Inference and maintenance planning of monitored structures through Markov chain Monte Carlo and deep reinforcement learning

Lathourakis, Christos; Andriotis, Charalampos; Cicirello, Alice

14th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP14
Dublin, Ireland, July 9-13, 2023

Inference and Maintenance Planning of Monitored Structures through Markov Chain Monte Carlo and Deep Reinforcement Learning

Christos Lathourakis
Research Assistant, Faculty of Civil Engineering, Delft University of Technology, Netherlands

Charalampos Andriotis
Assistant Professor, Faculty of Architecture & the Built Environment Delft University of Technology, Netherlands

Alice Cicirello
Associate Professor, Faculty of Civil Engineering, Delft University of Technology, Netherlands

ABSTRACT: A key computational challenge in maintenance planning for deteriorating structures is to concurrently secure (i) optimality of decisions over long planning horizons, and (ii) accuracy of real-time parameter updates in high-dimensional stochastic spaces. Both are often encumbered by the presence of discretized continuous-state models that describe the underlying deterioration processes, and the emergence of combinatorial decision spaces due to multi-component environments. Recent advances in Deep Reinforcement Learning (DRL) formulations for inspection and maintenance planning provide us with powerful frameworks to handle efficiently near-optimal decision-making in immense state and action spaces without the need for offline system knowledge. Moreover, Bayesian Model Updating (BMU), aided by advanced sampling methods, allows us to address dimensionality and accuracy issues related to discretized degradation processes. Building upon these concepts, we develop a joint framework in this work, coupling DRL, more specifically deep Q-learning and actor-critic algorithms, with BMU through Hamiltonian Monte Carlo. Single- and multi-component systems are examined, and it is shown that the proposed methodology yields reduced lifelong maintenance costs, and policies of high fidelity and sophistication compared to traditional optimized time- and condition-based maintenance strategies.

1. INTRODUCTION

Maintenance planning for deteriorating systems exposed to corrosive environments, e.g. coastal, marine, highly acidic conditions, is essential for resource-efficient management of structural risks. The most beneficial sequence of maintenance decisions strikes the best balance between life-cycle intervention costs and expected failure losses. This can be sought as the solution to an optimization problem, characterized by high complexity due to non-stationary environment dynamics, data and model uncertainties, and non-periodic actions over long planning horizons and multiple components. Therefore, classic and state-of-the-art threshold-based approaches (Frangopol et al. (1997)); Bayesian networks with risk-based thresholds (Straub and Faber (2005), Luque and Straub (2019)); renewal processes (Castanier et al. (2005)); and evolutionary optimization schemes for single- and multi-objective configurations (Yang and Frangopol (2019), Unal and Warn (2017)) manifest several limitations as they often rely on static optimization formulations and can not easily capture and control combinatorial system-level component interactions in a closed-loop fashion. Recently, Deep Reinforcement Learning (DRL) agent-based frameworks combined with Partially Observable Markov Decision Processes (POMDPs) princi-

ples, provided unmatched capabilities to tackle such problems in high-dimensional multi-component systems (Andriotis and Papakonstantinou (2019), Andriotis and Papakonstantinou (2021), Morato et al. (2023)). Large state-spaces are reparameterized through deep neural network architectures, and near-optimal strategies are derived by letting the agent(s) directly interact with the environment. To handle uncertainty in the deterioration process, current approaches successfully relied on environment dynamics described by dynamic Bayesian networks with discrete random variables (Morato et al. (2022), Luque and Straub (2019)) leading to closed-form updating strategies of the latent uncertain system parameters. However, this approximation can lead to inaccuracies since fine discretization is not always computationally feasible, whereas coarse discretization cannot capture the true nature of the deterioration process. Bayesian inference, aided by sampling techniques, is able to account for a continuous deterioration model. In particular, model updating techniques can update our knowledge of the uncertainties of latent variables, by combining physics-based models, and measurements carrying information on the uncertain system parameters (Lye et al. (2021), Kamariotis et al. (2022)).

In this work, an integrated framework is developed that combines DRL and BMU to determine an optimal sequence of maintenance decisions over the lifespan of continuously monitored deteriorating engineering systems. This approach exploits physics-based models of the system and of the deterioration process, and information obtained via vibration-based monitoring. Single- and multi-agent DRL architectures are considered, trained through Double Deep Q-Network (DDQN) and Proximal Policy Optimization (PPO). The updating of the uncertain continuous-valued system parameters is performed through Hamiltonian Monte Carlo (HMC) with No U-turn Sampling (NUTS). The applicability of the proposed workflow is investigated by considering a mass-spring system and a multi-component structural frame. It is then compared against optimized time- and condition-based heuristic approaches, with the obtained results confirming that the integrated framework can yield solutions of significant cost-efficiency and policy sophistication.

