Evaluating Meta-Reinforcement Learning through a HVAC Control Benchmark (Student Abstract)

Yashvir S. Grewal, Frits de Nijs, Sarah Goodwin
Monash University, Melbourne, Australia
ygre0001@student.monash.edu, {frits.nijs, sarah.goodwin}@monash.edu

Abstract

Meta-Reinforcement Learning (RL) algorithms promise to leverage prior task experience to quickly learn new unseen tasks. Unfortunately, evaluating meta-RL algorithms is complicated by a lack of suitable benchmarks. In this paper we propose adapting a challenging real-world heating, ventilation and air-conditioning (HVAC) control benchmark for meta-RL. Unlike existing benchmark problems, HVAC control has a broader task distribution, and sources of exogenous stochasticity from price and weather predictions which can be shared across task definitions. This can enable greater differentiation between the performance of current meta-RL approaches, and open the way for future research into algorithms that can adapt to entirely new tasks not sampled from the current task distribution.

Introduction

Yu et al. (2020) identify that the lack of realistic benchmarks and evaluation protocols hinders the progress of research in meta-RL methods. Current meta-RL approaches are typically evaluated on maze navigation tasks (Duan et al. 2016), locomotive tasks (Finn, Abbeel, and Levine 2017) and Bandit problems (Ritter et al. 2018). However, the task distributions of these benchmarks are too narrow, as the tasks are structurally similar and synthetic, lacking realistic uncertainty sources. In turn, this means that these benchmarks do not challenge algorithms to learn a shared structure across a wide task distribution. Thus, it becomes difficult to differentiate the performance of existing algorithms, which limits the development of new algorithms. This paper proposes a new meta-HVAC benchmark with a broader task distribution, that addresses the weaknesses in current meta-RL benchmarks. It is more aligned with real-world problems because it includes realistic stochasticity that we expect meta-RL algorithms to generalize over.

Meta-HVAC

The proposed meta-HVAC benchmark primarily responds to the problem of heating buildings in the context of fluctuating energy prices. We can frame this problem as a RL problem, where an agent controls a central-heating system, and has to make decisions about thermostat set points. The agent has to respond to a variety of exogenous factors. One factor is the real-time energy market, which typically has long periods of flat, stable low prices, interspersed with short periods where the price can spike to 3-4 times the average. Accordingly, consumers risk having to pay extremely high prices during periods of peak price. Other considerations the agent must regard are the occupancy levels of a building, as well as any grid limitations on power-supply. However, since RL-algorithms are highly sample-inefficient, they typically need to be trained on very accurate simulations of target buildings. This, however, is highly expensive, time-consuming and inefficient, rendering it infeasible to be deployed on a wider scale. A solution is to cast this problem as a meta-RL problem, whereby the agent is trained on a variety of simulated buildings and tasks. This enables the agent to learn a shared structure across buildings and tasks, which it can leverage to learn novel, unseen tasks more efficiently. This reduces the need for a simulator specifically tailored to the target building, making it more feasible for real-world application. Relevantly, the meta-HVAC benchmark can be used to train the agent on a variety of tasks. Most importantly, because this benchmark contains a broader task distribution, and is more closely aligned with real-world circumstances, it corrects deficiencies in existing meta-RL benchmarks.

Model	Price	Occupancy	Grid limits
1st-order	Flat tariff	Always	Unlimited
2nd-order	Day/night tariff	Scheduled	Planned
Multi-zone	Real-time det.	Predicted	On-line
Fluid dynamics	Real-time stoch.	-	-

Table 1: Identified variable factors of HVAC control tasks.

Meta-HVAC Tasks

Owing to its broadness and stochastic features, the meta-HVAC benchmark is an appropriate benchmark for evaluating meta-RL algorithms. In the meta-HVAC benchmark, the overarching task variables are model, price, occupancy and grid limits (as detailed in Table 1). Firstly, the different simulation **models** correspond to different transition dynamics of the Markov Decision Process (MDP), resulting in new tasks. Additionally, within each simulation model, different building configurations exist. Building configuration denotes the nature and size of the target enclosure. These configurations also determine the dynamics of the MDP, such that they can be classified as different tasks. In sum, the diversity of simulation models, combined with the different building configurations naturally broaden the benchmark's task distribution.

Secondly, choosing a different **price** variable results in a new objective for the agent. For example, selecting a 'flat tariff' price exclusively corresponds with the objective of maintaining a comfortable temperature. Contrastingly, a real-time stochastic price relates to the objective of balancing cost and comfort. Since this variable allows for different objectives, this expands the range of tasks, thereby broadening task distribution. Moreover, because the agent must respond to real-time stochastic prices, this introduces real-world uncertainty. The benchmark uses energy price data from AEMO[1], which varies across 5-minute intervals.

