

Deep Reinforcement Learning for Optimal Control of Induction Welding Process

Amit Surana, Abhijit Chakraborty, John Gangloff, Wenping Zhao

^a*RTX Technology Research Center, 411 Silver Lane, East Hartford, 06118, CT, USA*

Abstract

Optimizing induction welding (IW) process parameters for the application of joining thermoplastic composites is challenging as it requires achieving complex spatiotemporal thermal characteristics along the weld-line to obtain desired weld quality. We formulate an optimal control problem which captures these requirements and seeks to optimize the IW coil speed using a fast-acting dynamic IW process model. We develop a novel Deep Reinforcement Learning (DRL) framework to solve this computationally challenging control problem and demonstrate via simulation study that the learned DRL feedback control policy results in better spatiotemporal thermal characteristics as compared to the current state-of-the-art.

Keywords: Deep Reinforcement Learning, Induction Welding, Optimal Control.

1. Introduction

Lightweight carbon fiber-reinforced thermoplastic composites (CF-TPCs) are becoming increasingly popular for commercial aerospace industry due to improved toughness and better environmental resistance over traditional thermoset composite materials [1]. Currently produced TPCs have simple geometry due to the limited allowable deformation of the reinforcing fibers. On the other hand, large composite aircraft parts such as fuselage panels and fan cowl doors are characterized by complex geometries and variable skin thicknesses. These competing structural requirements makes the joining of these TPCs a critical step. There are three popular joining techniques: ultrasonic, resistance and electromagnetic induction welding. This paper will focus on the challenges associated with induction welding (IW) technique. In particular, the main challenge for IW is to achieve the appropriate spatiotemporal thermal behavior along the entire weld-line which is necessary to obtain desired weld quality. This can be particularly difficult to achieve due to the complexity of the part geometry. The success of IW for such applications, therefore requires the establishment of complex spatially dependent welding parameter settings over the length of each joint, i.e. a different recipe for each joint. In practice, this means time-consuming and expensive process parameter (e.g. speed, current, frequency of IW coil) exploration and selection, as well as limitations in laminate design changes once the parameter sets are established [2]. In situ process monitoring and control

provides an alternative approach whereby manufacturing process can be optimized online. Recent work by the researchers in [3] demonstrated that in-situ temperature monitoring of weld-line in a IW joining process can be used to provide live feedback to an operator, who can then modify the IW coil speed to maintain a pre-set temperature and improve weld quality. However, such an approach requiring manual intervention is not scalable, and a more practical automated feedback control framework is required.

In this paper we explore Deep Reinforcement Learning (DRL) framework for optimization of the IW process. Recently, DRL has emerged as a promising machine learning approach for automated manufacturing process control and planning, see for example [4, 5, 6, 7]. In particular, DRL framework has been successfully applied to high-dimensional and nonlinear optimal control problems. Moreover, once DRL control policy has been trained offline, it can be efficiently executed online, and thus provides an attractive approach for real-time process control.

The main contributions of this paper are as follows. Firstly, we formulate an optimal control problem for optimizing IW coil speed such that the desired spatiotemporal thermal behavior is achieved along the weld-line. This optimization problem involves temperature history dependent terms, and turns out to nonlinear, non-convex and non-smooth, and thus computationally challenging to solve. To address these challenges, we develop a novel framework to transform the history dependent

formulation into the standard Markov form, and apply proximal policy optimization (PPO) based DRL approach to solve the optimal control problem. Finally, via simulation studies we show that the resulting DRL control policy can result in better spatiotemporal thermal response compared to the current state-of-the-practise in which a constant speed profile is utilized.

2. Methodology

In this section we describe a fast acting dynamic IW model, formulate an optimal control problem, and develop a DRL based framework to solve the optimal control problem.