2. BACKGROUND

2.1. Sequential decision-making processes

2.1.1. Partially Observable Markov Decision Processes (POMDPs)

The problem of optimal stochastic control is handled using POMDPs, which address the uncertainties of planning maintenance strategies, including uncertain action outcomes and observations in a unified framework (Corotis et al. (2005), Papakonstantinou and Shinozuka (2014)). The basic POMDP components are the environment and the agent. In each decision step t, the decision-maker, cannot observe the exact state of the system, $s_t \in \mathcal{S}$, but forms a belief, b_t, i.e. a probability distribution over the system's states. Based on b_t it takes an action $a_t \in \mathcal{A}$, receives a reward $R_t(b_t, a_t) \in \mathcal{R}$ and an observation $o_{t+1} \in \mathcal{O}$, which is used to derive b_{t+1}, through a Bayesian update. The sequence of chosen actions defines the policy, π. An optimal policy aims to maximize the sum of the discounted rewards, i.e. the total return, G^π:

$$\begin{aligned} G^\pi &= R(b_t, a_t) + \ldots + \gamma^{T-t} R(b_T, a_T) \\ &= \sum_{i=t}^{T} \gamma^{i-t} R(b_i, a_i) \end{aligned} \tag{1}$$

where γ is a discount factor, denoting the importance of the current reward over future ones. The action-value function, $Q^\pi(b_t, a_t)$ is defined as the expected return over both $\mathcal{B}$ and $\mathcal{A}$.

$$Q^\pi(b_t, a_t) = \mathbb{E}_{b_{t+1}, a_{t+1}} \left[G^\pi \mid b_t, a_t \right] \tag{2}$$

Decomposing Q^π into the immediate reward and the discounted value of the successor belief-state leads to the following recursive formula:

$$Q^\pi(b_t, a_t) = R(b_t, a_t) + \gamma \mathbb{E}_{b_{t+1}, a_{t+1}} \left[Q^\pi(s_{t+1}, a_{t+1}) \right] \tag{3}$$

The value function $V^\pi(b_t)$ corresponds to the expected return starting from belief b_t and traversing life-cycle trajectories under policy π.

$$V^\pi(b_t) = \mathbb{E}_{a_t} \left[Q^\pi(b_t, a_t) \right] \tag{4}$$

During each decision step t, the next belief b_{t+1} is calculated using Bayes' rule:

$$\begin{aligned} b(s_{t+1}) &= \mathrm{P}(s_{t+1} \mid o_{t+1}, a_t, b_t) \\ &= \frac{\mathrm{P}(o_{t+1} \mid s_{t+1}, a_t)}{\mathrm{P}(o_{t+1} \mid b_t, a_t)} \sum_{s_t \in \mathcal{S}} \mathrm{P}(s_{t+1} \mid s_t, a_t)\, b(s_t) \end{aligned} \tag{5}$$

where $\sum_{s_t \in \mathcal{S}} \mathrm{P}(s_{t+1} \mid s_t, a_t)\, b(s_t)$ refers to the prior distribution, $\mathrm{P}(o_{t+1} \mid b_t, a_t)$ to the evidence, and $\mathrm{P}(o_{t+1} \mid s_{t+1}, a_t)$ to the likelihood function. Although POMDP solutions have been applied to investigate infrastructure maintenance, often along with point-based algorithms (Papakonstantinou et al. (2018), Andriotis et al. (2021)), they can face limitations when it comes to large and/or continuous action- and state-spaces. For continuous cases Eq (5) is transformed into:

$$b(s_{t+1}) = \frac{\mathrm{P}(o_{t+1} \mid s_{t+1}, a_t)}{\mathrm{P}(o_{t+1} \mid b_t, a_t)} \int_{s_t} \mathrm{P}(s_{t+1} \mid s_t, a_t)\, b(s_t) \mathrm{d}s_t \tag{6}$$

where P denotes probability density functions. The integral in Eq (6) contains the multiplication of two continuous distributions with no analytical expression, hence it can not be calculated in a closed form. The same integral appears in the calculation of the evidence, making it often intractable to define in a closed form, resorting to sampling techniques.

