ICOSSAR 2021

The 13th International Conference on Structural Safety and Reliability (ICOSSAR 2021), June 21-25, 2021, Shanghai, P.R. China
J. Li, Pol D. Spanos, J.B. Chen & Y.B. Peng (Eds)

Optimizing deep reinforcement learning policies for deteriorating systems considering ordered action structuring and value of information

C. P. Andriotis[a] and K. G. Papakonstantinou[b]

[a]*Faculty of Architecture & the Built Environment, TU Delft, The Netherlands, E-mail: c.andriotis@tudelft.nl*
[b]*Dept. of Civil & Environmental Engineering, Pennsylvania State University, USA, E-mail: kpapakon@psu.edu*

ABSTRACT: Inspection and maintenance (I&M) optimization entails many sources of computational complexity, among others, due to high-dimensional decision and state variables in multi-component systems, long planning horizons, stochasticity of objectives and constraints, and inherent uncertainties in measurements and models. This paper studies how the above can be addressed within the context of constrained Partially Observable Markov Decision Processes (POMDPs) and Deep Reinforcement Learning (DRL) in a unified fashion. Special emphasis is paid on how ordered action structuring of I&M actions can be exploited to decompose the respective policy parametrizations in actor-critic DRL schemes, resulting into fully decoupled maintenance and inspection actors. It is shown that the Value of Information (VoI) is naturally utilized in such POMDP control frameworks, as directly associated with the DRL advantage functions that emerge in the gradient computations of the inspection policy parameters. Overall, the presented approach, following the natural flow of engineering decisions, results in new architectural configurations for policy networks, facilitating more efficient training, while alleviating further the dimensionality burdens related to combinatorial definitions of I&M actions. The efficiency of the methodology is demonstrated in numerical experiments of a structural system subject to corrosion, where the optimization problem is formulated to concurrently account for state and model uncertainties as well as long-term probability of failure exceedance constraints. Results showcase that the obtained DRL policies considerably outperform standard decision rules.

KEYWORDS: inspection & maintenance; deep reinforcement learning; partially observable Markov decision processes; value of information; stochastic constraints; decision theory.

1 INTRODUCTION

Inspection and maintenance (I&M) planning of deteriorating systems can be defined as a dynamic programming problem of minimizing life-cycle risks and operational costs, through proper allocation of available resources. Formulating such a program involves two discrete tasks: modeling the deteriorating environment and optimizing actions over time. Modeling of environment transitions, as these are manifested due to chronic or abrupt stressors acting on structures, can be efficiently carried out by Bayesian networks, e.g. in (Straub, 2009; Andriotis & Papakonstantinou, 2018). Based on the constructed or learned Bayesian network, optimization is often conducted heuristically, through evaluation of responses under possible decision rules, with the best rule being eventually elected as the final policy. Such decision rules typically rely on threshold- or interval-based criteria (Straub & Faber, 2005; Sørensen, 2009; Nielsen & Sørensen, 2011). Other optimization formulations in the broader area of infrastructure I&M planning, focus on more principled mathematical programming processes for determination of optimal solutions, such as gradient-based, mixed integer programming, and evolutionary algorithms, e.g. (Nishijima, et al., 2009; Ouyang & Madanat, 2004; Su, et al., 2017; Yang & Frangopol, 2018).

Optimal I&M solutions from such formulations primarily ensue from static or quasi-static problem statements, thus, in principle, approximating the optimum that can be provided by a dynamic programming-based sequence of decisions. Bridging dynamic programming and Dynamic Bayesian Networks (DBNs), Partially Observable Markov Decision Processes have been successfully used for I&M planning (Jiang, et al., 2000; Papakonstantinou & Shinozuka, 2014; Schobi & Chatzi, 2016). The theoretical elegance of POMDPs is, however, not on par with the complexity of their accompanying solution techniques. Alleviating some of the emerging complexities, point-based value iteration is successfully implemented for optimization of medium-sized I&M problems (Papakonstantinou & Shinozuka, 2014; Papakonstantinou, et al., 2016 & 2018, Morato, et al., 2022a).

In all cases, from policy enumeration to static optimization formulations to dynamic programming, attempts to trace a globally optimal solution succumb to the burdens of dimensionality; long planning horizons; stochasticity of objectives and constraints; and integration of state and/or model updating, among others. To deal with such complex and general environments, a coupled Deep Reinforcement Learning (DRL) and POMDP framework has been introduced in (Andriotis & Papakonstantinou, 2019a, 2019b), shown to significantly outperform standard risk-, condition-, and time-based I&M rules in decision analysis of multi-component engineering systems. The integrated DRL-POMDP approach has the capacity to alleviate issues of dimensionality and uncertainty, approximating arbitrarily well globally optimal belief-conditioned sequential decision paths in long-horizon planning problems. This concept is extended in (Andriotis & Papakonstantinou, 2021) to facilitate incorporation of constraints that bound strictly, or in a probabilistic sense, relevant quantities of interest, such as risk, failure probability, system availability, budget-related costs, and other measures.

In this paper, the above-described multi-agent actor-critic DRL approach is further enhanced to provide training and planning of improved efficiency. Maintenance and inspection policies are decoupled and defined by separate actor networks, following the natural structuring of decisions inside each decision analysis step. Training of the two resulting policy networks (actors) is thereby performed independently. This architectural feature reduces the original output of the policy network that would be formed through combinations of I&M actions. Moreover, it uses distinct inputs, as it does not condition the two types of actions (maintenance and inspection) on the same prior probability distribution (belief) over states of structural health (as this may be represented by corrosion depth, crack size, or other relevant engineering metrics).

Parametric independence of inspection and maintenance policies is computationally reflected in the formed gradients of the actors. The inspection-actor and the maintenance-actor are trained with their own *advantage functions*, both being linked to the centralized life-cycle cost (parametrized by the critic network). For the inspection-actor network, it is shown that the respective advantage function assumes Value of Information (VoI) semantics, i.e., conditional VoI is directly leveraged for learning of inspection network weights during training. The role of VoI, as a metric quantifying the amount the decision-maker should be willing to pay for structural data prior to maintenance (Thöns & Faber, 2011; Pozzi & Der Kiureghian, 2011; Andriotis, et al., 2021), is therefore explicitly positioned within the framework of DRL-POMDP and within planning of the inspection policy.