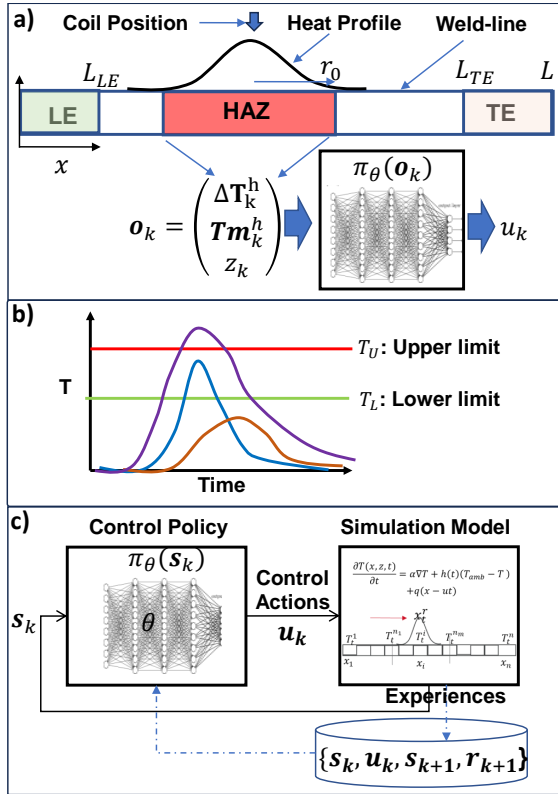


Figure 1: a) Schematic showing the leading edge (LE), trailing edge (TE) and heat affected zone (HAZ) along the weld-line. Also shown is the coil heat source profile at the coil location, and control policy parameterizations with input as the observation vector \mathbf{o}_k and output as the coil speed u_k . b) Notional temperature time curves (magenta, blue and orange) at three locations on a weld-line along with the desired minimum (green) and maximum (red) temperature limits. The magenta and orange curve lead to poor quality weld at their respective locations, while the blue curve results in a good quality weld. c) Schematic of the DRL framework.

2.1. Fast Acting Induction Welding Model

High-fidelity finite element (FE) simulation model for IW process already exists in literature, see [8] and the references therein. These high-fidelity models are not computationally amenable for real-time process optimization, and one has to rely on fast acting dynamic models. Such fast acting models are computationally cheap to execute at the expense of representing process physics only approximately. In this work we leverage a 1D spatio-temporal thermal simulation model developed in [2]. Specifically, a nonlinear 1D heat partial differential equation (PDE) is used to capture the spatiotemporal thermal behavior along the weld-line,

$$\frac{\partial T(x, t)}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} + q(x - x^r(t)) + \sigma\beta(T_a^4 - T^4) + h(x, x^r(t))(T_a - T), \quad (1)$$

where, $T(x, t)$ is the time-dependent temperature field, $x \in [0, L]$ with L being the length of the weld-line. We assume Dirichlet boundary condition with $T(0, t) = T(L, t) = T_a$, where T_a is the ambient temperature. Here, α is thermal diffusivity, σ is the Stefan-Boltzmann constant, and β is the emissivity which all depend on material properties of the TPC involved. The last two terms in (1) capture the radiative and convective heat transfer, respectively. The convective heat transfer coefficient $h(x, x^r(t))$ is position dependent and also depends on the coil position to capture IW configurations where forced convection is used to cool the already welded regions. To reflect that, we take,

$$h(x, x^r(t)) = \begin{cases} h_f & x \leq x^r(t), \\ h_0 & x > x^r(t), \end{cases} \quad (2)$$

where, h_0 is free convective heat transfer coefficient, and h_f is forced convective heat transfer coefficient. Following the approach used in Rosenthal model [9], we approximate the moving heat source term $q(x - x^r(t))$ (i.e. the induction coil) in form of a Gaussian heat profile,

$$q(x - x^r(t)) = q_0 e^{-\left(\frac{x - x^r(t)}{r_0}\right)^2}, \quad (3)$$

where, q_0 is the heat source peak intensity, r_0 defines the spatial extent of the heat affected zone (HAZ), and $x^r(t)$ indicates source location at time t , see Fig. 1a. The model parameters such as q_0 , r_0 and h_0 , h_f are typically not known exactly. They can be estimated via calibration process where a nonlinear least-squares optimization problem is solved to match the thermal history output of the model with experimental data, see [10] for the details.

