control Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany (e-mail: soeren.hohmann@kit.edu)*

Abstract: In this paper, a model-free adaptive control design for loop heat pipes (LHPs) based on the reinforcement learning (RL) method of deep deterministic policy gradient (DDPG) is presented. An LHP as a heat transport system combines complex, thermodynamic processes, which are not yet fully described in a dynamic control model over the entire LHP operating range for model-based control design. However, RL methods provide the controller with the ability to improve its control performance without a model by analyzing and rewarding the performance online. The aim of an LHP controller is to keep the LHP operating temperature as close as possible to the fixed setpoint temperature by additional heating, while the amount of heat to be transported and the temperature of the heat sink change over time. A validated numerical simulation of the LHP provides a safe, dynamic environment for the training of the learning controller. In comparison with the commonly used PI controller with a single temperature feedback, the control performance of the learning controller observing the same temperature achieves similar control results. Furthermore, multiple observations are easily incorporated into a model-free learning controller, whereby the additional feedback of further temperature measurements ensures an improved performance over the entire operating range.

Keywords: Reinforcement learning control, learning for control, model-free adaptive control, loop heat pipe, aerospace

1. INTRODUCTION

An important aspect of optimal operation of electronic components is their thermal control. In aerospace systems, the thermal control of electronic components is a challenge, since heat can only be released into space via radiation. For this reason, heat transport systems are used to control the temperatures of electronic components by transferring the excess heat to a remote radiator. For a reliable and efficient heat transport, loop heat pipes (LHPs) are widely used for their passive, two-phase working principle (Ku (1999)). The phase of the working fluid inside the LHP is changed through evaporation and condensation to achieve a higher heat transfer coefficient than with direct heat conduction. Without the use of power-consuming mechanical pumps, the mass flow in the pipes between the locally separated evaporator and the condenser are established by capillary forces in a fine-pored wick inside the evaporator (Maydanik (2005)). In order to keep the temperatures of the electronic components in a desired temperature corridor, the operating temperature of the LHP is controlled by additional heating. While the LHP operating temperature is adapted by the control heater, the heat transport of the LHP as its natural behavior is maintained (Ku (2008)). This natural behavior of the operating temperature depends mainly on two influences. The temperature at the radiator as the heat sink of the LHP changes due to variable insolation in the orbit. The operating statuses of the electronic components define the amount of excess heat that forms the heat load at the evaporator. Various control algorithms for the control heater were used in experiments to control different LHP temperatures with a single temperature feedback (heuristic PID controllers (Ku et al. (2011b), Ku et al. (2011a)) and two-point controllers (Khrustalev et al. (2014), Ku et al. (2011b))). However, the complexity of the LHP's working principle demands for more sophisticated control algorithms for the control heater to improve the performance of the operating temperature control against the external influences (Ku et al. (2011a)). The proportional time-delaying behavior of the CC temperature motivated the identification of a black-box model to study the LHP startup behavior in Huang et al. (2009) and the calculation of the PI parameters of a two-degree-of-freedom controller for improved disturbance response in Gellrich et al. (2018a). In Gellrich et al. (2018b), a PI controller was designed based on a nonlinear LHP model with only accurate dynamics for constant disturbances. Although the accuracy of this analytical LHP model and the performance of the model-based controller could be improved in Gellrich et al. (2019) by a nonlinear online parameter estimation and temperature prediction, the controller suffered from the same problem

as the aforementioned controllers: an increased sensitivity to heat load changes with decreasing sink temperature while controlling only one locally bounded LHP temperature measurement.