As an application of the developed approach, a multi-component steel truss structure subject to corrosion deterioration is examined. For the optimization problem, the failure risk over the service life is constrained, so that the structure adheres to prescribed safety levels, whereas the deterioration model is also uncertain and gets updated in time based on data. Results show the suggested approach to outperform standardly optimized decision rules. It is additionally shown that the new training scheme furnishes advantages of theoretical consistency and relies on more interpretable metrics, such as VoI, for policy training and action selection.

2 I&M OPTIMIZATION WITH POMDPs

The stochastic objective to be optimized in the I&M problem under consideration is the following:

$$V^{*}\left(\mathbf{b}_{0}\right)=\min_{\pi\in\Pi}\mathbb{E}\left[\sum_{t=0}^{T}\gamma^{t}c_{t}\left|a_{t}\sim\pi\left(o_{0:t},a_{0:t-1}\right)\right.\right] \qquad (1)$$

where V^{*} is the optimal expected life-cycle cost; $\mathbf{b}_0$ is the initial probability distribution of the system states/parameters; c_t is the cost at time step t; T is the length of service life; γ is a discount factor in (0,1); $a_t = (a_{I,t}, a_{M,t}) \in A = A_I \times A_M$ is an inspection ($a_{I,t}$) and maintenance ($a_{M,t}$) action at time step t; $o_t \in \Omega$ is an observation of structural health at time t; and $\pi \in \Pi$ is a policy (decision rule). Policy, π, is a function that maps past observations and actions to a new action. As such, at best, an optimal policy may consider the entire history of actions and observations up to each time step t, ($o_{0:t}$,$a_{0:t\text{-}1}$), to output an action a_t. Expectation of Eq. (1) is taken over possible states $s \in S$, as reached through transitions $e \in \mathcal{E}$, and the action outcomes.

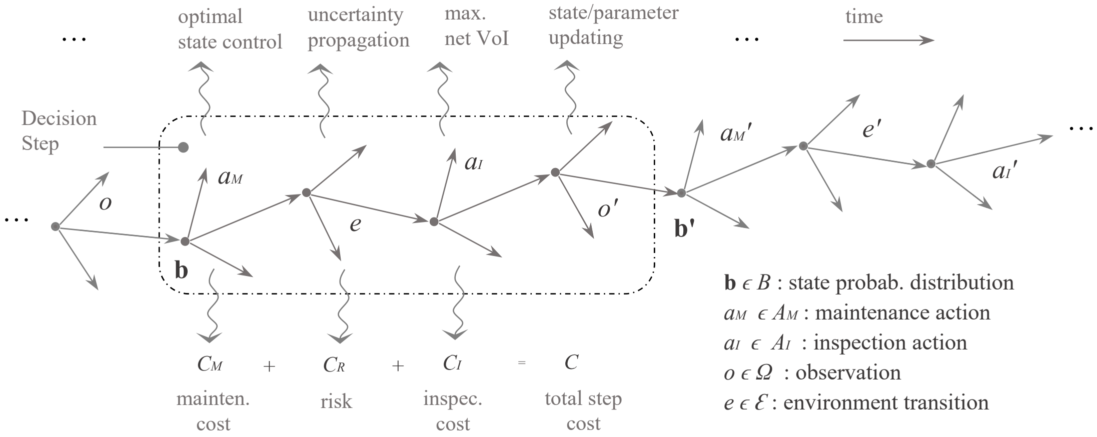
optimal state control
uncertainty propagation
max. net VoI
state/parameter updating
time
Decision Step
o
a_M
e
a_I
o'
$\mathbf{b}$
$\mathbf{b}'$
a_M'
e'
a_I'
C_M + C_R + C_I = C
mainten. cost
risk
inspec. cost
total step cost
$\mathbf{b} \in B$: state probab. distribution
$a_M \in A_M$: maintenance action
$a_I \in A_I$: inspection action
$o \in \Omega$: observation
$e \in \mathcal{E}$: environment transition

Figure 1. Recurrence of sequential decisions, random variables, and computational steps in inspection and maintenance planning of deteriorating engineering systems.

The cost function, c_t, depends on states and actions and is decomposed into sub-costs:

$$c(s,a_I,a_M)=\underbrace{c_M(a_M)}_{\text{maintenance}}+\underbrace{\gamma c_I(a_I)}_{\text{inspection}}+\underbrace{c_R(s,a_M)}_{\text{risk}} \tag{2}$$

Additional sub-costs can be included in this class of problems, such as scheduled shutdown costs and/or unavailability costs (Andriotis & Papakonstantinou, 2021). In case life-cycle is considered in its entirety, the cost function at time step 0 and T can also include initial and terminal costs such as design/construction and decommissioning costs, respectively (Morato, et al., 2022b). Such sub-costs are omitted here.

The sequence of random variable realizations and decisions within each step is shown in Figure 1. Accordingly, the computational tasks undertaken within each step consist of an optimal state control task; an uncertainty propagation (prediction) task; a Value of Information (VoI) maximization task; and a state/model updating task. POMDPs seek globally optimal solutions over a multi-step horizon, unifying the above tasks. In a POMDP, states representing the condition of structural health and/or model parameters are partially observable to the decision-maker, i.e., modeled as latent random variables (Papakonstantinou & Shinozuka, 2014; Schobi & Chatzi, 2016). Collected observations through inspections and monitoring are conditioned on these states and are used to update their respective priors. This forward filtering operation is herein described by a Bayesian update:

$$b'(s')=\Pr(s'|o',a_I,a_M,\mathbf{b})$$

$$\propto \Pr(o'|s',a_I)\Pr(s'|a_M,\mathbf{b}) \tag{3}$$

where $\mathbf{b}_t$ is the probability distribution over S at time t, and $(\cdot)'$ is $(\cdot)$ at the next time step. Based on this belief update, the Bellman equation, describing the optimal value of the objective function of Eq. (1), can be written as:

$$\begin{aligned}V^*(\mathbf{b}) &= \min_{a\in A}\Bigg\{\sum_{s\in S} b(s)c(s,a)+ \\ &\qquad \gamma\sum_{o\in\Omega}\Pr(o'|\mathbf{b},a)V^*(\mathbf{b}')\Bigg\} \\ &= \min_{a_M\in A_M}\Bigg\{\sum_{s\in S} b(s)(c_M+c_R)+\gamma V^*(\mathbf{b}^{a_M}) \\ &\qquad -\gamma\max_{a_I\in A_I} VoI_{net}(a_I)\Bigg\}\end{aligned} \tag{4}$$

where VoI_{net} denotes the net Value of Information (VoI) associated with inspection action a_I:

$$VoI_{net}(a_I)=V^*(\mathbf{b}^{a_M})-\mathbb{E}_o\left[V^*(\mathbf{b}')\right]-c_I \tag{5}$$

Thus, Eq. (4) describes that following execution of a maintenance action a_M, under an optimal POMDP policy, inspections a_I are chosen based on the net VoI (Andriotis, et al., 2021). Within the mathematical principles of POMDPs, this assures non-negativity of VoI both at every decision step and over the entire horizon of decisions.

2.1 Optimization under life-cycle risk constraints

Consideration of constraints in the decision optimization problem enables a policy to control the deterioration of structural health while satisfying pragmatic desired or required targets on some quantities of interest, as these for example relate to maintenance costs, risks, etc. These quantities are not necessarily bounded at every single step, in general, but rather over longer horizons. Some types of constraints may also be perceived only stochastically, as they depend on the underlying stochastic processes that govern the deterioration.

An example of such a type of constraints, which is of interest to this work, is bounding the failure probability or risk over the service life of a structure. The need for a stochastic formulation of this constraint type can be understood if one notes that the probability of failure over multiple time steps is updatable based on observations at instances where inspection actions are prescribed by the policy. Thereby, different realizations of a policy, producing different observation sequences, yield different trajectories of probability of failure. Accordingly, the expected cumulative risk is herein bounded as:

$$\begin{aligned}\mathfrak{R}_F^{\pi} &= \mathbb{E}\left[\sum_{t=0}^{T}\gamma^t c_R(s_t, a_{M,t})\right]\\ &= c_F\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t\left(P_{F_{t+1}|a_{0:t},o_{0:t}} - P_{F_t|a_{0:t},o_{0:t}}\right)\right] \qquad (6)\\ &= c_F\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t\Delta P_{F_t}\right] \leq \mathfrak{R}_{ult}\end{aligned}$$

where c$_F$ is the failure cost, P_{F_t} is the failure probability up to time t, and $\mathfrak{R}_{ult}$ is a prescribed life-cycle risk tolerance. A broader family of constraints and how these can be implemented in a POMDP/DRL framework is presented in more elaborate terms in (Andriotis & Papakonstantinou, 2021). Attaching the constraint of Eq. (6) at the objective function of Eq. (1), the following max-min optimization problem is eventually defined:

$$\begin{aligned}V^*(\mathbf{b}_0) &= \max_{\lambda\geq 0} V_{\lambda}^*(\mathbf{b}_0)\\ &= \max_{\lambda\geq 0}\min_{\pi\in\Pi}\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t\left(c_t + \lambda c_F\Delta P_{F_t}\right)\right. \qquad (7)\\ &\qquad \left.- \lambda\mathfrak{R}_{ult}^{\pi}\,\middle|\, a_t \sim \pi\left(o_{0:t}, a_{0:t-1}\right)\right]\end{aligned}$$

where λ is a Lagrange multiplier transforming the constrained problem into an unconstrained one. A globally optimal solution to the above problem is sought through the belief-based constrained multi-agent actor-critic DRL approach presented in the next section, for multi-component systems.

2.2 System-level control through multi-agent belief-based DRL

In the actor-critic DRL approach taken in this paper, following the deep multi-agent algorithmic schemes in (Andriotis & Papakonstantinou, 2019a; 2021), the value function of Eq. (7) is parametrized by a critic network, with parameters $\boldsymbol{\theta}_V$, which are gradually learned during training:

$$V_{\lambda}^{\pi}\left(\hat{\mathbf{b}}\right) \simeq V_{\lambda}^{\pi}\left(\hat{\mathbf{b}};\, \boldsymbol{\theta}_V\right) \qquad (8)$$

where $\hat{\mathbf{b}}$ is the system belief (typically consisting of all factored component and model parameter probability distributions). The multi-agent policy actor is also similarly parametrized, with each component being treated as a separate agent. Accordingly, policies of different components have their individual action outputs, which are conditionally independent given the current system belief:

$$\pi\left(\mathbf{a}\,|\,\hat{\mathbf{b}}\right) = \prod_{i=1}^{N_C}\pi_i\left(a^{(i)}\,|\,\hat{\mathbf{b}};\, \boldsymbol{\theta}_{\pi}\right) \qquad (9)$$

where N_c is the number of components (or control units), $\mathbf{a}$ is a vector of actions $a^{(i)}$, and $\boldsymbol{\theta}_{\pi}$ is the vector of the policy network parameters. The parameters of the actor and critic networks are updated based on their gradients. Based on Eq. (9), the actor gradient is given within the premises of the policy gradient theorem (Sutton, et al., 2000):

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_{\pi}} V_{\lambda}^{\pi}\left(\hat{\mathbf{b}}\right) = \mathbb{E}_{\mathcal{M}}\Bigg[& wA_{\lambda}^{\pi}\left(\hat{\mathbf{b}}, \mathbf{a};\, \boldsymbol{\theta}_V\right)\cdot\\ & \cdot\sum_{i=1}^{N_C}\nabla_{\boldsymbol{\theta}_{\pi}}\log\pi_i\left(a^{(i)}\,|\,\hat{\mathbf{b}};\, \boldsymbol{\theta}_{\pi}\right)\Bigg] \qquad (10)\end{aligned}$$

where w is an importance sampling weight, A_{λ}^{π} is the advantage function, and $\mathcal{M}$ is the experience replay containing information of past transitions and costs. Importance sampling is used because off-policy learning is used, i.e., experiences retrieved