2.1.2. Deep Reinforcement Learning

Recently developed DRL approaches offer major computational advantages to this end. By using model-free methods there is no need to have an analytical description of the transition dynamics, and with the POMDP functions being reparameterized and expressed in terms of some parameters θ, the computational cost of large state- and action-spaces is alleviated. Two major families of DRL approaches are incorporated in this work: deep Q-learning and actor-critic schemes. The former is used for single-agent solutions whereas the latter is also used for multi-agent solutions related to multi-component environments.

In deep Q-networks the training aims to determine the parameters θ of the Q-function $Q(b_t, a_t \mid \theta)$ that minimize the loss function:

$$L(\theta) = \mathbb{E}_{b_t, a_t}\left[\left(\left(y_t - Q(b_t, a_t \mid \theta)\right)\right)^2\right] \tag{7}$$

with y_t being the target for decision step t:

$$\begin{aligned} y_t &= R(b_t, a_t) \\ &+ \gamma Q\left(b_{t+1}, \arg\max Q\left(b_{t+1}, a_{t+1} \mid \theta\right) \mid \theta^-\right) \end{aligned} \tag{8}$$

Parameters θ^- correspond to a target network, which takes the values of the original one with a predefined delay. Moreover, a replay buffer is used in deep Q-networks, where (b_t, a_t, R_t, b_{t+1}) tuples are stored and then used in batch training. Each tuple is potentially used in many weight updates, leading to the use of non-consecutive uncorrelated samples, hence, reducing the variance through the updates. Using both the target and the original network for computing y_t, with parameters θ^- and θ, respectively, is a key concept of DDQN that reduces instabilities and avoids over-optimistic value estimates.

Actor-critic methods are based on computing the policy gradient:

$$g_\theta = \mathbb{E}_{b_t, a_t}\left[\sum_{t \geq 0} \nabla_\theta \log \pi(a_t \mid b_t, \theta)\, Q^\pi(b_t, a_t)\right] \tag{9}$$

To reduce the variance of the sampling-based estimator of Eq (9), a baseline is subtracted from the Q-function, introducing the advantage value, which refers to how advantageous a specific action is at the given belief:

$$A(b_t, a_t) = R(b_t, a_t) + \gamma V(b_{t+1}) - V(b_t) \tag{10}$$

To compute all terms in Eq (9), two different neural networks are used, which act as policy (actor) and value (critic) approximators.

To avoid large policy updates, the trust region methods were implemented, whereas off-policy methods were proposed to address sample complexity. In our analysis, we test the PPO introduced in Schulman et al. (2017). Denoting $r_t(\theta)$ as the ratio of the new over the old policy, $\pi_\theta / \pi_{\theta_{\text{old}}}$, the policy is updated based on the clipped objective:

$$\begin{aligned} L^{CLIP}(\theta) = \mathbb{E}\Big[&\min\big(r_t(\theta) A_t, \\ &\mathrm{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon\right) A_t\big)\Big] \end{aligned} \tag{11}$$

2.2. Bayesian Model Updating

Model updating constitutes an inverse problem, where observations of the system's behavior are used to update unknown system properties using Bayes' rule. An extensive review of model updating about damage assessment, including BMU can be found in Simoen et al. (2015).

When dealing with continuous parameters, the posterior distribution, i.e. the distribution of the updated system parameters, can not be expressed in a closed form, but only implicitly, point-wise, using a Monte Carlo approach or a numerical integration scheme. To overcome this obstacle, advanced sampling methods have been developed (Lye et al. (2021)). Two of the most well-known Markov Chain Monte Carlo (MCMC) algorithms, namely Metropolis-Hastings and Gibbs sampling, often fail to converge to the posterior distribution, especially for continuous model parameters. HMC approaches sampling by using an auxiliary variable scheme and simulating Hamiltonian dynamics (Neal et al. (2011)). HMC is not widely used due to its dependence on user-defined tuning parameters, but it provides the foundation for NUTS, a state-of-the-art self-tuning sampling algorithm introduced in Hoffman et al. (2014). NUTS is integrated into the proposed workflow.

3. PROPOSED FRAMEWORK

The conceptual breakdown of the complete problem along with the integration of the various methods and the interaction of the different computational blocks are displayed in Figure 1.

The belief b_t contains partially observable information over the continuous deterioration states. It can be the bin values that form a discrete distribution, its statistical moments, or even the needed parameters to form a known continuous distribution.

A schematic representation of the framework, including the neural network architecture, is depicted in Figure 2, for a Q-function approximation method and a multi-agent actor-critic method.

