A Comparison of Value-based and Policy-based Reinforcement Learning for Monitoring-informed Railway Maintenance Planning

GIACOMO ARCIERI, CYPRIEN HOELZL, OLIVER SCHWERY, DANIEL STRAUB, KONSTANTINOS G. PAPAKONSTANTINOU and ELENI CHATZI

ABSTRACT

Optimal maintenance planning for railway infrastructure and assets forms a complex sequential decision-making problem. Railways are naturally subject to deterioration, which can result in compromised service and increased safety risks and costs. Maintenance actions ought to be proactively planned to prevent the adverse effects of deterioration and the associated costs. Such predictive actions can be planned based on monitoring data, which are often indirect and noisy, thus offering an uncertain assessment of the railway condition. From a mathematical perspective, this forms a stochastic control problem under data uncertainty, which can be cast as a Partially Observable Markov Decision Process (POMDP). In this work, we model the real-world problem of railway optimal maintenance planning as a POMDP, with the problem parameters inferred from real-world monitoring data. The POMDP model serves to infer beliefs over a set of hidden states, which aim to capture the evolution of the underlying deterioration process. The maintenance optimization problem is here ultimately solved via the use of deep Reinforcement Learning (RL) techniques, which allow for a more flexible and broad search over the policy space when compared to classical POMDP solution algorithms. A comparison of value-based and policy-based RL methods is also offered, which exploit deep learning architectures to model either action-value functions (i.e., the expected returns from an action-state pair) or directly the policy. Our work shows how this complex planning problem can be effectively solved via deep RL to derive an optimized maintenance policy of railway tracks, demonstrated on real-world monitoring data, and offers insights into the solution provided by different classes of RL algorithms.

Giacomo Arcieri, PhD Student, Email: giacomo.arcieri@ibk.baug.ethz.ch. Department of Civil, Environmental and Geomatic Engineering, ETH Zurich, Zurich, Switzerland.
Cyprien Hoelzl. Department of Civil, Environmental and Geomatic Engineering, ETH Zurich, Zurich, Switzerland.
Oliver Schwery. Swiss Federal Railways SBB, Bern, Switzerland.
Daniel Straub. Engineering Risk Analysis Group, Technical University of Munich, Munich, Germany.
Konstantinos G. Papakonstantinou. Dept. of Civil and Environmental Engineering, Pennsylvania State Univ., PA, USA.
Eleni Chatzi. Department of Civil, Environmental and Geomatic Engineering, ETH Zurich, Zurich, Switzerland.

INTRODUCTION

Railway systems are subject to harsh environments and high repetitive loads, which lead to accelerated deterioration, resulting in a degraded level of service and increased risks and costs [1]. Maintenance actions, which may reflect various intervention intensities (e.g., minor/major repair or complete overhaul), must be planned throughout the operating life-cycle of the tracks, and in compliance with the current railway conditions, in order to prevent deterioration effects and associated costs [2,3]. The resulting optimal maintenance planning in railway systems forms a complex sequential decision-making problem.

The complexity is exacerbated due to the limited knowledge of the condition (state) of the railway infrastructure. This limitation is in part addressed by means of Structural Health Monitoring (SHM) tools [4–6], which however can only offer partial and/or noisy information about the actual condition of a system. For tackling such observation uncertainty, the problem can be cast into a Partially Observable Markov Decision Process (POMDP) framework [7]. POMDPs account for the incomplete knowledge of the system's states, thus termed *hidden states*, and the uncertainty associated with observations towards deriving optimal policies; namely a sequence of optimal decisions that minimize the total costs over a prescribed horizon, under stochastic environments.

However, POMDPs are notoriously difficult to solve and generally only approximate solution algorithms are available [8]. This difficulty can be attributed to the inherent uncertainty in the observations, which requires the decision maker to keep track of a history of past observations and actions for more efficient planning. Depending on the dimensionality of the observation and action spaces and the horizon of the problem, the complexity of the solution can abruptly escalate over time. This is referred to as the *curse of history* [7]. Under this premise, POMDPs would no longer satisfy the Markov property, which forms a main trait of Markov Decision Processes (MDPs) and allows for use of efficient solution algorithms with optimality guarantees.