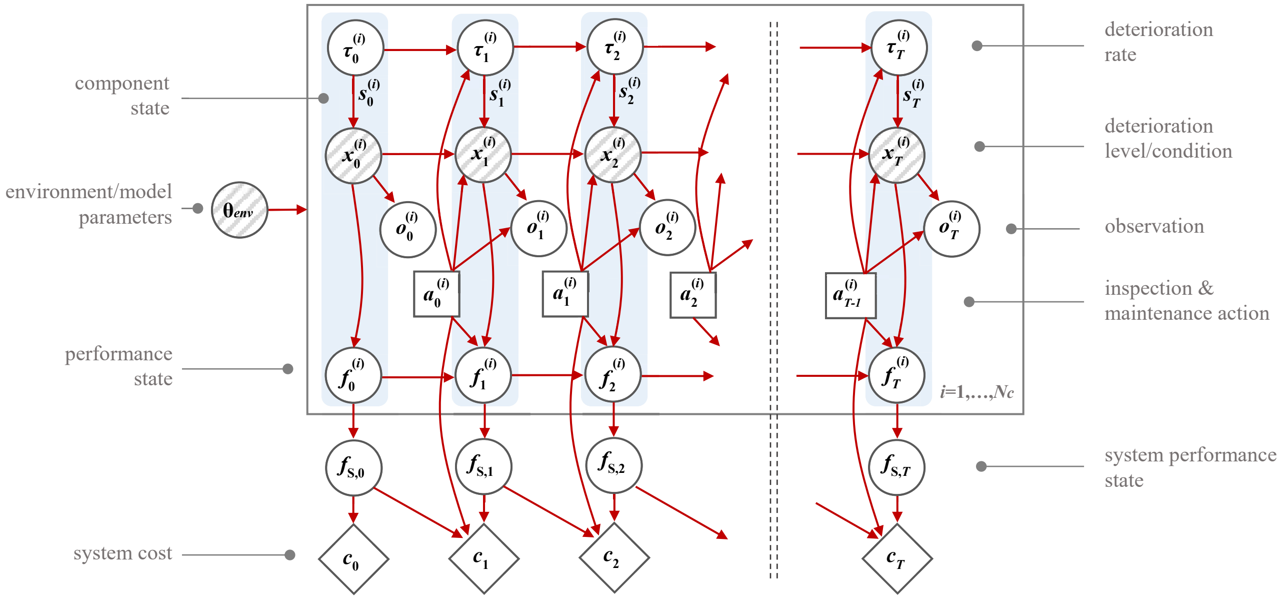

Figure 2. Underlying dynamic Bayesian network describing deterioration, structural and statistical dependencies among components, and effects of decisions on state and observation random variables.

from $\mathcal{M}$ were generated by policies other than the current one (in training time).

According to Eq. (7), and using the parametrized Lagrangian value function of Eq. (8), the advantage function is computed through the following relation:

$$A_\lambda^\pi\left(\hat{\mathbf{b}},\mathbf{a};\,\boldsymbol{\theta}_V\right) \simeq -\mathbb{E}_{\mathbf{s}}\left[c(\mathbf{s},\mathbf{a})\right] - \lambda c_F \Delta P_{F_t} \\ -\gamma V_\lambda^\pi\left(\hat{\mathbf{b}}';\,\boldsymbol{\theta}_V\right) + V_\lambda^\pi\left(\hat{\mathbf{b}};\,\boldsymbol{\theta}_V\right) \tag{11}$$

The critic network gradient is similarly given as:

$$\nabla_{\boldsymbol{\theta}_V} V_\lambda^\pi\left(\hat{\mathbf{b}}\right) = \mathbb{E}_{\mathcal{M}}\left[wA_\lambda^\pi\left(\hat{\mathbf{b}},\mathbf{a};\boldsymbol{\theta}_V\right)\nabla_{\boldsymbol{\theta}_V} V_\lambda^\pi\left(\hat{\mathbf{b}};\boldsymbol{\theta}_V\right)\right] \tag{12}$$

V_λ^π is linear with respect to λ and thus the respective gradient is:

$$\nabla_\lambda V_\lambda^\pi \simeq c_F \sum_{t=0}^{T} \gamma^t \Delta P_{F_t} - \Re_{ult}^\pi \tag{13}$$

Lagrange multipliers are updated using Eq. (13) in an on-policy manner at the end of every episode (service-life realization), and therefore, unlike for the other gradients, importance sampling is not required.

Based on the gradients of Eqs. (10), (12), and (13), respective parameters are updated through stochastic gradient descent. For λ=0, the constrained problem is now defined as an unconstrained one, with Eqs. (8)-(12) remaining unchanged. In that case, the gradient of Eq. (13) becomes irrelevant.

3 DECOUPLED INSPECTION AND MAINTENANCE ACTORS

Following the ordered structuring of inspection and maintenance actions within each decision step (Figure 1) the factored policy representation in Eq. (9) can be written as:

$$\pi\left(\mathbf{a}\,|\,\hat{\mathbf{b}}\right) = \underbrace{\prod_{i=1}^{N_{C,M}} \pi_M^{(i)}\left(a_M^{(i)}\,|\,\hat{\mathbf{b}}\right)}_{\pi_M:\text{ maintenance policy}} \underbrace{\prod_{i=1}^{N_{C,I}} \pi_I^{(i)}\left(a_I^{(i)}\,|\,\hat{\mathbf{b}}^{\mathbf{a}_M}\right)}_{\pi_I:\text{ inspection policy}} \tag{14}$$