To the best of the authors' knowledge, no nonlinear control model of the LHP exists to design a model-based controller with the feedback of multiple temperatures so far. That's why the authors strive for a model-free control design, where multiple measurements can be easily included as controller inputs. Model-free adaptive controls based on reinforcement learning (RL) have proven to be very effective in learning optimal control policies for complex nonlinear systems (Henze and Schoenmann (2003), Qi et al. (2019), Mitchell and Petzold (2018)). RL belongs to the field of machine learning that emphasizes how to act by following a policy based on the current environment to maximize the expected long-term benefit. Although there are many differences, RL is inspired by behavioral theory in psychology, where organisms gradually form expectations under the stimulation of rewards or punishments given by the environment, and produce habitual behaviors that can obtain the maximum benefit (Kaelbling et al. (1996)). Since the LHP temperatures react very sensitively to changes of the control heater output, the authors focus on the design of a deep deterministic policy gradient (DDPG) agent with continuous action space (Lillicrap et al. (2015)). Thus, the discretization problem arising from the curse of dimensionality with RL methods like Q-learning (Watkins and Dayan (1992)) or deep Q-network (DQN), introduced by Mnih et al. (2015), can be circumvented. In addition, the number of environment observations of DDPG agents can be easily increased with additional temperature measurements to improve the monitoring of the LHP for disturbance rejection.

In this work, two contributions are presented: First, the model-free control design for control heaters of LHPs. Second, the model-free control design with a multiple temperature feedback.

This paper is structured as follows: In Section 2, the LHP operating characteristics are presented. After the problem statement in Section 3, the implementations of the learning environment and the DDPG agents are described in Section 4. The performances of the learned controllers are compared and evaluated in Section 5, followed by the conclusions in Section 6.

2. LHP OPERATING CHARACTERISTICS

Before the problem statement is given, the LHP working cycle is explained in detail for a better understanding of the challenges in controlling the LHP. The working cycle of the LHP consists of multiple heat exchange processes between five LHP components. The visualization of these LHP components is presented in Fig. 1.

The excess heat of the distributed electronic components is transferred to the evaporator via two arterial heat pipes and vaporizes the liquid working fluid in the primary wick. Simultaneously, capillary forces arise at the liquid-vapor interface forcing a flow of vapor through the vapor line (VL) into the condenser. In the condenser, the vapor condensates as its phase returns from vapor to liquid while

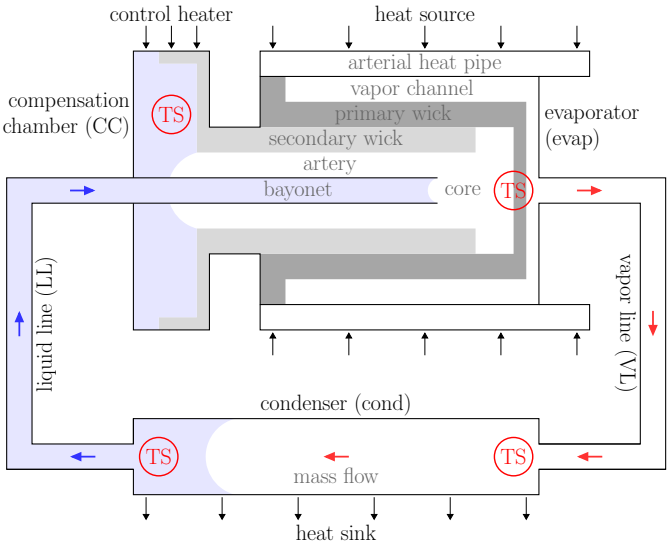

Fig. 1. Schematic of the LHP with four temperature sensors (TS) (cf. Ku (1999))

releasing its heat to the heat sink. The forming liquid is subcooled and flows through the liquid line (LL) and the compensation chamber (CC) into the evaporator core, where it supplies the primary wick with liquid again. A continuous heat transport depends on a permanently wetted primary wick. That's why a secondary wick between the CC and the evaporator ensures a liquid exchange between the CC and the primary wick. In transient states, the positions of the liquid-vapor interfaces of the two-phase CC and the two-phase condenser are counterbalanced by the liquid and vapor mass flows for a balanced mass distribution in the hermetically sealed LHP.

The saturated CC temperature governs the operating temperature (OT) of the LHP at the evaporator (Chernysheva et al. (2007)) establishing a balance between the subcooling of the entering liquid from the LL and the heat leakage as part of the heat load from the evaporator (Ku (2008)). Therefore, the OT of the LHP depends on both the heat load and the sink temperature. The typical U-shaped curve of the natural steady-state operating temperature (SSOT) is depicted in Fig. 2 in case of a higher ambient temperature T_{amb} at the LL than the sink temperature T_{sink} at the condenser.