For our application, we discretize the PDE (1) in space and time along the weld-line leading to,

$$\begin{aligned} T_{k+1}^i &= T_t^i + \alpha \Delta t \frac{T_k^{i+1} - 2T_k^i + T_k^{i-1}}{(\Delta x)^2} \\ &+ \Delta t q_0 e^{-\left(\frac{x_i - x_k^r}{r_0}\right)^2} + \Delta t \sigma \beta (T_{amb}^4 - (T_k^i)^4) \\ &+ \Delta t h(i \Delta x, x_k^r) (T_a - T_k^i), \end{aligned} \quad (4)$$

$$x_{k+1}^r = x_k^r + u_k \Delta t, \quad (5)$$

where, Δx is the spatial grid size, Δt is the time step, T_k^i is temperature at the grid location $x_i = i \Delta x$, $i = 1, \dots, N$ at time $t = k \Delta t$, and x_k^r is the coil position at time $t = k \Delta t$ moving with speed u_k . Since we use Dirichlet condition at the boundaries, that implies $T_{-1}^k = T_a$ and $T_t^{N+1} = T_a$. We express (4)-(5) more compactly as,

$$\mathbf{T}_{k+1} = \mathbf{f}(\mathbf{T}_k, x_k^r, u_k), \quad (6)$$

$$x_{k+1}^r = x_k^r + u_k \Delta t, \quad (7)$$

where, $\mathbf{T}_k = (T_k^1, \dots, T_k^N)'$ denotes a vector of temperature values along the grid line, and \mathbf{f} represents the right hand side of (4) expressed in a vector form. Here, \prime denotes vector transpose.

2.2. Optimal Control Problem Formulation

Given the goal is to design coil speed sequence $\{u_0, u_1, \dots, u_K\}$ such that the maximum temperature at every location on the weld-line reaches within the desired range $[T_L, T_U]$, we formulate following optimal control problem:

$$\max_{\{u_k, \mathbf{T}_k, x_k^r\}_{k \in \mathcal{I}_K}} \sum_{k=0}^K \sum_{i=1}^N \frac{1}{2} \left(1 + \frac{T_k^i - T_L}{T_U - T_L} \right) \mathbb{I}_{\{T_k^i > T_L\}}, \quad (8)$$

subject to following constraints:

$$\mathbf{T}_{k+1} = \mathbf{f}(\mathbf{T}_k, x_k^r, u_k), \quad k \in \mathcal{I}_K, \quad (9)$$

$$x_{k+1}^r = x_k^r + u_k \Delta t, \quad k \in \mathcal{I}_K, \quad (10)$$

$$\max_{k \in \mathcal{I}_K} T_k^i \leq T_U, \quad i \in \mathcal{I}_g, \quad (11)$$

$$\max_{k \in \mathcal{I}_K} T_k^i \geq T_L, \quad i \in \mathcal{I}_s, \quad (12)$$

$$u_k \in U, \quad k \in \mathcal{I}_K, \quad (13)$$

$$L \leq x_{K+1}^r \leq L + \delta, \quad (14)$$

where, $\mathcal{I}_K = \{0, \dots, K\}$ is the time horizon, $\mathcal{I}_s \subset \mathcal{I}_g = \{1, \dots, N\}$ is subset of grid locations and $\mathbb{I}_{a>b}$ is an indicator function with value 1 if $a > b$ and 0 otherwise. Each term in the objective function (8) increases linearly as the maximum temperature reaches the upper limit T_U , and attains a zero value for grid location where the maximum temperature does not exceed

the lower limit T_L . Thus, the objective function attains higher values when the temperature dwells for longer duration near the upper temperature limit T_U while not exceeding it (due to constraint (11)). The constraints (9-10) correspond to the dynamic evolution as discussed in the previous Section 2.1; constraint (11) enforces that the maximum temperature is below the upper threshold T_U everywhere along the weld-line; constraint (12) enforces that the maximum temperature is above the lower threshold T_U over the specified region \mathcal{I}_s along the weld-line; constraint (13) enforces that the coil speed is in the feasible range $U = [u_{min}, u_{max}]$; and constraint (14) makes sure that at the final time step $K + 1$ the coil has reached the end of the weld-line. Here δ is a slack parameter which can be set to $u_{max} \Delta t$, and ensures that the coil does not keep moving after it has reached the end of the weld-line.

