Fault Detection and Response for Safe Control of Artificial Muscles in Soft Robots

Ran Jing[1*], *Graduate Student Member, IEEE*, Charles Van Hook[1*], Ilyoung Yang[1], Andrew P. Sabelhaus [1,2], *Member, IEEE*

***Abstract*— Robots built from soft materials have the potential for intuitively-safer interactions with humans and the environment. However, soft robots' embodiments have many sources of failure that could lead to unsafe conditions in closed-loop control, such as degradation of sensors or fracture of actuators. This article proposes a fault detection system for sensors attached to artificial muscle actuators that satisfies a formal safety condition. Our approach combines redundant sensing, model-based state estimation, and Gaussian process regression to determine when one sensor's reading statistically diverges from another, indicating a fault condition. We apply the approach to electrothermal shape memory alloy (SMA) artificial muscles, demonstrating that our method prevents the overheating and fire damage risk that could otherwise occur. Experiments show that when the muscle's nominal sensor (temperature via a thermocouple) is fractured from the robot, the redundant sensor (electrical resistance) combined with our method prevents violation of state constraints. Deploying this system in real-world human-robot interaction could help make soft robots more robust and reliable.**

***Index Terms*— Fault Detection, Soft Robotics, Eletrothermal Actuators, Fault Tolerant Control.**

I. Introduction

Soft robots have the potential to overcome fundamental challenges in human-robot interaction by reducing reliance on modeling and sensing. For a soft robot, mistakes are often less severe: for example, soft materials apply lower forces upon impact with humans [1]. This concept has been extended to formally-safe control in soft robots, such as invariance of critical states [2], limits on forces [3], [4], or anti-collision [5]. These techniques sidestep the challenges [6] with feedback control of soft systems [7] by optimizing for constraint satisfaction rather than tracking error.

However, safety-verified control systems can fail in practice, particularly with soft robots. Soft actuators can degrade [8] or fail catastrophically due to puncture [9] or overheating [10], and are prone to model mismatch [11]. These phenomena are *faults*: a component of the system behaves differently than its model for controllers, therefore drives the system into an unsafe state. For example, a temperature sensor might disconnect from its fixture [12] in a thermal shape memory alloy (SMA) artificial muscle when in contact with a human (Fig. 1). No prior work incorporates detection and recovery of such faults into formal safety for soft robot control, as prior work is not verifiable [13]–[15], or only focuses on tolerance to partial actuation failures [16], [17].

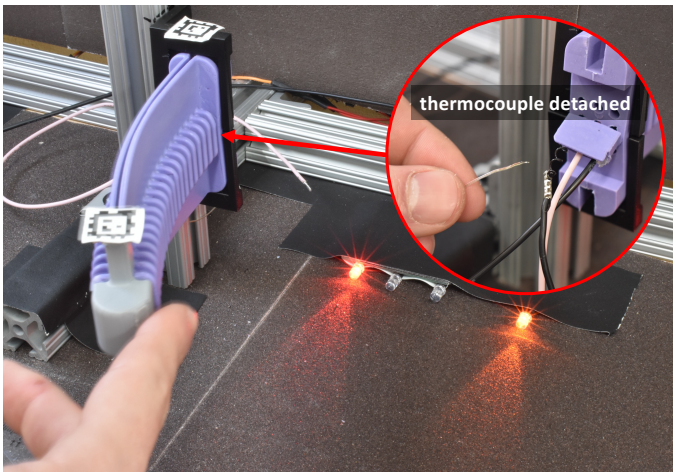

Fig. 1. Our approach detects faults in a sensor in soft robots, such as fracture of a thermocouple in an electrothermal artificial muscle, and responds in closed-loop to ensure invariance of safety-critical states. Experiments show 100% recall on fault triggers (red LED).

This manuscript proposes redundant self-sensing in soft robots to detect and respond to faults. Our insight is that smart actuators often have additional signals that indicate their internal state, such as electrical resistance in shape memory materials [18]. Since the relationship between these redundant signals and actuator state is often challenging to model, we propose a learning method (Gaussian Process Regression [19]) to predict the state. Our approach detects faults by comparing predictions against a nominal sensor. We provide a proof of set invariance under certain conditions. The method is applied to the shape memory muscle in Fig. 1, demonstrating a 100% recall on detection of catastrophic fracture of a thermal sensor.

This article contributes:

- A formalization of sensor faultiness, via divergence in nominal vs. redundant estimators, as a *safety margin*,
- An analysis of this method's assumptions in a formal proof of invariance during safety-critical control,
- A validation on electrothermal soft robot muscles in both simulation and hardware.

To our knowledge, this manuscript establishes the first generalizable approach to detecting and responding to faults in soft artificial muscles that maintains a safety verification.

This work was in part supported by the U.S. National Science Foundation under Award No. 2340111 and 2209783.

[1]R. Jing, C. Van Hook, I. Yang, and A.P. Sabelhaus are with the Department of Mechanical Engineering, Boston University, Boston MA, USA. {rjing, cvanhook, alviny21, asabelha}@bu.edu.

[2]A.P. Sabelhaus is also with the Division of Systems Engineering, Boston University, Boston MA, USA.

*Equal Contribution.

II. Background and Prior Work

Many past techniques have proven stability of feedback control with redundant sensing in the presence of faults for well-characterized systems [20], as well as responses that satisfy a set-invariance safety condition [21], [22]. These methods work well with actuator failure in traditional machines, but unfortunately, accurate low-order models are often unavailable in smart materials for soft robots [10], particularly when hysteresis is present [12].