For multi-component cases, Q-function approximation can become impractical due to the immense amount of possible actions that are formed in a combinatoric fashion. This is not the case for actor-critic algorithms, which return the probability distribution of the actions as an output, instead of the action-state value function. Thus, assuming that the actions of the system's components are conditionally independent, the policy derived from an actor-network, $\pi_\theta(\underline{a_t} \mid b_t)$, can be decomposed and expressed as the product of multiple policies which refer to each of the n components individually, $a_t^1, a_t^2, \dots a_t^n$ instead

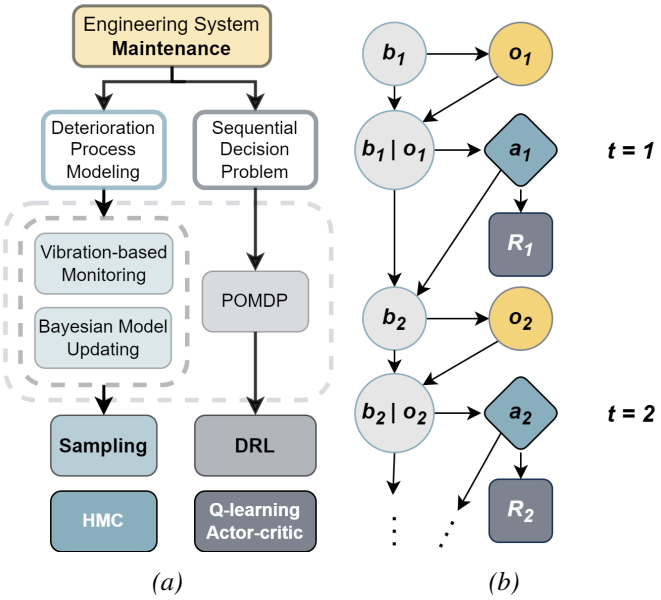

Figure 1: (a) Problem conceptual breakdown and (b) probabilistic POMDP decision graph.

of the full action vector, $\underline{a_t}$. We can, therefore, write:

$$\begin{aligned}\pi_\theta(\underline{a_t} \mid b_t) &= \pi_\theta(a_t^1 \mid b_t) \cdot \pi_\theta(a_t^2 \mid b_t) \dots \pi_\theta(a_t^n \mid b_t) \\ &= \prod_{i=1}^{n} \pi_\theta(a_t^i \mid b_t) \end{aligned} \tag{12}$$

The performance of the proposed framework is highlighted using benchmarks. A fine grid of deterioration thresholds and maintenance time intervals was exhaustively tested, to determine the most beneficial damage state and/or time to act.

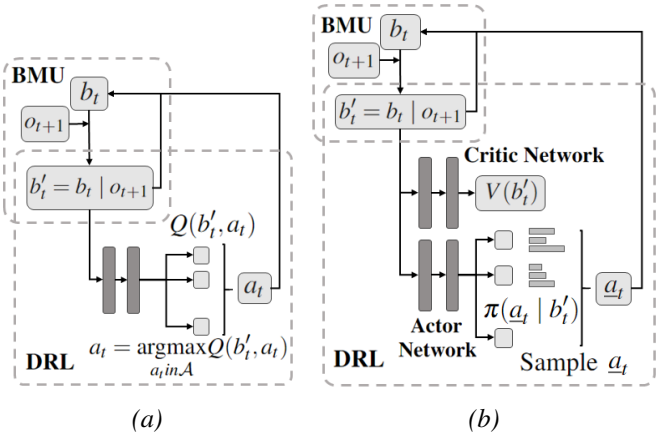

Figure 2: Neural network architectures for a single- and multi-component system (a) Q-function approximation for discrete action space (b) Actor-critic for centralized states and centralized discrete actions

4. NUMERICAL EXPERIMENTS

4.1. Single Degree of Freedom (SDOF) oscillator

The system's deteriorating parameter is the spring stiffness, with its deterioration for a given age, or deterioration rate, τ, following the power law:

$$D(\tau) = A\,\tau^B \tag{13}$$

$A \sim \ln\mathcal{N}(0.008, 0.004)$, $B \sim \mathcal{N}(1.5, 0.5)$, are responsible for the model's uncertainty, being updated during each step. As in Kamariotis et al. (2022), the stiffness $K(\tau)$ at a given deterioration rate, τ, is:

$$K(\tau) = \frac{K_0}{1+D(\tau)} = \frac{K_0}{1+A\,\tau^B} \tag{14}$$

A clear distinction should be made between the decision step t, i.e. the running time variable of the system's lifespan, increasing with a unit step, and the deterioration rate τ, which characterizes the system's age that can be affected by the agent's actions.