Finally, note that above optimization problem is high dimensional (given $N, K \gg 1$), nonlinear, non-convex and non-smooth, the final time step K is not known a priori, and finding an initial feasible solution can be difficult. This makes it a computationally challenging problem to solve via traditional optimal control approaches such as model predictive control [11]. We explore DRL framework to address these challenges.

2.3. Deep Reinforcement Learning

2.3.1. DRL Overview

DRL has recently shown unprecedented success in dealing with high dimensional nonlinear optimal control/decision making problems [12, 13, 14, 15, 16, 17]. DRL is a reinforcement learning approach which is used to compute optimal control/decision policy for decision making problems modeled via Markov Decision Processes (MDP). In DRL, the value function/policy are parameterized in form of neural networks (NN) and trained using data sampled via interaction with an environment (e.g., simulation model or real system).

For our application we use proximal policy optimization (PPO) [16] which is the state-of-the-art DRL method. PPO uses two NNs: the actor or the control policy network $\pi_\theta(\mathbf{u}|\mathbf{s})$ with parameters θ outputs the conditional probability of taking an action \mathbf{u} given the state \mathbf{s} , and the critic network $V_\phi(\mathbf{s})$ with parameters ϕ returns the corresponding expectation of the discounted long-term reward. PPO being a model-free on-policy method uses a type of policy gradient training that alternates between sampling data via interacting with an environment (e.g., simulation model) and optimizing the actor/critic networks parameters via stochastic gradient approach. PPO is a simplified form of Trust Region

Policy Optimisation (TRPO) [17] approach and uses a clipped surrogate objective function for optimizing the actor network to improve training stability by limiting the size of the policy change at each learning iteration.

In each learning episode agent interaction with the environment generates samples $\{\mathbf{s}_k, \mathbf{s}_{k+1}, \mathbf{u}_k, r_{k+1}\}_{t=0}^P$, which includes the current state \mathbf{s}_k , the next state $\mathbf{s}_{k+1} \sim \mathcal{T}(\cdot|\mathbf{s}_k, \mathbf{u}_k)$ resulting from taking the control action \mathbf{u}_k sampled from the current policy $\mathbf{u}_k \sim \pi_\theta(\cdot|\mathbf{s}_k)$, and the corresponding immediate reward r_{k+1} for transitioning from \mathbf{s}_k to \mathbf{s}_{k+1} (see Fig. 1c). Note, here \mathcal{T} is the transition function (e.g., implemented via a simulation model (4-5) which is Markov, i.e., depends only on the current state \mathbf{s}_k and action \mathbf{u}_k). For each pair $(\mathbf{s}_k, \mathbf{u}_k)$, these samples are used to evaluate the corresponding return $G_k = D_k + V_\phi(\mathbf{s}_k)$ and the generalized advantage function (GAE) D_k ,

$$D_k = \sum_{j=k}^{k+H} (\gamma\lambda)^{j-k} (r_{j+1} + d\gamma(V_\phi(\mathbf{s}_{j+1}) - V_\phi(\mathbf{s}_j))), \quad (15)$$

over selected experience horizon H . Here γ is the discount factor, λ is a smoothing GAE factor and $d = 1$ unless \mathbf{s}_{k+H} is terminal state in which case $d = 0$. A mini-batch data of size M is then sampled from these experience tuples $\{\mathbf{s}_i, \mathbf{u}_i, G_i, D_i\}_{i=1}^M$, and is used to update the critic parameters ϕ by minimizing the loss $L_c(\phi) = \frac{1}{2M} \sum_{i=1}^M (G_i - V_\phi(\mathbf{s}_i))^2$. Similarly, the actor parameters θ are updated by minimizing the loss function $L_a(\theta)$,

$$L_a(\theta) = \frac{1}{2M} \sum_{i=1}^M (-\min(b_i(\theta)D_i, c_i(\theta)D_i) + w\mathcal{H}(\pi_\theta(\cdot|\mathbf{s}_i))), \quad (16)$$

where,

$$b_i(\theta) = \frac{\pi_\theta(\mathbf{u}_i|\mathbf{s}_i)}{\pi_{\theta_{old}}(\mathbf{u}_i|\mathbf{s}_i)}, \quad c_i(\theta) = \max(\min(b_i(\theta), 1+\epsilon), 1-\epsilon), \quad (17)$$

with θ_{old} being the actor parameters from the previous learning epoch, and c_i is the clipping function with clip factor $\epsilon < 1$. Here, \mathcal{H} is the entropy loss function which promotes exploration, and $w \geq 0$ is a weight factor for the entropy loss.