The evaporation of the working fluid in the primary wick determines the mass flow rate proportionally. For this reason, the residence time of the subcooled liquid in the LL decreases with increasing heat load leading to a decreasing total heat input from the ambient to the liquid, when the ambient temperature is higher than the sink temperature. At the same time, the liquid-vapor interface in the condenser moves towards the condenser outlet and thereby shortens the subcooling length in the condenser. As a result, the minimum $T_{ot,min}$ of the SSOT is reached at the heat load $\dot{Q}_{ot,min}$ because of maximal subcooling in the condenser and minimal heat gain in the LL. Higher heat loads lead to an increasing SSOT due to a further decreasing subcooling and an increasing temperature at the condenser outlet correspondingly. Two LHP modes are distinguished at the heat load $\dot{Q}_{trans}$: variable thermal conductance mode and fixed thermal conductance mode.

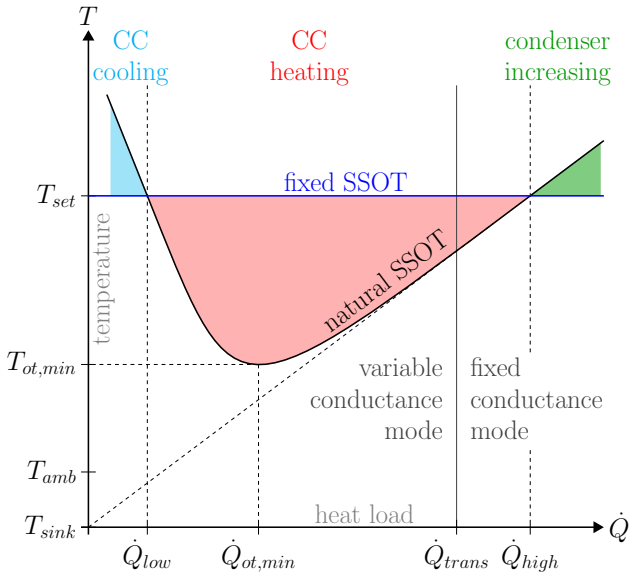

Fig. 2. Steady-state operating temperature (SSOT) of the LHP (cf. Ku (2008) and Chuang (2003))

At this point, the overall LHP heat transfer coefficient is the highest, because the condenser is fully used (Chuang (2003)). At higher heat loads than $\dot{Q}_{trans}$, the SSOT increases accordingly. Although the heat load changes over time, moving from one SSOT to another, the overall goal is to fix the temperature of the electronic components at a constant temperature, which requires a constant OT of the LHP as heat transport system at a desired setpoint temperature T_{set}. Thus, the output power $\dot{Q}_{CC}$ of a control heater on the CC influences the power balance in the CC between the heat loads $\dot{Q}_{low}$ and $\dot{Q}_{high}$. At lower heat loads, the CC must be cooled. At higher heat loads, the condenser length must be enlarged to increase the heat exchange with the heat sink.

The control method with a control heater regulating the CC temperature provides stable temperatures and less temperature oscillations at low heat loads (Ku et al. (2011a)). Due to the nonlinear behavior of the LHP, a learning controller for the control heater must be supplied with comprehensive information about the operating status of the LHP with the aid of appropriate temperature measurements, since only noninvasive surface measurements are possible for hermetically sealed LHPs. In addition, the dynamic training environment must also cover the whole operating range of the LHP in order to achieve a functioning learning controller in all operating points.

3. PROBLEM STATEMENT

Common controls for the OT of the LHP are restricted to the sole feedback of the CC temperature due to the simplicity of the control structures and the available control models. They focus only on the control error between the setpoint temperature and the CC temperature as single controller input in order to keep the OT inside a small corridor around the desired setpoint temperature against varying operating conditions. The major cause for varying operating conditions are the sink temperature and the heat load as time-variant disturbances, which have a direct effect on LHP parameters (Gellrich et al. (2019)). To improve the performance of the OT control, the controller's information about the LHP must be improved. Indeed, further temperature measurements around the loop for operation monitoring are available. These can be used by the controller to determine the LHP operating status closer to the disturbance impact points than with only the CC temperature measurement (see Fig. 1). So far, temperature sensors at the evaporator (closer to the heat load), at the condenser inlet, and the condenser outlet (closer to the heat sink) are unused by the controller, but could provide the necessary information for increased performance.