Parametrizing the two policies with different networks, i.e., having $\boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_I$, and substituting Eq. (14) in Eq. (10), it immediately follows that the gradient is decomposed into two parts. Based on this and keeping only the advantage parts that are relevant to each decision, the maintenance and inspection actors have the following gradients:

$$\nabla_{\boldsymbol{\theta}_M} V_\lambda^\pi = \mathbb{E}_{\mathcal{M}}\left[w_M A_{M,\lambda}^\pi\left(\hat{\mathbf{b}},\mathbf{a}_M;\,\boldsymbol{\theta}_V\right)\cdot \\ \cdot\sum_{i=1}^{N_{C,M}} \nabla_{\boldsymbol{\theta}_M} \log \pi_M^{(i)}\left(a_M^{(i)}\,|\,\hat{\mathbf{b}};\,\boldsymbol{\theta}_M\right)\right] \tag{15}$$

$$\nabla_{\boldsymbol{\theta}_I} V_\lambda^\pi = \mathbb{E}_{\mathcal{M}}\Bigg[w_I A_{I,\lambda}^\pi\left(\hat{\mathbf{b}}^{\mathbf{a}_M}, \mathbf{a}_I; \boldsymbol{\theta}_V\right) \cdot \\ \cdot \sum_{i=1}^{N_{C,I}} \nabla_{\boldsymbol{\theta}_I} \log \pi_I^{(i)}\left(a_I^{(i)} \,|\, \hat{\mathbf{b}}^{\mathbf{a}_M}; \boldsymbol{\theta}_I\right)\Bigg] \qquad (16)$$

The advantage functions, $A_{M,\lambda}^\pi$, $A_{I,\lambda}^\pi$, which are necessary for the training of the maintenance and inspection policy networks, respectively, assume the following forms:

$$A_{M,\lambda}^\pi\left(\hat{\mathbf{b}}, \mathbf{a}_M; \boldsymbol{\theta}_V\right) \simeq -\mathbb{E}_{\mathbf{s}}\left[c_M(\mathbf{s},\mathbf{a})\right] - \lambda c_F \Delta P_{F_t} \\ -\gamma V_\lambda^\pi\left(\hat{\mathbf{b}}^{\mathbf{a}_M}; \boldsymbol{\theta}_V\right) + V_\lambda^\pi\left(\hat{\mathbf{b}}; \boldsymbol{\theta}_V\right) \qquad (17)$$

$$A_{I,\lambda}^\pi\left(\hat{\mathbf{b}}^{\mathbf{a}_M}, \mathbf{a}_I; \boldsymbol{\theta}_V\right) \simeq -c_I(\mathbf{a}_I) - \\ V_\lambda^\pi\left(\hat{\mathbf{b}}'; \boldsymbol{\theta}_V\right) + V_\lambda^\pi\left(\hat{\mathbf{b}}^{\mathbf{a}_M}; \boldsymbol{\theta}_V\right) \qquad (18)$$

Following standard nomenclature in structural reliability literature regarding information value, e.g. in (Straub, 2014; Konakli, et al., 2015), it can be readily noted that the inspection advantage is the net value of Conditional VoI (CVoI) at every time step *t*:

$$CVoI_{net} = A_{I,\lambda}^\pi\left(\hat{\mathbf{b}}^{\mathbf{a}_M}, \mathbf{a}_I\right) \qquad (19)$$

Taking the expected value over all possible observations, one can compute VoI:

$$VoI_{net} = \mathbb{E}_{\mathbf{o}}\left[A_{I,\lambda}^\pi\left(\hat{\mathbf{b}}^{\mathbf{a}_M}, \mathbf{a}_I\right)\right] \qquad (20)$$

As shown in Eq. (4), VoI is inherently present in POMDPs as the selection criterion of inspection actions at every step. This property can be also naturally utilized in DRL to avoid separate related neural networks and inspection parametrizations, thus only parametrizing the maintenance policy and choosing the inspection that maximizes net VoI, using Eqs. (5) and (20). This option would, however, require an accurate estimate of VoI at every step, which in multi-component systems is computationally hard (expectation over component observations, **o**).

By parametrizing instead the inspection network and using CVoI to train the parameters, this optimal inspection behavior is gradually approximated through gradient descent. In similar terms, the advantage function backpropagated for the maintenance actor can be interpreted as net conditional value of maintenance.

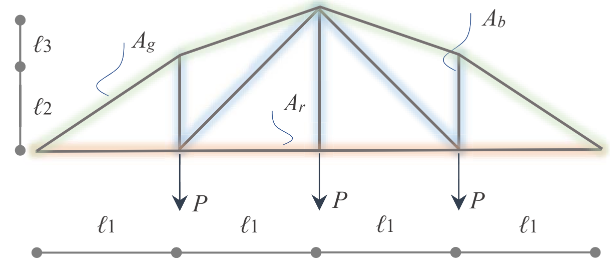

Figure 3. Multi-component deteriorating truss structure.

Table 1. Environment parameters and random variables.

Name	Type [a]	Values [b]
Annual corrosion depth, δx (cm)	*Gam*($\delta\tilde{x}\cdot\kappa,\kappa$), see Eq. (21)	0: 0.02: 1.0
Initial corrosion depth, x_0 (cm)	*Exp*(2)	0: 0.02: 1.0
Gamma parameter, m_{70} (mm)	*Uni*([4,6])	4: 0.5 :6
Gamma parameter, σ_{70} (mm)	*deterministic*	0.2 m_{70}
Gamma parameter, β	*deterministic*	1.5
Observation, o (cm)	*Norm*(x,6%·x)	0: 0.02: 1.0
Load, P (kN)	*Weib*(330,72)	0: 5: 380
Nominal section areas, A_r, A_g, A_b (m^2)	*deterministic*	32e-4, 38e-4, 26e-4
Yield strength, σ_y (MPa)	*deterministic*	355.0
Lengths, ℓ_1, ℓ_2, ℓ_3 (m)	*deterministic*	6.0, 3.0, 4.5
Costs, c_I, c_{mat}, c_{bas}, c_F (c_{reb})[c]	*deterministic*	0.01, 0.6, 0.02, 1.0
Discount, γ	*deterministic*	0.95

[a] *Gam*, *Exp*, *Uni*, *Norm*, and *Weib*, are gamma, exponential, uniform, normal, and Weibull distributions, respectively.