A monitoring system is assumed to provide the decision-maker with noisy measurements that can be used to extract the system's eigenfrequency for every decision step t. Numerically, these values are obtained by using sampled values for A, B. This measurement is further contaminated with Gaussian white noise, through the coefficient ε_{obs}, yielding the observation used for the updating process.

- Model output: $\omega(\tau) = \sqrt{\dfrac{K(\tau)}{m}} = \sqrt{\dfrac{K_0}{m\,(1+A\,\tau^B)}}$
- From measurement: $\hat{\omega} = \omega(\tau)|_{\text{sampled } A,B}$
- Observation: $O = \hat{\omega} + \mathcal{N}(0, \varepsilon_{\text{obs}} \cdot \hat{\omega})$

Three actions are considered, namely "*do nothing*", "*partial repair*", and "*replacement*". Regarding the deterioration rate τ, doing nothing does not affect it, a partial repair reduces it by two steps, and a total replacement resets it to zero. The reward that the agent receives in every decision step, i.e. the costs of maintenance, consists of the cost of the chosen action and the cost of the risk of failure.

$$R(b_t, a_t) = C_{a_t}(a_t) + C_{\text{risk}}(b_t) \tag{15}$$

The partial observability of the system's damage in every decision step is expressed by having continuous distributions instead of deterministic values or discrete distributions for the system parameters A, B, and subsequently for the deterioration $D(\tau)$.

4.1.1. Results

The DRL agents are trained with DDQN and PPO, obtaining Q-learning and actor-critic policies, respectively. Both approaches outperform benchmark decision rules, with the exact details about the costs and the achieved reduction shown in Table 1. Apart from the lower mean cost, the heuristic approach resulted also in a greater standard deviation, meaning that the stochasticity of the environment can lead to worse performance and higher maintenance costs when following a threshold-based policy.

Valuable conclusions are drawn from the realization of the learned policies. Due to the high stochasticity, each episode is considerably different, thus, multiple policies are plotted for both algorithms, in Figures 3, 4, along with optimal repair and replace thresholds derived through exhaustive realizations.

Policy trends are identified, highlighting the agent's ability to diverge from traditional strategies and act at unexpected deterioration stages. Actor-critic acted more consistently compared to deep Q-learning, with limited policies passing the replace heuristic value and most of the repair actions located lower than the heuristic's repair threshold. In deep Q-learning, the agent was stricter during the first decision steps, which aligns with real-life policies about brand-new components. The damage values lie mostly below the replace and even the repair threshold, especially when further in the examined time horizon. This difference between the two algorithms can be justified by the way each agent chooses actions. In DDQN the agent picks deterministically the most beneficial action, based solely on the action-state value functions $Q(a_t, b_t)$. Therefore, early interventions are not considered/explored as much by the decision-maker. However, the PPO agent, even for high damage values, chooses the action based on a probability distribution, i.e. $\pi(a_t \mid b_t)$, which allows any action, no matter how "good" or "bad" it is, to be picked, leading to more conservative on average, yet similarly optimal policies.

Table 1: Costs achieved after 50 policy realizations.

Algorithm	**Mean**	**St. Dev.**	**Decrease**
Benchmark	100.0%	100.0%	-
DDQN	79.0%	65.2%	21.0%
PPO	80.0%	57.3%	20.0%