The first goal of this work is the design of a model-free RL agent for the control heater that performs at least as well as the commonly used PI controller. The agent learns an optimal behavior in dependence on the CC temperature. An appropriate dynamic training environment is designed that covers the entire operating range. By maximizing the reward for keeping the CC temperature as close as possible to the setpoint temperature, the improved performance of the controller against time-variant disturbances is achieved. The second goal is to extend the state-of-the-art single temperature feedback by the available temperature measurements to enhance the controller's information about the system for improved disturbance rejection.

4. IMPLEMENTATION

The challenge of finding the optimal policy with RL is given by the trade-off between exploration for better actions in the future and exploitation of successful actions in the past to maximize future rewards. While dynamic programming algorithms in optimal control require prior knowledge of the environment's dynamics to learn the optimal policy, temporal-difference (TD) methods do not need a dynamic model of the environment (Richard S. Sutton (2018)). In addition, on-policy TD methods cannot separate exploration from learning, whereas off-policy TD methods can (Singh et al. (2000)). Q-learning (Watkins and Dayan (1992)) is such a model-free, off-policy TD method that solves the Riccati equation online to learn the optimal policy of the agent while controlling the environment (Lewis and Vrabie (2009)). As well as its extension to deep neural networks (DNN) for function approximation, introduced by Mnih et al. (2015) as deep Q-network (DQN), both algorithms train agents with discrete action outputs. Since small changes of the control heater result in strong temperature changes, the design of deep deterministic policy gradient (DDPG) agents based on Lillicrap et al. (2015) with a continuous action space is preferred.

The RL control problem with the LHP is visualized as Markov Decision Process (MDP) (van Otterlo and Wiering (2012)) in Fig. 3. The validated numerical simulation of an LHP on a test bench in MATLAB, based on Meinicke et al. (2019) and extended for controller validation in Gellrich et al. (2019), is taken as the safe environment in the MDP. It uses the finite-difference method to solve the partial differential equations of the condenser and an iterative solution method to calculate the temperatures at the four sensors. The agent is given by the control heater, which applies the power $\dot{Q}_{cc,t}$ as action to the CC of the LHP in

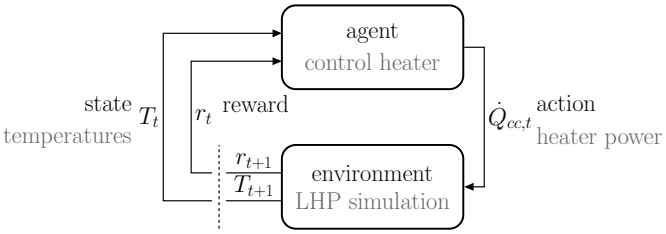

Fig. 3. RL control problem with the LHP as MDP (cf. Richard S. Sutton (2018))

the environment to control the OT of the LHP. The agent observes the resulting state T_{t+1} of the environment via the measurable temperatures of the LHP and receives a reward r_{t+1} for its action. The evaluation of the state and the action is stored in a buffer and used to improve the policy of the agent. Based on the modified policy, the agent decides for his next action $\dot{Q}_{cc,t+1}$. This updating process is repeated several times in an episode for several episodes in a training until the agent determines a satisfactory policy. After the training process, the control heater is expected to be able to deliver the proper power at any state of the environment to maintain the CC temperature at the desired setpoint temperature T_{set}, although the LHP is disturbed by the sink temperature T_{sink} and the heat load $\dot{Q}_{load}$ in the environment. Considering the real LHP, the ranges of the input parameters are given by:

$$30\,\mathrm{W} \leq \dot{Q}_{load} \leq 75\,\mathrm{W}, \tag{1}$$
$$-15\,^\circ\mathrm{C} \leq T_{sink} \leq 15\,^\circ\mathrm{C}, \tag{2}$$
$$0\,\mathrm{W} \leq \dot{Q}_{cc} \leq 10\,\mathrm{W}, \tag{3}$$
$$T_{set} = 27\,^\circ\mathrm{C}. \tag{4}$$