[b] For random variables, reported values include minimum: step: maximum.

[c] c_{reb}: entire structure rebuild cost; c_{mat}: material cost of intact structural volume; c_{bas}: base cost of replacement campaign incurred when $N_{C,M}$>0; $\mathfrak{R}_{ult}$ =0.05c_{reb}.

4 RESULTS

4.1 Modeling of the deteriorating environment

Scheduling of I&M actions is considered for a multi-component steel structure subject to deterioration. The structure consists of 13 truss members, which incur cross section losses due to operation in a corroding environment for a service horizon of 50 years. I&M actions are taken once

Table 2. Description of optimized baseline policies.

Policy Acronym	Decision variables	Optimal values
TPI: Time Periodic Inspection	inspection time interval, t_I ; section-loss replacement threshold, o_R	t_I* =5 (years); o_R* = 6 (mm)
RBI: Risk-Based Inspection	system PoF[a] inspection threshold, p_{th} ; section-loss replacement threshold, o_R	p_{th}*= 8e-3; o_R*= 8 (mm)
TPI-RBP[b]: TPI & Risk-Based Prioritization	number of components to be maintained, n_{pr} ; t_I ; o_R	n_{pr}* = 9 t_I* = 4 (years); o_R*= 6 (mm)
RBI-RBP[b]: RBI & Risk-Based Prioritization	number of components to be maintained, n_{pr} ; p_{th} ; o_R	n_{pr}* = 8; p_{th}* = 6e-3; o_R*= 6 (mm)

[a] Probability of failure of structural system.

[b] RBP is conducted based on PoF of individual components.

Table 3. Life-cycle costs of baseline and DRL policies [a].

Cost	TPI	RBI	TPI-RBP	RBI-RBP	DRL
I&M	**2.18**	**1.82**	**1.84**	**1.57**	**1.00**
Insp.	0.20	0.21	0.24	0.20	0.25
Maint.	1.98	1.61	1.60	1.37	0.75
Risk	0.27	0.29	0.30	0.30	0.29

[a] Computed based on 10^3 policy realizations and normalized with respect to the DRL policy. Normalized 95% confidence intervals are tighter than 0.1. For all policies constraint violation estimate is lower than 3.5%.

every year. The deterioration process DBN and the truss are shown in Figures 2 and 3, respectively.

In the DBN of Figure 2, deterioration rate, $\tau^{(i)}$, corresponds to component age; deterioration level, $x^{(i)}$, corresponds to corrosion depth (assumed uniform over the cross section); component performance state, $f^{\,(i)}$, indicates the failure of a component (binary); and system performance, f_s, indicates system failure (binary). The overall system belief is defined based on the above variables together with the environment parameters, $\boldsymbol{\theta}_{env}$. The depth of corrosion of each member is modeled as a gamma process, whose mean follows a power law (Frangopol, et al., 2004), as is typical for the assumed type of stressor and material:

$$\begin{aligned} \tilde{x}_\tau &= r\tau^\beta / \kappa \\ x_\tau - x_{\tau-1} &\sim Gamma\left(\cdot \mid r\tau^\beta - r\left(\tau-1\right)^\beta, \kappa\right) \end{aligned} \qquad (21)$$

Environment parameters, $\boldsymbol{\theta}_{env}$, comprise (β, r, κ) of Eq. (21) together with load, P. Given β, parameters r, κ are determined through known mean corrosion depth (m_{70}) values and the respective standard deviation (σ_{70}) after exposure of 70 years. The state of each member also includes its age. Age is considered, without loss of generality, to be a fully observable variable reflecting the rate of deterioration, evolving deterministically based on the selected maintenance action, in contrast to the corrosion depth variable which is stochastic and latent, only perceivable through observations collected at time instances of inspection visits.

The global environment parameter m_{70} of the gamma process is also a latent variable. As illustrated in Figure 2, it is assumed that m_{70} is an overarching parameter, shared among components, since all member stochastic deterioration processes ensue from the same environment. This parameter is also identifiable in inference time, i.e., during the deployment phase of the life-cycle policy, based on direct cross-section thickness measurements (observations). Discretization and probabilistic assumptions for the corrosion modeling can be found in Table 1. At inspection times, all components are inspected ($N_{C,I}$=1). Maintenance actions exist for each member ($N_{C,M}$=13), including do-nothing and perfect-repairs (replacement).

Component failure occurs when the cross section normal stress, σ, exceeds the material yield stress, σ_y, or the cross section area loss exceeds 50%. System failure occurs upon failure of at least one member (series system assumption). Members under compression are not prone to buckling due to appropriate slenderness ratio. The same value of corrosion penetration depth inflicts roughly equal percentage of cross section loss to all members, due to the equally thick hollow circular cross sections used for the structural members (1cm). Parameter details for the environment are presented in Table 1.

4.2 *Deep network parametrization and training*

On the basis of the derivations of Section 3, separate inspection and maintenance actors are utilized, whereas the critic network corresponds to a parametrization of the Lagrangian value function, allowing us to evaluate the actors' advantage functions through a surrogate of the life-cycle cost. For the maintenance actor, the used parametrization introduces independency among components, in accordance with the deep decentralized multi-agent actor critic architecture in (Andriotis & Papakonstantinou, 2021).

Each actor has 2x100 hidden layers, mapping the system belief to a binary output, whereas the critic is parametrized with 2x300 hidden layers. For the inspection actor, the binary output applies to all members. The involved networks were trained with Keras with Tensorflow backend version 1.5.0.