Once the DRL policy π_θ is trained, it can be used to compute the feedback control actions $\mathbf{u} \sim \pi_\theta(\cdot|\mathbf{s})$ based on the current state \mathbf{s} , see Fig. 1c.

2.3.2. DRL Formulation of IW Optimal Control Problem

The standard DRL framework as discussed in the previous section cannot be directly applied to solve the optimal control problem (8) since: i) it requires keeping

track of entire temperature-time history at each grid location leading to a non-Markov setting; and ii) involves state/action constraints (11-14). Note that the dynamic constraints (9-10) are naturally handled by the DRL framework. To address these challenges, we propose a novel DRL formulation involving carefully selected state definition, reward function and episode termination conditions.

From the form of history dependent quantities appearing in (8), it can be noticed that to transform the problem to a standard Markov form it is sufficient to keep track of maximum temperature up till current time step k , i.e.,

$$Tm_k^i = \max\{T_j^i : j = 0, \dots, k\}, \quad (18)$$

at each grid location $i \in \mathcal{I}_g$. The state vector \mathbf{s}_k then comprises of the components $\mathbf{T}_k, \mathbf{T}_{k-1}, \mathbf{Tm}_k, x_k^r$, where $\mathbf{Tm}_k = (Tm_k^1, \dots, Tm_k^N)'$ and note we have included both the current temperature vector \mathbf{T}_k and the temperature vector at pervious time step \mathbf{T}_{k-1} . Then given the coil speed u_k , the state vector $\mathbf{s}_k = (\mathbf{T}'_k, \mathbf{T}'_{k-1}, \mathbf{Tm}'_k, x_k^r)'$ evolves to $\mathbf{s}_{k+1} = (\mathbf{T}'_{k+1}, \mathbf{T}'_k, \mathbf{Tm}'_{k+1}, x_{k+1}^r)'$ as follows,

$$\mathbf{T}_{k+1} = \mathbf{f}(\mathbf{T}_k, x_k^r, u_k), \quad (19)$$

$$\mathbf{Tm}_{k+1} = \max\{\mathbf{Tm}_k, \mathbf{T}_{k+1}\}, \quad (20)$$

$$x_{k+1}^r = x_k^r + u_k \Delta t, \quad (21)$$

where, given two vectors $\mathbf{w}_1, \mathbf{w}_2 \in R^N$, $\max\{\mathbf{w}_1, \mathbf{w}_2\} = (\max(w_{11}, w_{21}), \max(w_{12}, w_{22}), \dots, \max(w_{1N}, w_{2N}))'$ is an element wise maximum.

Given this Markovian state evolution formulation, we next define observation vector, action space, reward function, and episode termination condition. The schematic of the DRL framework is illustrated in the Fig. 1, where Fig. 1a shows the observation and action space, Fig. 1b gives examples of temperature-time curves to help motivate the reward function definition, and Fig. 1c shows the DRL training process as outlined in the previous section.

Observations and Actions. Since the heating effect of coil is local, rather than using the full state \mathbf{s}_k for the weld-line as input for the actor/critic networks in PPO, a smaller observation vector \mathbf{o}_k is chosen to capture the thermal state in HAZ around the current location x_k^r of the coil, see Fig. 1a. The HAZ zone is defined as a region spanning $[\max(0, x_k^r - L_{HAZ}), \min(L, x_k^r + L_{HAZ})]$ which can be translated into a set of grid indices $\mathcal{I}_k^h \subset \mathcal{I}_g$. For instance, L_{HAZ} can be chosen to be r_0 , the extant of spread of Gaussian heat source, see Eqn. (3). Furthermore, we expect that the control policy will be different

near the leading edge (LE) and trailing edge (TE) compared to the middle sections of the weld-line. We define LE and TE regions as $[0, L_{LE}]$ and $[L_{TE}, L]$, respectively (see Fig. 1a).