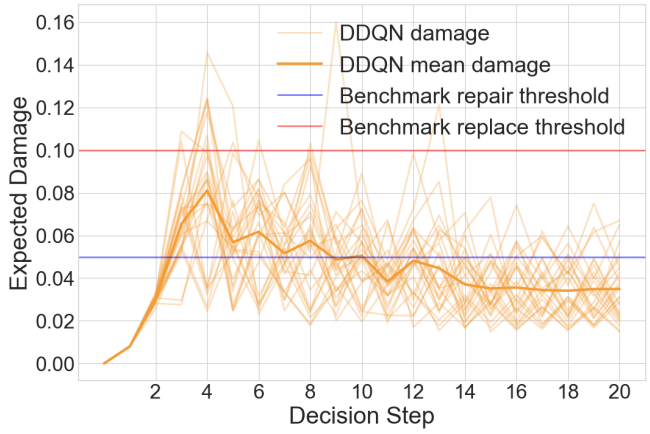

Figure 3: 50 deep Q-network policy realizations

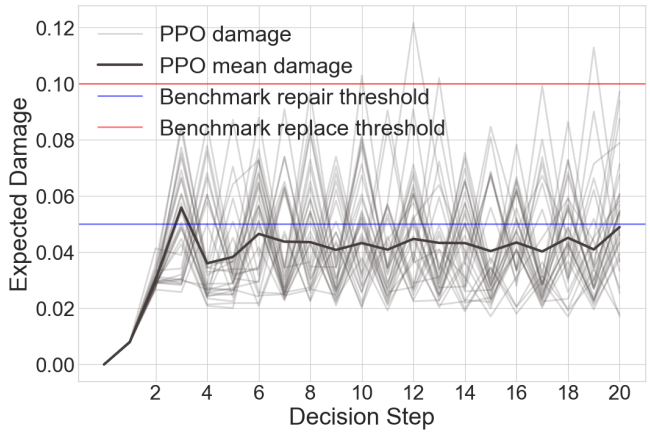

Figure 4: 50 actor-critic policy realizations

Another interesting outcome is the updating process toward discovering the true parameters of the environment (in this case $B_{\text{true}} = 2.0$). In Figure 5, the evolution of parameter B is plotted for 4 different policy realizations based on near-optimal weights. Incorporating more observations generated using the "true" values for B reduces the uncertainty, with the inferred B converging to the ground truth over time with a decreasing variance.

Figure 5: Updating parameter B in inference time.

4.2. Three-storey 2D frame

A multi-component structure is now considered, i.e. the three-storey plane frame illustrated in Figure 6. The structure is exposed to a corrosive environment, that causes section losses to the 6 vertical elements of the frame. The damage increment of each component for every deterioration rate τ_i, follows a Gamma process of shape $v(\tau_i)$ and scale u:

$$\Delta D_i \sim \text{Ga}(v(\tau_i) - v(\tau_i - 1), u)\,, \quad i = 1, \ldots, 6 \quad (16)$$

The scale u is assumed known and constant for all components, while the shape $v(\tau)$ is a stochastic parameter described by the power law $v(\tau) = A\,\tau^B$. A, B are random variables (as defined for the SDOF case) shared by all components. In each decision step t, these are sampled from the distributions $\text{P}(A), \text{P}(B)$, for every component, leading to different Gamma step distributions. Thus, the total damage for each component at any given decision step is the sum of all the Gamma distributions of the intermediate damage increments. It is known that the sum of gamma (v_t, u) random variables has a gamma $(\sum v_t, u)$ distribution, hence:

$$D_{t,i}^{\text{tot}} \sim \sum_{\tau=1}^{t} \text{Ga}\big(\cdot \mid A_{\tau,i}\,\tau^{B_{\tau,i}} - A_{\tau,i}\,(\tau-1)^{B_{\tau,i}}, u\big)$$
$$= \text{Ga}\big(\cdot \mid \sum_{\tau=1}^{t} A_{\tau,i}\,\tau^{B_{\tau,i}} - A_{\tau,i}\,(\tau-1)^{B_{\tau,i}}, u\big) \quad (17)$$

The partially observable state contains the shape parameter $v_{t,i}$ and the deterioration rate τ_i for every i component. Similarly to the SDOF case, the first

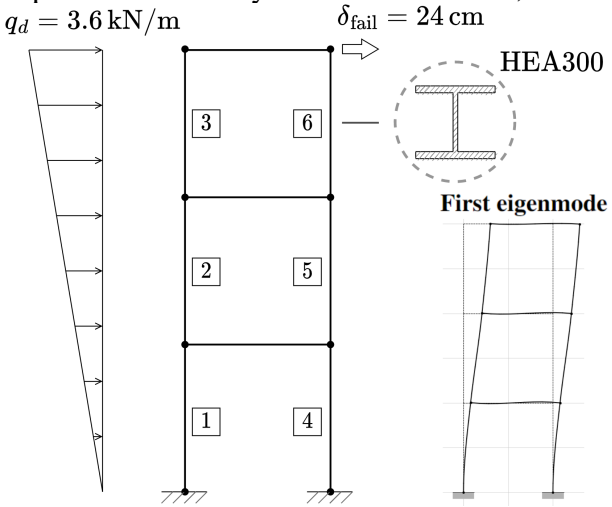

Figure 6: Three-storey plane frame

eigenmode extracted from measurements is used as an observation. Its similarity with the eigenmode of the deteriorated frame, results in the updating of the stochastic parameters A, B, through NUTS. Sampled damage values are used to numerically create the noisy measurements, which are further contaminated with noise to create the observations, i.e. the modal displacements. The same actions, as in the SDOF case, are considered, referring to each component, composing the 6×1 action vector, $\underline{a_t}$.

4.2.1. Results

Despite the underlying uncertainties that lead to a different deterioration for every episode, an indicative policy realization is presented in Figure 7. The agent is stricter when it comes to the deterioration of the base columns, which is a reasonable strategy since the middle and upper columns contribute less to the global failure of the frame.