To alleviate these problems, a common approach is the introduction of the *belief* variable b in the POMDP framework [9], namely a probability distribution over the hidden states inferred via Bayesian updates when new observations and actions are available. The belief reflects a sufficient statistics of the history of actions and observations, in the sense that knowledge of b_t bears an equivalent amount of information as the history of all past observations and actions until time t. Solving a POMDP is equivalent to solving the planning problem defined over the belief space, termed the belief-MDP [10]. While the belief-MDP can still be considered a complex problem to be solved, this tremendously simplifies the solution of the original POMDP and the Markov property is satisfied anew.

An efficient approach for the solution of POMDP problems, in the context of maintenance planning, has been introduced in Andriotis and Papakonstantinou [11]. The input space is mapped from the observations and actions to the belief variable, i.e., the POMDP problem is shifted into the belief-MDP, by means of a model of the environment and Bayes theorem. The beliefs are then used as inputs of neural networks, which solve the problem, i.e., learn an optimal policy, through deep Reinforcement Learning (RL) algorithms. This approach has led to the solution of particularly complex POMDP problems [11–13].

The choice of the deep RL algorithm is, however, still critical for the solution of

the problem. The spectrum of available algorithms is vast, comprising different families of methods. Focusing on model-free RL algorithms [14], i.e., excluding use of model-based RL methods [15], which would introduce further challenges in the context of partial observability, these can be divided into: i) value-based methods, which learn the expected value (return) from a given state of the problem and are limited to discrete control settings, and ii) policy-based methods, which directly parametrize the policy and search the optimal solution via gradient descent, enabling also the solution of continuous control tasks. The latter methods can be enhanced with the use of value networks (critic), inspired by value-based algorithms, to stabilize the policy network (actor) training, resulting into the actor-critic approach. However, even within the same family of methods, notoriously no algorithm is consistently better than the others on an extended range of tasks [16, 17]. In maintenance planning, where the action space is generally discrete, both families of methods are permissible, rendering the choice of the RL algorithm even more complex.

In this work, we model a real-world problem of railway optimal maintenance planning as a POMDP, with the problem parameters being inferred from real-world monitoring data. This has been accomplished in previous work [18, 19] by means of a hidden Markov model conditioned on actions. The model then serves to infer beliefs over the actual hidden states which are meant to reflect the best available knowledge over the state (condition) of the monitored railway infrastructure. In this case, we employ observations that allow to indirectly assess the condition of the rails and substructure. The maintenance optimization problem is here ultimately solved by deep RL techniques in order to provide the sought policy. A comparison of value-based and policy-based RL methods is also offered. In particular, a state-of-the-art algorithm is chosen for each class, namely Rainbow [20] and Policy Proximal Optimization (PPO) [21] for value-based and policy-based methods, respectively. Specifically, PPO is implemented with the clipped surrogate objective and the actor-critic style. Despite the discrete control setting, where value-based methods are often expected to exhibit better performance than policy-based methods, our results show that PPO is able to deliver a robust performance, outperforming Rainbow in the considered real-world problem.

BACKGROUND AND METHODS FUNDAMENTALS

Markov Decision Processes

MDPs [22] are formulated as sequential stochastic control problems defined by the tuple $\langle S, A, R, T, H, \gamma \rangle$. At a given time-step t, the decision maker (or agent) observes the state of the problem (or *environment*) $s_t \in S$, selects action $a_t \in A$, and receives the reward $r_t = R(s_t, a_t)$. The environment then transitions to the new state s_{t+1} according to the stochastic transition model T. The objective is to find the policy π^* which maximizes the expected sum of rewards, namely $\pi^* = \arg\max_\pi \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t r_t\right]$.