The agent is trained with the policy-based DDPG algorithm of Lillicrap et al. (2015) provided by the Reinforcement Learning Toolbox in MATLAB (MathWorks (2019)). Because of its actor-critic structure, two DNNs are designed. The actor network approximates the policy function and the critic network approximates the Q-value function, which measures the performance of the policy function. The pictorial representation of both networks is shown in Fig. 4.

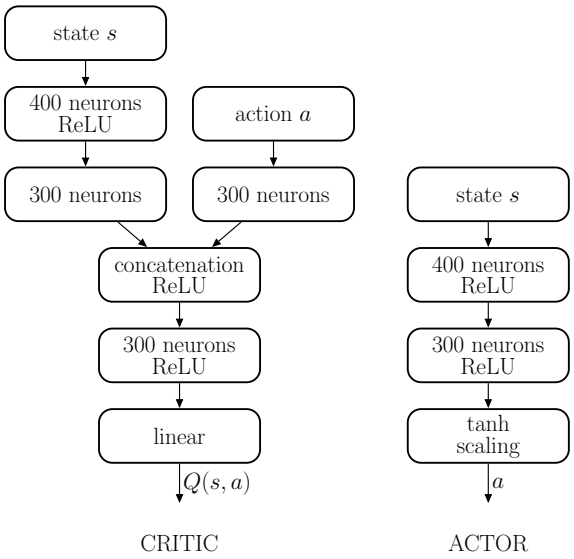

Fig. 4. DNN structures of the actor-critic DDPG agent

The actor network and the critic network consist of fully connected layers with multiple neurons and rectified linear units (ReLU) as activation functions, both commonly used in DNNs. The single output of the critic network is calculated linearly by a fully connected layer with one neuron. A hyperbolic tangent (tanh) layer and a scaling layer form the final layers of the actor network to consider a continuous action space within the limitation (3) of the control heater. Since the actor network outputs an action directly, a stochastic noise model is added to ensure exploration. In MATLAB, the Ornstein-Uhlenbeck noise (Uhlenbeck and Ornstein (1930)) is used to influence the exploration behavior of DDPG agents by tuning the mean and the variance of the noise signal. For numerical stability in the DDPG learning process, both the critic and the actor network are copied, so that two target networks track both learning networks smoothly. Furthermore, the networks are updated by sampling minibatches randomly from an experience replay buffer (Lillicrap et al. (2015)). The chosen hyperparameters for the DDPG training process are listed in Table 1.

Table 1. DDPG hyperparameter

Parameter	Value
learning rate of actor/critic	10^{-4}
sample time	1 s
target smooth factor	10^{-3}
experience buffer length	10^6
minibatch size	64
noise variance	0.5
noise mean	0.0
noise variance decay rate	10^{-5}

The reward function has a great impact on the learning result. For the control design, a linear reward function of the control error as difference between the CC temperature $T_{cc,t}$ and the setpoint temperature T_{set} is chosen as

$$r_t = -10 \cdot |T_{cc,t} - T_{set}|\,. \tag{5}$$

For comparison, two agents of the control heater are trained for OT control. The first agent, single DDPG (sDDPG), observes the state of the environment through the same signals as the commonly used PI controller, which are the setpoint temperature, the current and the last CC temperature. The second agent, multiple DDPG (mDDPG), receives the same signals, but additionally the three temperature measurements at the evaporator, the condenser inlet, and the condenser outlet (see Fig. 1), which are already available for operation monitoring. The training profile, depicted in Fig. 5, is constructed in such a way that the entire LHP operating range is covered dynamically, as stipulated in Sec. 3.