Lastly, in Figure 8, policy realizations are plotted for different levels of training. The agent initially allows the deterioration to grow significantly, following completely uninformed strategies. This trend changes over the course of training episodes, with the agent ending up limiting the damage to lower values, thereby reducing the risk of failure and subsequently the total maintenance cost.

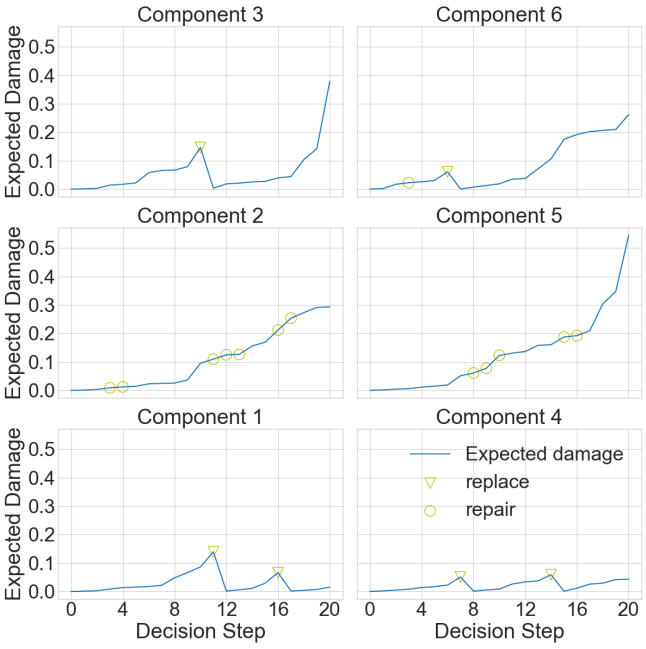

Figure 7: Policy realization for all components

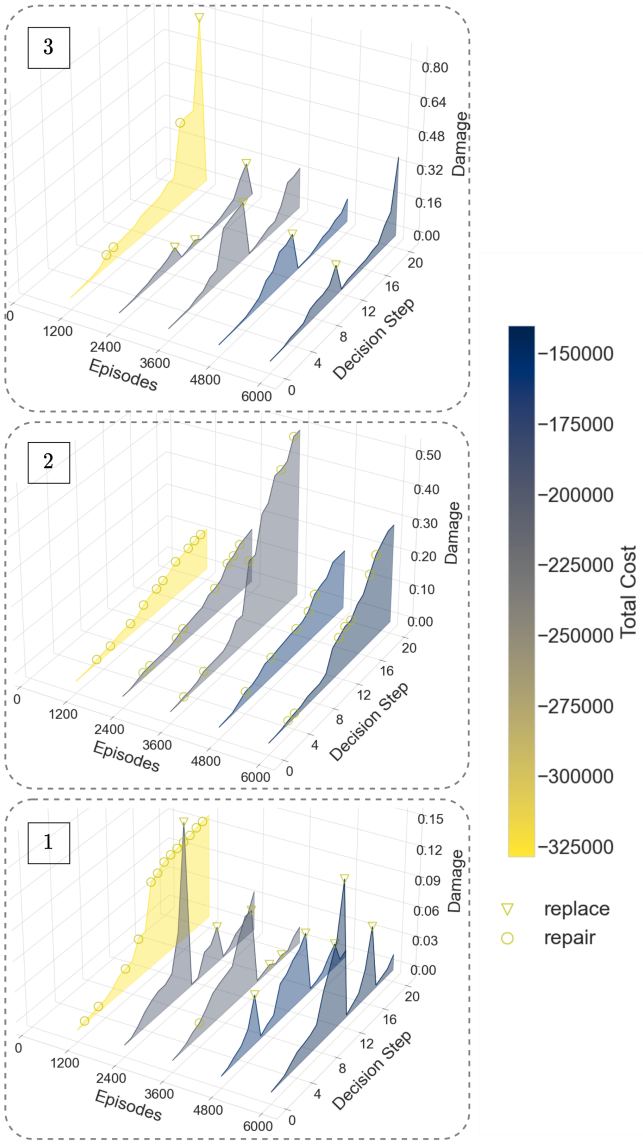

Figure 8: Policy realizations for all components for different training episodes

5. CONCLUSIONS

A joint Bayesian Model Updating (BMU) and Deep Reinforcement Learning (DRL) framework is developed to determine optimal maintenance strategies. Hamiltonian Monte Carlo (HMC) with No U-turn Sampling (NUTS) is coupled with two different DRL approaches, namely deep Q-learning and actor-critic, which are both able to outperform traditional maintenance approaches in terms of maintenance costs and policy sophistication over the structural service life. Continuous inference through NUTS was incorporated instead of simplified state discretization assumptions, thus embracing utmost modeling fidelity for the studied prob-

lems. Bayesian updating was identified as the main computational bottleneck in the joint maintenance planning and parameter inference problem, and not DRL. Further advances in Bayesian inference techniques, therefore, control the biggest lever in boosting the overall sequential optimization process, even when advanced DRL algorithms are involved.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the TU Delft AI Labs program.