Value functions

Value-based methods are based on the value function $V^\pi(s_t)$ [23], which outputs the

expected sum of rewards of policy π from a certain state:

$$V^{\pi}(s_t) = \mathbb{E}_{a_t \sim \pi(s_t)} \left[R(s_t, a_t) + \gamma \sum_{s_{t+1} \in S} p(s_{t+1}|s_t, a_t) V^{\pi}(s_{t+1}) \right] \tag{1}$$

The state values of the optimal policy π^* are given by $V^{\pi^*}(s_t) = \max_{\pi} V^{\pi}(s_t)$. The expected sum of rewards of policy π from a certain state-action pair is provided by the Q-function, which can be written in the form of the Bellman equation [24]:

$$Q^{\pi}(s_t, a_t) = R(s_t, a_t) + \gamma \sum_{s_{t+1} \in S} p(s_{t+1}|s_t, a_t) Q^{\pi}(s_{t+1}, \pi(s_{t+1})) \tag{2}$$

From the expectation of the Q-function it is possible to retrieve the value-function, i.e., $V^{\pi}(s_t) = \mathbb{E}_{a_t \sim \pi(s_t)} Q^{\pi}(s_t, a)$. If a policy π coincides with the optimal policy π^*, then the maximization operator replaces the expectation in the previous formulas. Finally, the advantage value function estimates the advantage of selecting action a_t, instead of following the current policy π:

$$A^{\pi}(s_t, a_t) = Q^{\pi}(s_t, a_t) - V^{\pi}(s_t) \tag{3}$$

The current policy π can be recursively improved via reinforcement learning via Temporal Difference (TD) methods [25]:

$$Q^{\pi}(s_t, a_t) \leftarrow Q^{\pi}(s_t, a_t) + \alpha \delta \tag{4}$$

where α is the learning rate, $\delta = y - Q^{\pi}(s_t, a_t)$ are the TD errors, and y is the vector of targets. For instance, in the Q-learning algorithm $y = r_t + \gamma \max_a Q^{\pi}(s_{t+1}, a)$.

In deep RL algorithms, Q-values are parametrized by a neural network, characterized by a vector of parameters θ. In the seminal Deep Q-Network (DQN) [26] algorithm, the expectations are computed through a batch of tuples (s_t, a_t, r_t, s_{t+1}) collected on a replay buffer D and a second neural network θ' is used for the target Q-values. Namely, the DQN algorithm minimizes the following loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim D} \left[\left(r_t + \gamma \max_a Q_{\theta'}(s_{t+1}, a) - Q_{\theta}(s_t, a_t) \right)^2 \right] \tag{5}$$

Rainbow

Rainbow [20], which is considered the state-of-the-art value-based method, combines six different extensions into the DQN algorithm, leading to tremendously improved data efficiency and final performance. These extensions are summarized as follows:

- **Double Q-learning** allows to alleviate the maximization bias by decoupling the action selection from the estimation of its action value through two separate Q-networks.
- **Prioritized replay** boosts the learning by sampling batches from the replay buffer proportionally to the TD error, instead of uniform sampling.

- **Dueling networks** are used for parallel estimation of state values and advantage values, which are then combined by a special aggregator for more accurate action value estimates.
- **Multi-step learning** leads to significantly improved performance by exploiting n-step target returns instead of the original single step.
- **Distributional RL** is used to learn approximate distributions of state action values by discretizing the distribution in N "atoms", instead of single expected return estimates.
- **Noisy Nets** propose a noisy linear layer, whose noise degree is learned during training, to favour state-conditional exploration instead of fixed ϵ-greedy policies.

Proximal Policy Optimization

PPO [21] is considered the state-of-the-art policy-based method, namely a class of algorithms that directly parametrize the policy π_θ via neural networks with parameters θ. PPO is rooted in trust region methods, albeit comprising a tremendously simpler implementation. Contrary to other policy-based methods, which aim to keep the parameters of the updated policy network θ' close to θ via the learning step size of gradient descent, trust region methods [27] aim to keep the updated policy $\pi_{\theta'}$ close to π_θ in order to achieve monotonic improvements. This is achieved by adding a Kullback-Leibler (KL) divergence constraint, transforming the RL problem into a constrained optimization problem, which is non-trivial to implement.