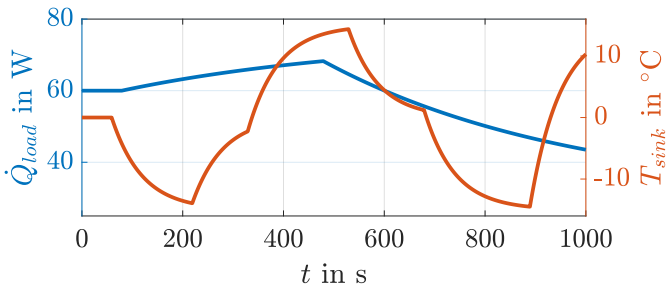

Fig. 5. DDPG training profile

5. NUMERICAL VALIDATION

The validation of the trained agents is performed by comparing the control performances of both agents when controlling the environment with the benchmark profile of the disturbances in Fig. 6.

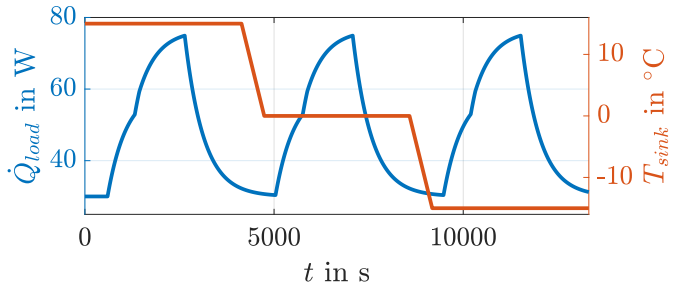

Fig. 6. DDPG benchmark profile

To validate the advantage of model-free controllers and the feedback of multiple temperatures, the performances of both agents are compared to the commonly used PI controller, which is model-based designed for the numerical simulation in Gellrich et al. (2019). Fig. 7 shows the relevant variables of the three controllers: the CC temperature T_{cc} in the upper subplot, the control heater output $\dot{Q}_{cc}$ in the middle subplot, and the difference $\Delta\dot{Q}_{cc}$ of the control heater output to the control heater output of the PI controller in the lower subplot. The sDDPG in red and the mDDPG in yellow achieve similar control performances as the PI controller in blue. All three controllers keep the CC temperature in a narrow corridor of $\Delta\dot{Q}_{cc} = \pm 0.5\,\mathrm{K}$ around the setpoint temperature. The aforementioned sensitivity of the LHP to control heater output changes in Sec. 1 becomes obvious in Fig. 7. Since the control heater outputs in the middle subplot are close to each other, small differences in the control heater outputs result, as shown in the lower subplot. Nevertheless, the controlled CC temperatures clearly show different profiles. At the start, both learning controllers react with a spike in the control heater output to the initial conditions of the numerical simulation, which results in a temperature offset compared to the smoother start of the PI controller. Although the heat load changes and the sink temperature decreases, the maximal spikes of the mDDPG-controlled CC temperature stay in the same corridor, whereas the CC temperatures controlled by the sDDPG and the PI controller tend to higher spikes with lower sink temperatures. To quantify the control performance, the maximal absolute deviation (MAD) and the root mean square error (RMSE) of the three controllers over the entire time frame between the CC temperatures and the setpoint temperature are both listed in Table 2.

Table 2. Maximal absolute deviation (MAD) and root mean square error (RMSE) of the three controllers