Given these considerations, we define $\mathbf{o}_k = g(\mathbf{s}_k)$ as a vector $\mathbf{o}_k = ((\Delta \mathbf{T}^h)', (\mathbf{Tm}^h)', z_k)'$ comprising of the following components extracted from the state \mathbf{s}_k : maximum temperature $\mathbf{Tm}^h = \{Tm_k^i, i \in \mathcal{I}^h\}$ and time derivatives of temperature $\Delta \mathbf{T}^h = \{\frac{T_k^i - T_{k-1}^i}{\Delta t}, i \in \mathcal{I}^h\}$, and the coil position z_k which is defined as,

$$z_k = \begin{cases} -1 & x_k^r \in [0, L_{LE}], \\ 1 & x_k^r \in [L_{TE}, L], \\ 0 & \text{otherwise.} \end{cases} \quad (22)$$

To capture the constraint (13), we restrict the action u_k to a discrete set of values $\{u_1^d, \dots, u_m^d\} \subset U$.

Reward Function. Motivated by the objective functional form in (8) we construct a reward shaping function,

$$r(\mathbf{s}_k) = \frac{1}{N} \sum_{i=1}^N r_g(T_k^i), \quad (23)$$

which for each grid location assigns a reward,

$$r_g(T) = \begin{cases} 0 & T < T_L, \\ \frac{1}{2} \left(1 + \frac{T - T_L}{T_U - T_L}\right) & T_L \leq T \leq T_U, \\ -1 & T > T_U. \end{cases} \quad (24)$$

Episode Termination Condition. We model the constraints (11), (12) and (14) in form of episode termination conditions which ends a learning episode if any of these constraints is violated, thereby indirectly penalizing control policy leading to an undesirable temperature-time behavior. Given the current state $\mathbf{s}_k = (\mathbf{T}'_k, \mathbf{T}'_{k-1}, \mathbf{Tm}'_k, x_k^r)'$, an episode is terminated if any of the following conditions are met:

- Coil reaches the end of the weld-line, i.e., $x_k^r > L$ (corresponding to the constraint (14)).
- If maximum temperature Tm_k^i reaches beyond upper limit, i.e. $Tm_k^i > T_U$ at any grid location $i = 1, \dots, N$ (corresponding to the constraint (11)), implying that the welding has failed (see the magenta temperature-time curve in the Fig. 1b).
- If maximum temperature Tm_k^i starts reducing before hitting the lower limit, i.e.

$$Tm_k^i < T_L, \quad \frac{T_k^i - T_{k-1}^i}{\Delta t} < 0, \quad (25)$$

at any grid location $i \in \mathcal{I}_s$ (corresponding to the constraint (12)), again implying that the welding has failed (see the orange temperature-time curve in the Fig. 1b).

For reference also shown in the Fig. 1b is a blue temperature-time curve which does not violate these conditions.