PPO ensures the closeness of the new policy by simply optimizing the following (unconstrained) clipped surrogate objective:

$$\mathcal{L}^{CLIP}(\theta') = \mathbb{E}\left[\min\left(\frac{\pi_{\theta'}(s)}{\pi_\theta(s)}A^{\pi_\theta}, \text{clip}\left(\frac{\pi_{\theta'}(s)}{\pi_\theta(s)}, 1-\epsilon, 1+\epsilon\right)A^{\pi_\theta}\right)\right] \tag{6}$$

where clip$(\cdot)$ is a function that clips the ratio between new and old policies by the specified $(1-\epsilon, 1+\epsilon)$ bounds, and ϵ is the clipping hyperparameter. Despite the simpler implementation, PPO enables the largest possible improvement without the risk of performance collapse by bounding the permitted changes in the policy. As a result, PPO is known to be a robust algorithm, significantly reducing the detrimental variance that notoriously affects policy gradient methods. Finally, PPO trains a stochastic policy by predicting a probability distribution over actions. The latter are sampled from the learned distributions during training to promote exploration, while the best action (i.e., the one with highest probability) can be used at testing time.

Partially Observable Markov Decision Processes

POMDPs are a generalization of the MDP framework, with uncertainty incorporated into the observations and defined by the tuple $\langle S, A, Z, R, T, O, b_0, H, \gamma\rangle$. The state $s_t \in S$ is no longer accessed by the agent, who only receives the observation $z_t \in Z$ produced according to the stochastic observation model O. The agent thus forms its belief, which was initialized in b_0, with the new observation (and past action) according

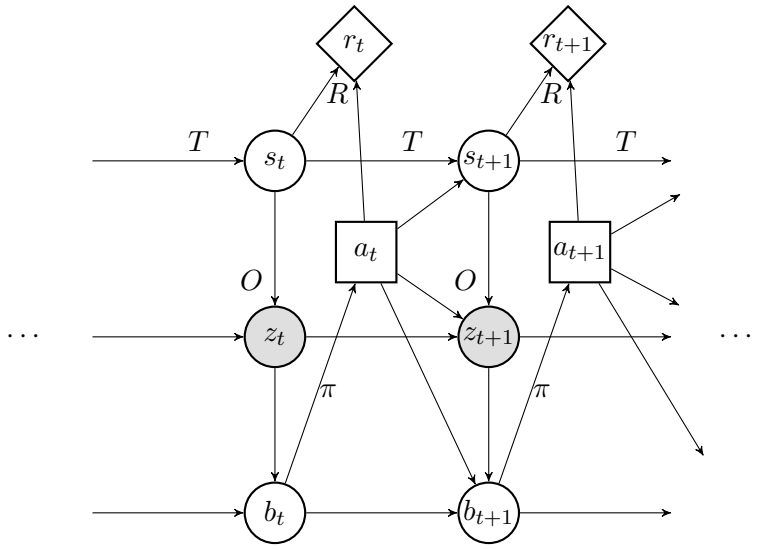

Figure 1. Probabilistic graphical model of a POMDP.

to Equation 7. The agent then selects action $a_t \in A$ based on the updated belief b_t. The objective is thus to find the optimal policy that maximizes the expected returns, which maps beliefs to actions. The sequential decision-making problem of the POMDP is displayed in the probabilistic graphical model in Figure 1.