controller	MAD in K	RMSE in K
PI	0.48	0.12
sDDPG	0.46	0.14
mDDPG	0.16	0.08

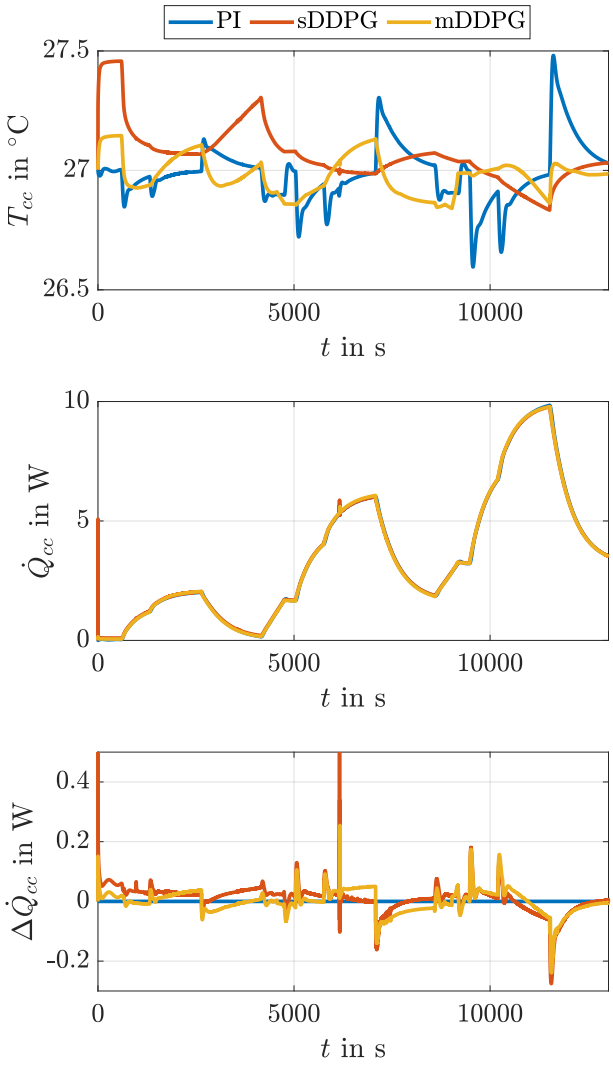

Fig. 7. Comparison of the three control heater controllers: CC temperature T_{cc}, control heater output $\dot{Q}_{cc}$, and control heater output difference $\Delta\dot{Q}_{cc}$ of the PI controller in blue, the sDDPG in red, and the mDDPG in yellow.

Table 2 indicates that the MAD of the CC temperature from the setpoint temperature is decreased by the learning controllers compared to the model-based PI controller. Furthermore, the RMSE of the mDDPG is smaller than the RMSE of the PI controller. The slightly higher RMSE of the sDDPG results from the initial control heater output spike. However, the benefit of additional temperature measurements for the OT control of LHPs is verified in Fig. 7 and in Table 2.

6. CONCLUSIONS

In this paper, two learning controllers for the control heater of an LHP based on DDPG have been designed and trained to control the LHP operating temperature at a fixed setpoint temperature despite time-variant disturbances. The control performances of the learning controllers have proven to be as good as the control performance of the commonly used, model-based PI controller.

By extending the number of observations of one controller agent with further available temperature measurements, the disturbance rejection of the control heater to date has been improved. The additional information about the state of the LHP has lifted the sensitivity of the control heater controller to heat load changes at lower sink temperatures. For an overall guaranty of the functionality and stability of the learning controllers in all operating points of the LHP, large sets of training data are required, which leads to long measurement times, since the temperature processes are very slow. For this reason, the state-of-the-art analytical models of the LHP have to be improved, especially at the condenser, to model the dynamics of all measured LHP temperatures. Then, the validated benefit of a multiple temperature feedback, shown in this work, can be applied in model-based controllers with proven stability to facilitate the adaption to different LHP designs.

REFERENCES

Chernysheva, M.A., Vershinin, S.V., and Maydanik, Y.F. (2007). Operating temperature and distribution of a working fluid in LHP. *International Journal of Heat and Mass Transfer*, 50(13), 2704–2713. doi: 10.1016/j.ijheatmasstransfer.2006.11.020.

Chuang, P.Y. (2003). *An improved steady-state model of loop heat pipe based on experimental and theoretical analyses*. Ph.D. thesis, Pennsylvania State University, Department of Mechanical and Nuclear Engineering.

Gellrich, T., Meinicke, S., Knipper, P., Hohmann, S., and Wetzel, T. (2018a). Two-degree-of-freedom heater control of a loop heat pipe based on stationary modeling. In *48th International Conference on Environmental Systems, Albuquerque, New Mexico, USA*.

Gellrich, T., Schuermann, T., Hobus, F., and Hohmann, S. (2018b). Model-based heater control design for loop heat pipes. In *2nd IEEE Conference on Control Technology and Applications (CCTA)*. IEEE. doi: 10.1109/ccta.2018.8511470.