3. Simulation Results

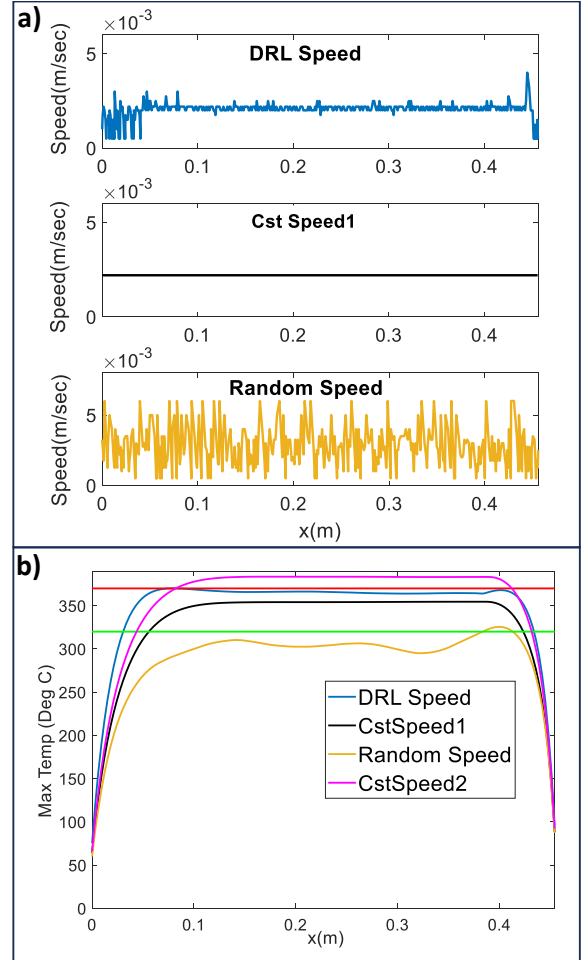


Figure 2: a) DRL, constant speed and random speed profiles along the weld-line. b) Maximum temperature reached along the weld-line using these different speed profiles. Also shown are the lower (green) and upper (red) temperature limits.

For simulation demonstration we use a flat single lap and curved double joint architecture for roller-based heat up for the IW experimental setup [10]. The chosen material system-of-interest for this work is the

Toray Cetex TC1225 LMPAEEK thermoplastic, intermediate modulus carbon fiber composite tape system. The minimum and maximum temperature limit for desirable weld for this material system was taken to be $T_L = 325^\circ\text{C}$ and $T_U = 375^\circ\text{C}$, respectively. The length of the panel was $L = 0.45\text{m}$ to match the experimental setup. To define the LE and TE regions we assumed $L_{LE} = 0.25L$ and $L_{TE} = 0.75L$, respectively. The coil speed was constrained to lie within the range $[0.5, 6] \times 10^{-3}$ m/sec. We used $\Delta t = 0.1\text{sec}$ and $\Delta x = 0.003\text{m}$ for the model discretization in equations (4-5). The model parameters $\alpha, \beta, h_f, h_o, q_0, r_0$ were calibrated to the experimental setup using empirically collected thermocouple temperature data, see [10] for the details.

We use the PPO implementation in MATLAB R2021b for training the DRL control policy on the calibrated model. The hyperparameters used were tuned manually with following values leading to good training performance: discount factor $\gamma = 0.99$, GAEFactor $\lambda = 0.95$, clipfactor $\epsilon = 0.15$, cross-entropy loss weight $w = 0.01$, mini-batch size $M = 256$, and experience horizon $H = 1024$. For each episode the initial temperature values were assumed to be uniform $T_0^i = T_0, i = 1, \dots, N$ along the weld-line where, T_0 was sampled from a normal distribution with mean as the room temperature $T_a = 25^\circ\text{C}$ and variance 5%. While executing DRL feedback control policy during testing we assume that the observation vector (see Section 2.3.2) is available from the simulation model. In real setup, this information will need to be derived from insitu sensors such as thermocouples, pyrometers or thermal cameras.

Figure 2 shows the performance of the trained DRL control policy and comparison with constant speed solution currently used in practice, and a random speed profile. Figure 2a shows the speed profile along the weld-line, and the Fig. 2b shows the maximum temperature reached at each grid location on the weld-line for each of three approaches. The DRL speed profile may appear random at first glance (and therefore we included a random speed profile for comparison) but has clear trends: lower speed at leading edge and medium speed at middle/trailing edge locations. Indeed, as can be seen, the blue curve which represents maximum temperature achieved by DRL control policy is closest to upper limit as desired, compared to random speed profile which is not able to achieve desired maximum temperature along the weld-line. Furthermore, compared to the constant speed profile (CstSpeed1, see the black curve in the Fig. 2b) the DRL control policy achieves slightly better heating characteristics also at the leading

and trailing edges. To achieve similar behavior at the leading/trailing edge using uniform speed profile, one will need to further reduce the speed but that causes the maximum temperature to exceed in the middle section of the weld-line (CstSpeed2, see the magenta curve in the Fig. 2b). The DRL policy is able to automatically learn the nonlinear spatiotemporal effect of panel edges/boundary conditions and panel geometry on the TPC thermal response, and compensate for it by adjusting the speed profile accordingly. Overall it is promising that DRL learns non-intuitive speed profile modulations which pushes the temperature closer to upper limit while still not exceeding it.