$$\begin{aligned} b(s_t) &= \frac{p(z_t \mid s_t, a_{t-1})}{p(z_t \mid \mathbf{b}, a_{t-1})} \sum_{s_{t-1} \in S} p(s_t \mid s_{t-1}, a_{t-1}) b(s_{t-1}) \\ p(z_t \mid \mathbf{b}, a_{t-1}) &= \sum_{s_t \in S} p(z_t \mid s_t, a_{t-1}) \sum_{s_{t-1} \in S} p(s_t \mid s_{t-1}, a_{t-1}) b(s_{t-1}) \end{aligned} \tag{7}$$

In this work, the agent, represented by neural networks, does not receive the observations z, but only the beliefs b, which can thus be considered as directly computed by the environment. As such, the agent tackles a problem that can be considered as an MDP over the (continuous) belief space, i.e., the states s of the MDP framework coincide with the beliefs b.

THE RAILWAY OPTIMAL MAINTENANCE

In this work, we are concerned with the real-world problem of optimal maintenance. The railway track, displayed in Figure 2, is composed of a superstructure (rails, sleepers, and ballast) and the substructure. The latter fulfills several functions, such as sustaining the superstructure, acting as a filter to prevent infiltration of fine material, and favoring the water runoff from the track. As such, the substructure plays a key role in the degradation process of the track. In addition, replacement of the substructure costs about twice the amount of a superstructure renewal [1], significantly affecting the life-cycle costs of the track.

Accurate maintenance of the substructure is key to extend the lifetime of the track, reduce the operating costs, and ensure an adequate service. When the deterioration level

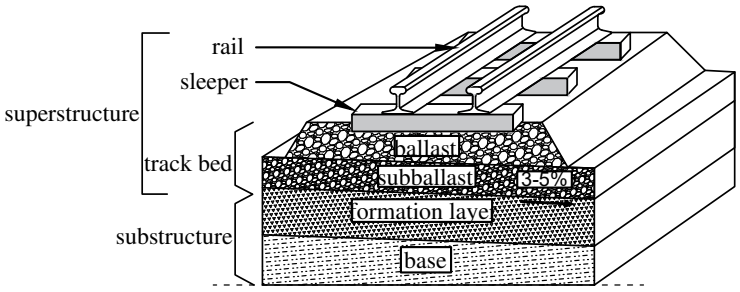

Figure 2. Structure of the railway track.

is estimated as moderate, a so-called tamping maintenance action is often applied. Otherwise, if the substructure is in poor condition, a replacement action is instead deemed as suitable.

The direct observation of the condition of a complex system, such as a railway network, is a task with inherent uncertainties. The common approach of visual inspection offers more of a qualitative and often biased assessment. More recently, railway managers exploit indirect information that can be extracted via use of SHM tools. In this work, we rely on use of the *fractal value* indicator [28], which is computed from on-board monitoring measurements. A specialized diagnostic vehicle carries out the measurement of the so-called longitudinal level, which essentially translates into an indicator of the regularity of the rail profile. The fractal value is then computed via filtering these measurements within pre-specified wavelength ranges. In this work, we employ the long-wave (25-70 m) fractal values, which have been shown to comprise an indirect indicator of substructure deterioration [1] and are used by railway authorities to support track maintenance [28]. Through a collaboration with the Swiss Federal Railways (SBB), we have access to almost 10 years of fractal values collected over the Swiss railway network. In addition, the ZMON (ZustandsMONitoring—condition monitoring) database of the SBB offers access to logs where remedial actions [29], including tamping and renewal, are recorded over the examined time period for the investigated tracks.

While fractal values constitute a valid indicator for decision support, it should be noted that these do not directly measure the substructure deterioration but only form a correlation therewith. In addition, the collected data is affected by noise, associated with measurement imprecision and disturbances, and possible errors during data storage. As such, the railway optimal maintenance forms a natural application of the POMDP modeling, where the uncertain observations are represented by the fractal values, which are defined over the continuous negative set of real numbers. The hidden states are discretized in 4 possible ground truth railway conditions. Three possible actions are considered, namely the action do-nothing a_0, and the aforementioned tamping and renewal maintenance actions a_1 and a_2, respectively. The (negative) rewards represent the costs associated with actions and states and are reported in Table I in general cost units. The transition dynamics T and the observation generating process O have been inferred from the aforementioned real-world collected data in Arcieri et al. [18, 19] to build the POMDP model of the problem. Finally, one time-step of the problem equals 6

TABLE I. COSTS OF THE POMDP MODEL.