4.3 *Policy evaluations and comparisons*

To assess the quality of the learned DRL policy we compare against different baselines that are built and optimized as per risk-, condition-, and time-

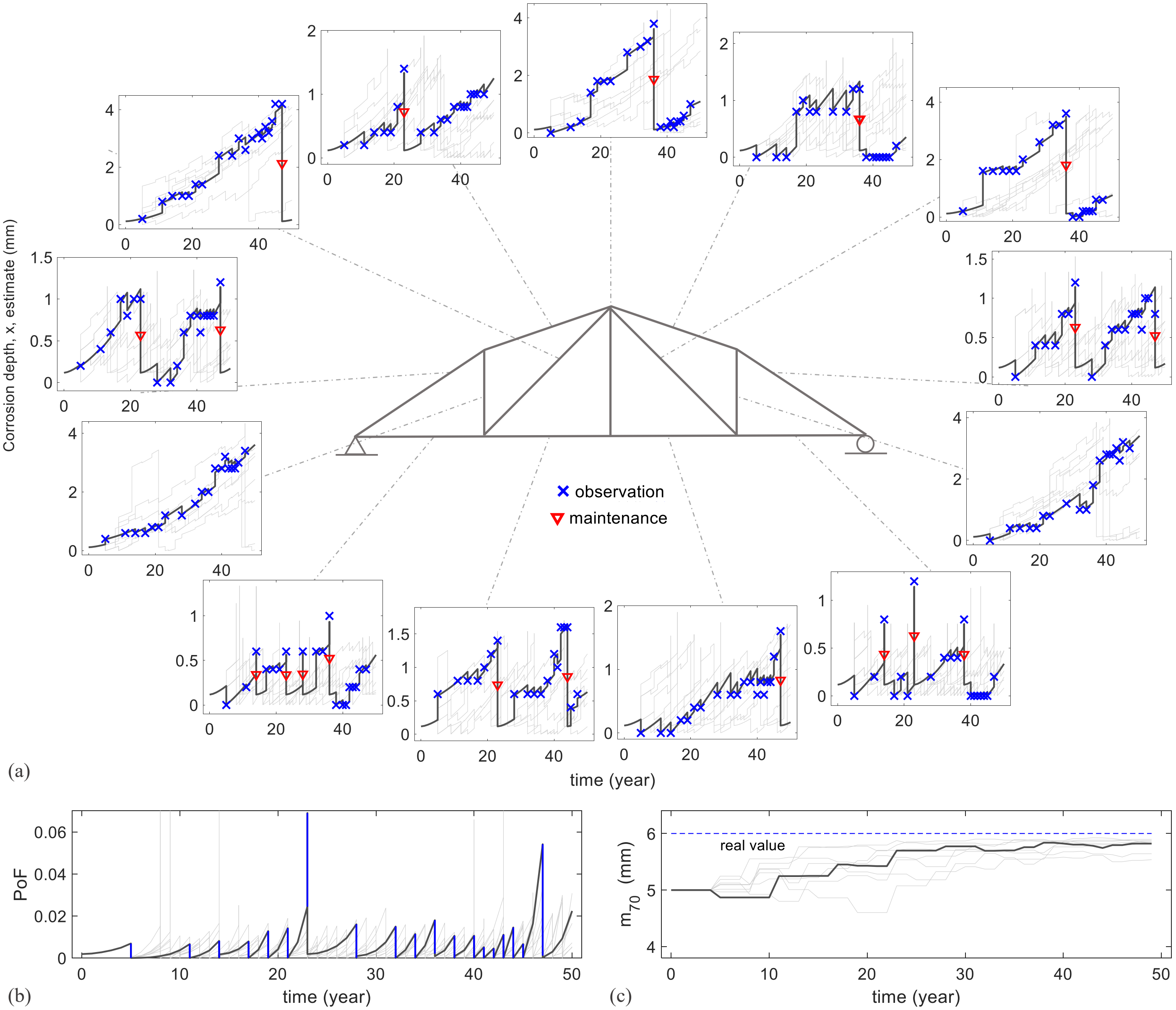

Figure 4. Policy realization of the trained multi-agent actor-critic DRL networks. (a) Corrosion depth mean estimates based on history of previous actions and observations; (b) System probability of failure (PoF), with the observation-driven update part indicated in blue; (c) Gamma process environment parameter identification.

based assumptions. These policies are succinctly described in Table 2. The life-cycle costs of all policies, including the DRL one, are reported in Table 3. It can be observed that the DRL policy reaches a 57% lower life-cycle cost than the optimized risk-based policy with component prioritization. This outcome is made possible by the dynamic and adaptive nature of DRL policies, which are not relying on static thresholds and/or conditioning of actions on noisy observation outcomes, but rather perform a direct mapping from the dynamically evolving posterior state and model parameter beliefs to I&M actions.

Indicative realizations of the DRL policy are shown in Figure 4. For these realizations, the real environment parameter is assumed to be m_{70}=6 mm, indicating operation of the structure under severely corrosive conditions. Surrounding plots in Figure 4(a) show mean estimated corrosion depths for all members, as computed throughout the service life based on actions and observations and the corresponding belief vectors at each time step. The system probability of failure and gamma process parameter updates are also depicted in Figures 4(b) and 4(c). In the highlighted policy realization, it is observed that 19 inspections were necessary during the planning horizon of 50 years, taken at years 5, 11, 14, 17, 19, 21, 23, 28, 32, 34, 36, 38, 40-45, and 47. Maintenance actions (component perfect-repairs) are dynamically selected for each component based on both its individual corrosion depth estimates (inferred via inspections) and system-level scheduling considerations.

It can be observed, for example, in Figure 4(a), that for the lower left member, although maintenance is taken in year 23, the updated corrosion estimate after inspection at year 28 dictates maintenance at that time too. Later, although equal levels of corrosion are reached or exceeded, repair is postponed until more components can be included in the intervention (year 36), as there is a base cost associated with maintenance campaigns. It can be, therefore, seen that component-level adaptability regarding maintenance activity is not irrelevant to system-level considerations. To reduce single-member interventions, maintenance activity is automatically grouped. For example, 6 members are maintained in years 23, 4 members in years 36 and 47, and 2 members in year 14. This pattern is observed in all realizations without, however, suggesting fixed, time-based interventions (see other realizations in grey). It is instead noticed to be a mixture of opportunistic criteria and time-based maintenance towards achieving optimality. Component classes are also discovered as reflected through similar policy patterns, e.g., internal members are seen to mainly require no more than one repair during the service life, typically after 30 years. However, the actual timing of the respective visit is again adjusted so that single-component interventions are avoided. This is an adaptability feature driven by the unique observation sequences of the life-cycle realizations.