4. Conclusions

We formulated an optimal control problem for optimizing coil speed for an IW process modeled via 1D heat equation with a moving heat source. To address computational challenges associated with solving the optimal control problem, we applied state-of-the-art DRL framework which involved a novel approach to transform the history dependent problem formulation into the standard Markov form. Numerical study demonstrated that the DRL control policy results in a better performance, i.e. desired temperature-time behavior along the weld-line, compared to the current state-of-the-practise of using a pre-determined uniform coil speed.

In the future we plan to experimentally implement and validate the proposed framework. In that regard, it will also be necessary to explore preferably non-contact real-time in situ sensing approaches and integrate with the feedback DRL policy. It will also be worthwhile to extend the framework, e.g. via domain randomization/transfer learning approaches, so that DRL policy reliably generalizes across different panel/weld-line geometries, boundary conditions and material properties.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Ad-

vanced Manufacturing Office, Award Number DE-EE0009398. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

- [1] T. Ahmed, D. Stavrov, H. Bersee, A. Beukers, Induction welding of thermoplastic composites—an overview, *Composites Part A: Applied Science and Manufacturing* 37 (10) (2006) 1638–1651.
- [2] W. Zhao, J. Alms, B. Blakeslee, A. Chakraborty, J. Gangloff, M. Klecka, J. Mendoza, Z. Wang, L. Xing, Automated induction welding of large thermoplastic composite structure, *SAMPE*, <https://doi.org/10.33599/nasampe/s.22.0867> (2022).
- [3] N. A. A. Rahim, J. Pandher, N. Coppola, V. Penumetsa, M. van Tooren, In-situ monitoring and control of induction welding in thermoplastic composites using high definition fiber optic sensors, *The Composites and Advanced Materials Expo, CAMX, Anaheim, CA* (2019).
- [4] T. Toner, M. Saez, D. M. Tilbury, K. Barton, Opportunities and challenges in applying reinforcement learning to robotic manipulation: An industrial case study, *Manufacturing Letters* 35 (2023) 1019–1030.
- [5] M. Panzer, B. Bender, Deep reinforcement learning in production systems: a systematic literature review, *International Journal of Production Research* 60 (13) (2022) 4316–4341.
- [6] A. Surana, K. Reddy, Guided policy search based control of a high dimensional advanced manufacturing process, in: *2022 IEEE Conference on Control Technology and Applications (CCTA), IEEE, 2022*, pp. 1415–1420.
- [7] A. Surana, Reinforcement learning: An industrial perspective, in: *Handbook of Reinforcement Learning and Control*, Springer, 2021, pp. 647–672.
- [8] F. Lionetto, S. Pappadà, G. Buccoliero, A. Maffezzoli, Finite element modeling of continuous induction welding of thermoplastic matrix composites, *Materials & Design* 120 (2017) 212–221.
- [9] T. Eagar, N. Tsai, et al., Temperature fields produced by traveling distributed heat sources, *Welding Journal* 62 (12) (1983) 346–355.
- [10] J. Gangloff, W. Zhao, S. Sarkar, S. Mondal, L. Xing, A. Chakraborty, A. Surana, B. Bedard, J. Alms, Multi-source machine learning and thermoplastics enhanced aerospace manufacturing, *SAMPE*, <https://doi.org/10.33599/nasampe/s.23.0021> (2023).
- [11] J. A. Rossiter, *Model-based predictive control: a practical approach*, CRC press, 2017.
- [12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, *Nature* 518 (7540) (2015) 529–533.
- [13] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of go without human knowledge, *nature* 550 (7676) (2017) 354–359.
- [14] Y. Duan, X. Chen, R. Houthoofd, J. Schulman, P. Abbeel, Benchmarking deep reinforcement learning for continuous control, *CoRR abs/1604.06778* (2016).
- [15] S. Levine, C. Finn, T. Darrell, P. Abbeel, End-to-end training of deep visuomotor policies, *The Journal of Machine Learning Research* 17 (1) (2016) 1334–1373.
- [16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, *arXiv preprint arXiv:1707.06347* (2017).
- [17] J. Schulman, S. Levine, P. Abbeel, M. Jordan, P. Moritz, Trust region policy optimization, in: *International conference on machine learning*, PMLR, 2015, pp. 1889–1897.