State condition	s_0	s_1	s_2	s_3
Maintenance action				
a_0	0	0	0	0
a_1	-50	-50	-50	-50
a_2	$-2,050$	$-2,710$	$-3,370$	$-4,050$
Condition cost	-100	-200	$-1,000$	$-8,000$

months, since this is the average frequency of the fractal value collection, and the considered horizon of the planning problem is 50 time-steps. The defining problem traits have been elicited from the SBB (the Swiss Federal Railways) so as to adhere to real-world characteristics.

EVALUATION

The problem of optimal planning of railway optimal maintenance under data uncertainty, herein modelled as a POMDP, is solved by means of the RL methods previously described, namely PPO and Rainbow. The hyperparameters have been optimized via a thorough grid-search, with the best performing values reported in Table II.

For both algorithms, 2 million training time-steps are sampled from the environment (i.e., 40,000 training episodes). Every 4,000 training time-steps, network/policy updates are performed on the collected data. Every 5 updates, an evaluation iteration of the best learned policy (i.e., without stochasticity of the policy) is performed over 100,000 testing episodes. Summary statistics of the best evaluation iteration are reported for both algorithms in Table III. In addition, the optimal solution computed by assuming full observability of the hidden states is also reported as an upper bound of the best possible solution that can be achieved under data uncertainty.

Although value-based methods are usually often expected to yield better performance than policy-based methods on discrete control tasks, PPO shows a robust and

[1]Intervals of [time-step, learning rate]

TABLE II. BEST HYPERPARAMETERS FROM THE GRID-SEARCH OPTIMIZATION.

Hyperparmeter	PPO	Rainbow
Hidden layers	2	2
Hidden size	256	256
Learning rate	1e-4	[[0, 1e-3], [200000, 1e-4], [2000000, 1e-5]][1]
Activation	ReLu	ReLu
Clip parameter	0.1	-
Number atoms	-	100
V_{min}	-	-25,000
V_{max}	-	-10,000
Target update frequency	-	50

TABLE III. PERFORMANCE OF THE BEST MODEL INFERRED DURING THE TRAINING PROCESS, EVALUATED OVER 100,000 SIMULATIONS.

Method	Avg. performance	SE	Max	Min
Optimal MDP	-13,315	27	-5,000	-93,980
PPO	-14,519	36	-5,050	-119,750
Rainbow	-14,830	37	-5,050	-127,400

significantly better solution than Rainbow on this optimal maintenance problem. As such, it delivers a maintenance planning policy of railway assets, based on the fractal value indicator, which is not far from the optimal solution of the fully observable problem, which is based on the (unrealistic) assumption of directly observing the deterioration state s_t.

Finally, Figure 3 shows two trials of the maintenance actions planned by the RL agent trained with PPO (left) and the one trained with Rainbow (right). From bottom to top: the observations (fractal values); the beliefs computed via Bayes' formula and used as inputs by the RL agents; the true hidden states, which are not accessed by the agent and/or the model; the actions planned by the RL agents.

CONCLUSIONS

In this work, a railway optimal maintenance problem is modelled as a POMDP. The problem, whose transition dynamics and observation generating process have been inferred from real-world monitoring data, is solved by transforming the POMDP into the