7. REFERENCES

Andriotis, C. and Papakonstantinou, K. (2019). "Managing engineering systems with large state and action spaces through deep reinforcement learning." *Reliability Engineering & System Safety*, 191, 106483.

Andriotis, C. and Papakonstantinou, K. (2021). "Deep reinforcement learning driven inspection and maintenance planning under incomplete information and constraints." *Reliability Engineering & System Safety*, 212, 107551.

Andriotis, C. P., Papakonstantinou, K. G., and Chatzi, E. N. (2021). "Value of structural health information in partially observable stochastic environments." *Structural Safety*, 93, 102072.

Castanier, B., Grall, A., and Bérenguer, C. (2005). "A condition-based maintenance policy with non-periodic inspections for a two-unit series system." *Reliability Engineering & System Safety*, 87(1), 109–120.

Corotis, R. B., Hugh Ellis, J., and Jiang, M. (2005). "Modeling of risk-based inspection, maintenance and life-cycle cost with partially observable markov decision processes." *Structure and Infrastructure Engineering*, 1(1), 75–84.

Frangopol, D. M., Lin, K.-Y., and Estes, A. C. (1997). "Life-cycle cost design of deteriorating structure." *Journal of structural engineering*, 123(10), 1390.

Hoffman, M. D., Gelman, A., et al. (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.*, 15(1), 1593–1623.

Kamariotis, A., Chatzi, E., and Straub, D. (2022). "Value of information from vibration-based structural health monitoring extracted via bayesian model updating." *Mechanical Systems and Signal Processing*, 166, 108465.

Luque, J. and Straub, D. (2019). "Risk-based optimal inspection strategies for structural systems using dynamic bayesian networks." *Structural Safety*, 68–80.

Lye, A., Cicirello, A., and Patelli, E. (2021). "Sampling methods for solving Bayesian model updating problems: A tutorial." *Mechanical Systems and Signal Processing*, 159, 107760.

Morato, P., Papakonstantinou, K., Andriotis, C., Nielsen, J., and Rigo, P. (2022). "Optimal inspection and maintenance planning for deteriorating structural components through dynamic Bayesian networks and markov decision processes." *Structural Safety*, 94, 102140.

Morato, P. G., Andriotis, C. P., Papakonstantinou, K. G., and Rigo, P. (2023). "Inference and dynamic decision-making for deteriorating systems with probabilistic dependencies through Bayesian networks and deep reinforcement learning." *Reliability Engineering & System Safety*, 109144.

Neal, R. M. et al. (2011). "MCMC using Hamiltonian dynamics." *Handbook of markov chain monte carlo*, 2(11), 2.

Papakonstantinou, K. G., Andriotis, C. P., and Shinozuka, M. (2018). "POMDP and MOMDP solutions for structural life-cycle cost minimization under partial and mixed observability." *Structure and Infrastructure Engineering*, 14(7), 869–882.

Papakonstantinou, K. G. and Shinozuka, M. (2014). "Planning structural inspection and maintenance policies via dynamic programming and markov processes. part i: Theory." *Reliability Engineering & System Safety*, 130, 202–213.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). "Proximal policy optimization algorithms." *arXiv preprint arXiv:1707.06347*.

Simoen, E., De Roeck, G., and Lombaert, G. (2015). "Dealing with uncertainty in model updating for damage assessment: A review." *Mechanical Systems and Signal Processing*, 56, 123–149.

Straub, D. and Faber, M. H. (2005). "Risk based inspection planning for structural systems." *Structural safety*, 27(4), 335–355.

Unal, M. and Warn, G. P. (2017). "A set-based approach to support decision-making on the restoration of infrastructure networks." *Earthquake Spectra*, 33(2), 781–801.

Yang, D. Y. and Frangopol, D. M. (2019). "Life-cycle management of deteriorating civil infrastructure considering resilience to lifetime hazards: A general approach based on renewal-reward processes." *Reliability Engineering & System Safety*, 183, 197–212.