Gellrich, T., Zhang, X., Schwab, S., and Hohmann, S. (2019). Nonlinear model identification adaptive heater control design for loop heat pipes. In *3rd IEEE Conference on Control Technology and Applications (CCTA)*.

Henze, G. and Schoenmann, J. (2003). Evaluation of reinforcement learning control for thermal energy storage systems. *HVAC&R Research*, 9(3), 259–275. doi: 10.1080/10789669.2003.10391069.

Huang, B.J., Huang, H.H., and Liang, T.L. (2009). System dynamics model and startup behavior of loop heat pipe. *Applied Thermal Engineering*, 29(14), 2999–3005. doi: 10.1016/j.applthermaleng.2009.03.015.

Kaelbling, L.P., Littman, M.L., and Moore, A.W. (1996). Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4, 237–285. doi:10.1613/jair.301.

Khrustalev, D., Stouffer, C., Ku, J., Hamilton, J., and Anderson, M. (2014). Temperature control with two parallel small loop heat pipes for GLM program. *Frontiers in Heat Pipes*, 5(9).

Ku, J., Paiva, K., and Mantelli, M. (2011a). Loop heat pipe operation using heat source temperature for set point control. Technical Report 20110015274, NASA Technical Reports Server.

Ku, J. (1999). Operating characteristics of loop heat pipes. In *SAE Technical Paper No. 1999-01-2007*. SAE International. doi:10.4271/1999-01-2007.

Ku, J. (2008). Methods of controlling the loop heat pipe operating temperature. In *SAE Technical Paper No. 2008-01-1998*. SAE International. doi:10.4271/2008-01-1998.

Ku, J., Paiva, K., and Mantelli, M. (2011b). Loop heat pipe transient behavior using heat source temperature for set point control with thermoelectric converter on reservoir. In *9th Annual International Energy Conversion Engineering Conference*. American Institute of Aeronautics and Astronautics.

Lewis, F.L. and Vrabie, D. (2009). Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, 9(3), 32–50. doi:10.1109/mcas.2009.933854.

Lillicrap, T.P., Hunt, J.J., Pritzel, A.e., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv e-prints*, arXiv:1509.02971.

MathWorks (2019). *Reinforcement Learning Toolbox user's guide*. The MathWorks, Inc., Natick, Massachusetts, USA.

Maydanik, Y. (2005). Loop heat pipes. *Applied Thermal Engineering*, 25, 635–657.

Meinicke, S., Knipper, P., Helfenritter, C., and Wetzel, T. (2019). A lean approach of modeling the transient thermal characteristics of loop heat pipes based on experimental investigations. *Applied Thermal Engineering*, 147, 895–907. doi: 10.1016/j.applthermaleng.2018.10.123.

Mitchell, B.A. and Petzold, L.R. (2018). Control of neural systems at multiple scales using model-free, deep reinforcement learning. *Scientific Reports*, 8(1). doi: 10.1038/s41598-018-29134-x.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533. doi:10.1038/nature14236.

Qi, X., Luo, Y., Wu, G., Boriboonsomsin, K., and Barth, M. (2019). Deep reinforcement learning enabled self-learning control for energy efficient driving. *Transportation Research Part C: Emerging Technologies*, 99, 67–81. doi:10.1016/j.trc.2018.12.018.

Richard S. Sutton, A.G.B. (2018). *Reinforcement learning*. The MIT Press.

Singh, S., Jaakkola, T., Littman, M.L., and Szepesvári, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3), 287–308. doi:10.1023/A:1007678930559.

Uhlenbeck, G.E. and Ornstein, L.S. (1930). On the theory of the Brownian motion. *Physical Review*, 36(5), 823–841. doi:10.1103/physrev.36.823.

van Otterlo, M. and Wiering, M. (2012). Reinforcement learning and Markov decision processes. In *Adaptation, Learning, and Optimization*, 3–42. Springer Berlin Heidelberg. doi:10.1007/978-3-642-27645-3.

Watkins, C.J.C.H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3-4), 279–292. doi: 10.1007/bf00992698.