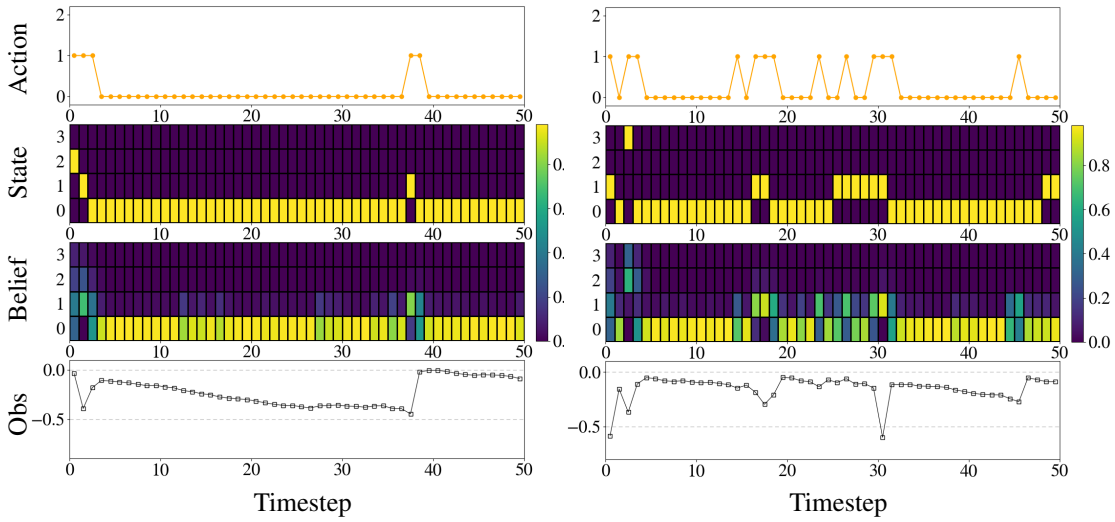

Figure 3. Policy simulations of the maintenance actions planned by the RL agents trained with PPO (left) and Rainbow (right). From bottom to top: the observations (fractal values); the beliefs, namely a probability distribution over hidden states, computed via Bayes' formula and used as inputs by the RL agents; the true hidden states, which are not accessed by the agent and/or the model; the actions planned by the RL agents.

equivalent belief-MDP problem, which is ultimately solved via deep RL techniques. A comparison of the solution offered by two RL algorithms is presented, namely PPO and Rainbow, which form state-of-the-art solutions for policy-based and value-based RL methods, respectively. Despite the discrete control setting, which is often expected to favor value-based algorithms, PPO delivers a robust and superior performance to Rainbow.

The solution methods presented here focused on model-free RL schemes. Future work will explore deep model-based RL solutions [15], involving different algorithms and new challenges to be addressed, in particular how to reliably learn the POMDP structure under data uncertainty.

ACKNOWLEDGMENT

The authors wish to acknowledge the support of the Swiss Federal Railways (SBB) as part of the ETH Mobility Initiative project REASSESS.

REFERENCES

1. Hoelzl, C., V. Dertimanis, E. Chatzi, D. Winklehner, S. Züger, and A. Oprandi. 2021. "Data driven condition assessment of railway infrastructure," in *Bridge Maintenance, Safety, Management, Life-Cycle Sustainability and Innovations*, CRC Press, pp. 3251–3259.
2. Neuhold, J., M. Landgraf, S. Marschnig, and P. Veit. 2020. "Measurement data-driven lifecycle management of railway track," *Transportation Research Record*, 2674(11):685–696.
3. Landgraf, M., M. Zeiner, D. Knabl, and F. Corman. 2022. "Environmental impacts and associated costs of railway turnouts based on Austrian data," *Transportation Research Part D: Transport and Environment*, 103:103168.
4. Farrar, C. R. and K. Worden. 2012. *Structural health monitoring: a machine learning perspective*, John Wiley & Sons.
5. Kamariotis, A., E. N. Chatzi, and D. Straub. 2022. "Value of information from vibration-based structural health monitoring extracted via Bayesian model updating," *Mechanical Systems and Signal Processing*, 166:108465.
6. Kamariotis, A., E. Chatzi, and D. Straub. 2023. "A framework for quantifying the value of vibration-based structural health monitoring," *Mechanical Systems and Signal Processing*, 184:109708.
7. Papakonstantinou, K. G. and M. Shinozuka. 2014. "Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part I: Theory," *Reliability Engineering & System Safety*, 130:202–213.
8. Papakonstantinou, K. G. and M. Shinozuka. 2014. "Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part II: POMDP implementation," *Reliability Engineering & System Safety*, 130:214–224.
9. Kaelbling, L. P., M. L. Littman, and A. R. Cassandra. 1998. "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, 101(1-2):99–134.
10. Andriotis, C. P., K. G. Papakonstantinou, and E. N. Chatzi. 2021. "Value of structural health information in partially observable stochastic environments," *Structural Safety*, 93:102072.
11. Andriotis, C. P. and K. G. Papakonstantinou. 2019. "Managing engineering systems with large state and action spaces through deep reinforcement learning," *Reliability Engineering & System Safety*, 191:106483.