5 CONCLUSIONS

An actor-critic Deep Reinforcement Learning (DRL) architecture and training approach is presented in this paper. Following the ordered action structuring of I&M actions in decision analysis of deteriorating structures, decoupled inspection/maintenance actor networks are devised. That is, the developed architecture recurrently conditions maintenance and inspection decisions on post-inspection and post-maintenance posterior beliefs, respectively. The two networks are trained based on their own distinct advantage functions. In the case of the inspection policy network, the advantage function coincides with the net conditional Value of Information (VoI), a metric that can objectively guide inspection decision updates in both learning and deployment times. Thereby, under the learned policy, the inspection plan incorporates the inspection selection criterion inherently present in Partially Observable Markov Decision Processes (POMDPs), i.e., the maximization of the net VoI. Following this intuitive formulation regarding decisions, this approach is found to provide adept planning solutions for a long-horizon problem of a multi-component deteriorating structure under corrosion. The decoupled DRL-POMDP policy outperforms by at least 57% standard baselines following condition-, risk-, and time-based assumptions.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the U.S. National Science Foundation under CAREER Grant No. 1751941 and LEAP-HI Grant No. 2053620, and the Center for Integrated Asset Management for Multimodal Transportation Infrastructure Systems, 2018 U.S. DOT Region 3 University Center. Dr. Andriotis would further like to acknowledge the support of the TU Delft AI Labs program.

REFERENCES

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

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

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

Andriotis, C. P. & Papakonstantinou, K. G., 2019b. *Life-cycle policies for large engineering systems under complete and partial observability.* 13th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP), Seoul, South Korea.

Andriotis, C. P. & Papakonstantinou, K. G., 2018. Extended and generalized fragility functions. *Journal of Engineering Mechanics,* 144(9), p. 04018087.

Frangopol, D. M., Kallen, M. & Noortwijk, J., 2004. Probabilistic models for life-cycle performance of deteriorating structures: review and future directions. *Progress in Structural Engineering and Materials,* 6(4), pp. 197-212.

Jiang, M., Corotis, R. & Ellis, J., 2000. Optimal life-cycle costing with partial observability. *Journal of infrastructure systems,* 6(2), pp. 56-66.

Konakli, K., Sudret, B. & Faber, M., 2015. Numerical investigations into the value of information in lifecycle analysis of structural systems. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering,* 2(3), p. B4015007.

Morato, P. G., Papakonstantinou, K. G., Andriotis, C. P., Nielsen, J. S. & Rigo 2022a. Optimal inspection and maintenance planning for deteriorating structural components through dynamic Bayesian networks and Markov decision processes. *Structural Safety*, 94, p. 102140.

Morato, P. G., Papakonstantinou, K. G., Andriotis, C. P. & Rigo, P., 2022b. *Managing offshore wind turbines through Markov decision processes and dynamic Bayesian networks.* 13th International Conference on Structural Safety & Reliability (ICOSSAR), Shanghai, China.

Nielsen, J. S. & Sørensen, J. D., 2011. On risk-based operation and maintenance of offshore wind turbine components. *Reliability Engineering & System Safety,* 96(1), pp. 218-29.

Nishijima, K., Maes, M. A., Goyet, J. & Faber, M. H., 2009. Constrained optimization of component reliabilities in complex systems. *Structural Safety,* 31(2), pp. 168-78.

Ouyang, Y. & Madanat, S., 2004. Optimal scheduling of rehabilitation activities for multiple pavement facilities: exact and approximate solutions. *Transportation Research Part A: Policy and Practice,* 38(5), pp. 347-65.

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

Papakonstantinou, K. G., Andriotis, C. P. & Shinozuka, M., 2016. *Point-based POMDP solvers for lifecycle cost minimization of deteriorating structures.* 5th International Symposium on Life-Cycle Civil Engineering (IALCCE), Delft, The Netherelands.

Papakonstantinou, K. G. & Shinozuka, M., 2014. Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part I: Theory. *Reliability Engineering & System Safety,* 130, p. 202-213.

Papakonstantinou, K. G & Shinozuka, M., 2014. Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part II: POMDP implementation. *Reliability Engineering & System Safety,* 130, p. 214-224.

Pozzi, M. & Der Kiureghian, A., 2011. *Assessing the value of information for long-term structural health monitoring.* Health Monitoring of Structural and Biological Systems, International Society for Optics and Photonics.

Schobi, R. & Chatzi, E. N., 2016. Maintenance planning using continuous-state partially observable Markov decision processes and non-linear action models. *Structure and Infrastructure Engineering,* 12(8), p. 977-994.

Sørensen, J., 2009. Framework for risk-based planning of operation and maintenance for offshore wind turbines. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology,* 12(5), p. 493-506.

Straub, D., 2009. Stochastic modeling of deterioration processes through dynamic Bayesian networks. *Journal of Engineering Mechanics,* 135(10), pp. 1089-1099.

Straub, D., 2014. Value of information analysis with structural reliability methods. *Structural Safety,* 49, pp. 75-85.

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

Sutton, R. S., McAllester, D. A., Singh, S. P. & Mansour, Y., 2000. *Policy gradient methods for reinforcement learning with function approximation.* Advances in Neural Information Processing Systems, pp. 1057-1063.

Su, Z., Jamshidi, A., Núñez, A., Baldi, S. & De Schutter, B., 2017. Multi-level condition-based maintenance planning for railway infrastructures–A scenario-based chance-constrained approach. *Transportation Research Part C: Emerging Technologies,* 27(4), pp. 92-123.

Thöns, S. & Faber, M. H., 2011. *Assessing the value of structural health monitoring.* 11th International Conference on Structural Safety & Reliability (ICOSSAR), New York, USA.

Yang, D. & Frangopol, D., 2018. Probabilistic optimization framework for inspection/repair planning of fatigue-critical details using dynamic Bayesian natworks. *Computers and Structures,* 198, pp. 40-50.