Thirdly, varying **occupancy** levels — always, scheduled and predicted occupancy — also influence the dynamics of the MDP, but only for a specific period during an episode. Lastly, **grid limits** refer to periods when electricity cannot be accessed, such as planned and sporadic instances of non-supply, as compared to unlimited supply. Because occupancy levels and grid limits are beyond the agent's control, this lends exogeneity to the benchmark, making it more realistic. Additionally, sampling from the four overarching variables—model, price, occupancy and grid limit—as well as selecting a specific building configuration within the simulation model, results in a broader task distribution.

Empirical Results

We use RL^2 (Duan et al. 2016; Wang et al. 2016), an LSTM-based on-policy meta-RL algorithm. The meta-training task distribution consisted of six building configurations of 1st-order models, and two building configurations of 2nd-order models, totalling eight buildings.[2] Three of the 1st-order buildings, and one of the 2nd-order buildings had a flat tariff price task. The remaining four buildings had a real-time stochastic price task. We sampled one unseen building from both the first and second order models, to calculate the mean reward over 50 episodes. The occupancy and grid limits were set at 'always' and 'unlimited', respectively, during both training and testing. Performance shown in Fig. 1, test-time learning curves in Fig. 2. For comparison, we used a task specific policy trained with ACKTR algorithm.

Conclusion and Future Work

At the time of writing this paper we are in the process of extending this benchmark to multi-task RL problems. Moreover, a critical assumption is that the training-task distribution should be the *same* as the testing-task distribution. However, an open problem is designing algorithms which do not rely upon this assumption. As shown in Fig. 1, when the agent's training-task distribution did not include any 2nd-order model buildings, it performed poorly when tested on a 2nd-order model building, more specifically when the task had a real-time stoch. price. Therefore, a future research endeavour is designing algorithms that do not require the test-task to be sampled from the same distribution as the training-task. Accordingly, the meta-HVAC benchmark paves the way for developing this class of algorithms.

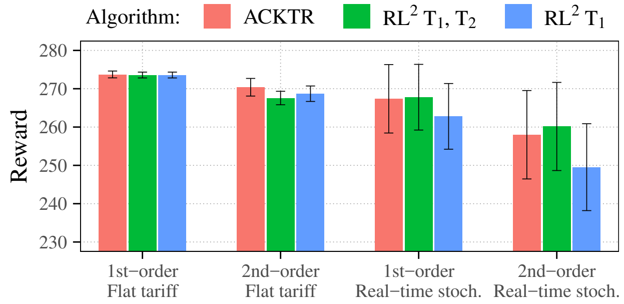

Figure 1: Per-task performance of the learned policy. Single-task learner ACKTR for reference against RL^2, trained with and without 2nd-order models in the training tasks.

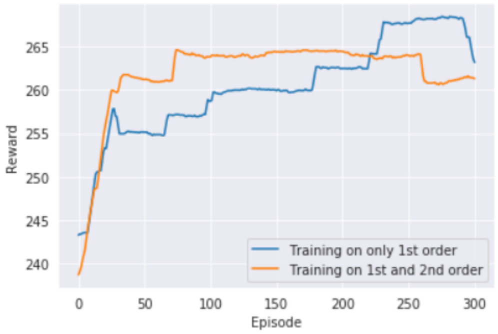

Figure 2: Learning curve of RL^2 on unseen 2nd-order model, when trained with and without 2nd-order models in tasks.

References

Duan, Y.; Schulman, J.; Chen, X.; Bartlett, P. L.; Sutskever, I.; and Abbeel, P. 2016. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. URL arXiv:1611.02779v2.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proc. 34th International Conference on Machine Learning*, vol. 70:1126–1135.

Ritter, S.; Wang, J.; Kurth-Nelson, Z.; Jayakumar, S.; Blundell, C.; Pascanu, R.; and Botvinick, M. 2018. Been There, Done That: Meta-Learning with Episodic Recall. In *Proc. Machine Learning Research*, vol. 80:4354–4363.

Wang, J. X.; Kurth-Nelson, Z.; Tirumala, D.; Soyer, H.; Leibo, J. Z.; Munos, R.; Blundell, C.; Kumaran, D.; and Botvinick, M. 2016. Learning to Reinforcement Learn. URL arXiv:1611.05763.

Yu, T.; Quillen, D.; He, Z.; Julian, R.; Hausman, K.; Finn, C.; and Levine, S. 2020. Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning. In *Proc. Machine Learning Research*, vol. 100:1094–1100.

[1] https://aemo.com.au/en/energy-systems/electricity/national-electricity-market-nem/data-nem/data-dashboard-nem

[2] Our code may be found at:
https://github.com/yashvirsinghgrewal-crypto/energyexperiments