12. Andriotis, C. P. and K. G. Papakonstantinou. 2021. "Deep reinforcement learning driven inspection and maintenance planning under incomplete information and constraints," *Reliability Engineering & System Safety*, 212:107551.
13. Morato, P. G., C. P. Andriotis, K. G. Papakonstantinou, and P. Rigo. 2023. "Inference and dynamic decision-making for deteriorating systems with probabilistic dependencies through Bayesian networks and deep reinforcement learning," *Reliability Engineering & System Safety*:109144.
14. Arulkumaran, K., M. P. Deisenroth, M. Brundage, and A. A. Bharath. 2017. "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, 34(6):26–38.
15. Arcieri, G., D. Wölfle, and E. Chatzi. 2021. "Which Model to Trust: Assessing the Influence of Models on the Performance of Reinforcement Learning Algorithms for Continuous Control Tasks," *arXiv preprint arXiv:2110.13079*, doi:10.48550/arXiv.2110.13079.
16. Duan, Y., X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. 2016. "Benchmarking deep reinforcement learning for continuous control," in *International Conference on Machine Learning*, PMLR, pp. 1329–1338.
17. Henderson, P., R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. 2018. "Deep reinforcement learning that matters," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32.
18. Arcieri, G., C. Hoelzl, O. Schwery, D. Straub, K. G. Papakonstantinou, and E. Chatzi. 2023. "Bridging POMDPs and Bayesian decision making for robust maintenance planning under model uncertainty: An application to railway systems," *Reliability Engineering & System Safety*:109496.
19. Arcieri, G., C. Hoelzl, O. Schwery, D. Straub, K. G. Papakonstantinou, and E. Chatzi. 2023. "POMDP inference and robust solution via deep reinforcement learning: An application to railway optimal maintenance," *arXiv preprint arXiv:2307.08082*, doi:10.48550/arXiv.2307.08082.
20. Hessel, M., J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver. 2018. "Rainbow: Combining improvements in deep reinforcement learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32.
21. Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*.
22. Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons.
23. Bertsekas, D. 2012. *Dynamic programming and optimal control: Volume I*, vol. 1, Athena scientific.
24. Bellman, R. 1966. "Dynamic programming," *Science*, 153(3731):34–37.
25. Sutton, R. S. and A. G. Barto. 2018. *Reinforcement learning: An introduction*, MIT press.
26. Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. 2015. "Human-level control through deep reinforcement learning," *Nature*, 518(7540):529–533.
27. Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz. 2015. "Trust region policy optimization," in *International Conference on Machine Learning*, PMLR, pp. 1889–1897.
28. Landgraf, M. and F. Hansmann. 2019. "Fractal analysis as an innovative approach for evaluating the condition of railway tracks," *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 233, ISSN 20413017.
29. Hoelzl, C., G. Arcieri, L. Ancu, S. Banaszak, A. Kollros, V. Dertimanis, and E. Chatzi. 2023. "Fusing Expert Knowledge with Monitoring Data for Condition Assessment of Railway Welds," *Sensors*, 23(5):2672.