In contrast, machine learning can detect faults in systems that are difficult to model [23]. When a learned fault detector is sufficiently accurate, control barrier functions and their neural variants could be used for verifiable safety [24]–[26]. However, learned soft robot models are notoriously imperfect due mismatch between infinite-dimensional state spaces and finite-dimensional sensing [27].

Rather than directly classifying a signal as faulty or not [28] or relying on model-based approaches, we propose a conservative combination of all the above. Our approach assumes a nominal model for the nominal sensor, possibly faulty, and uses learning for the redundant self-sensing smart material. We rely on the well-defined statistical properties of Gaussian Process Regression [19], which captures the quality of feature selection in the redundant sensor signals.

III. Problem Setup and Approach

Our problem assumes an existing soft robot limb under feedback control [2], [8]. Fig. 2 shows the assumptions as well as application in a shape-memory alloy soft limb with redundant sensing as the muscle's electrical resistance [18].

A. Equations of Motion and Existing Conditions

Consider a robot actuated by artificial muscles whose states are safety-critical. If the full state of the robot is $\mathbf{x} = \begin{bmatrix} \dots & w \end{bmatrix}^\top \in \mathbb{R}^N$, we denote $w \in \mathbb{R}$ the artificial muscle state without loss of generality (i.e., our method can be applied to M-many artificial muscles independently, $w_1 \dots w_M$). Many soft artificial muscles are such scalar systems [29]. The safety constraint is $w \in \mathcal{S} \subset \mathbb{R}$.

We assume there is some nominal dynamics of the actuator (muscle) state, with a scalar control input,

$$w(t+1) = f(w(t), u(t)) \tag{1}$$

and that there exists a feedback controller that ensures safety under idealized measurements of w, applying $u(t) = u^*(t)$,

$$u^*(\mathbf{x}) : \; w(0) \in \mathcal{S} \Rightarrow w(t) \in \mathcal{S} \; \forall t. \tag{2}$$

Nominal sensors give measurements as $\mathbf{x}_t$, which contains w_t modeled as additive noise: $w_t = w(t) + \varepsilon_w$, with $\varepsilon_w \sim \mathcal{N}(0, \sigma_w^2)$. The idealized control law is assumed to be probabilistically safe to n-many standard deviations:

Assumption 1: $u = u^*(\mathbf{x})|_{w=w_t} \Rightarrow w \in \mathcal{S}$ for $w_t \geq w(t) - n\sigma_w$, i.e., the true state is under-measured by n standard deviations. We are implicitly choosing an upper bound as the safe criterion, $\mathcal{S} = \{w | w < w^{max}\}$, and that the actuator dynamics are a *monotonic control system:*

Assumption 2: $u_1 < u_2 \Rightarrow w_1(t+1) < w_2(t+1)$ where $w_i(t+1) = f(w(t), u_i)$.

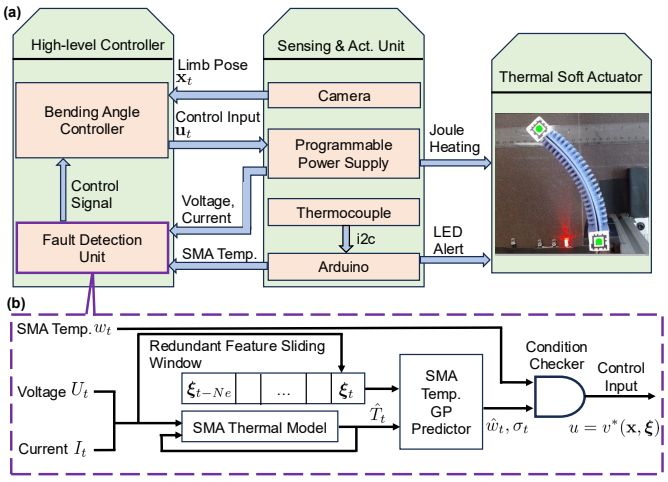

Fig. 2. This manuscript assumes a system architecture for a soft robotic limb with (a) controllers for the limb's pose states $\mathbf{x_t}$ using control inputs to the artificial muscles $\mathbf{u_t}$, estimating a nominal signal $\boldsymbol{w_t}$ for the muscles' low-level states, and redundant sensors $\boldsymbol{\xi_t}$. For shape memory muscles in a planar limb, these are correspondingly the limb's bending angle, electrical power (voltage), muscle temperature, and muscle electrical resistance. Our fault detection system (b) estimates a $\hat{\boldsymbol{w}}_{\boldsymbol{t}}$ using Gaussian Process Regression over a featurized sequence $\boldsymbol{\xi}_{\boldsymbol{t-N_D:t}}$ and applies a safety-ensured input $\bar{\mathbf{u}}$ if estimates diverge larger than some variance.

This is a common setup in artificial muscles: higher voltages lead to more force in dielectrics, higher pressures similarly in pneumatics, and larger currents to higher temperatures in electrothermal muscles.

B. Sensor Fault Problem Statement

Now consider a fault condition on the sensors, as in:

$$w_t = w_t^F < w(t) - n\sigma_w, \tag{3}$$

i.e., our sensor reports a faulty reading lower than can be bounded by the nominal safe controller.

The proof for the nominal controller then fails: $u^*(\mathbf{x})|_{w=w_t^F} \not\Rightarrow w(t+1) \in \mathcal{S}$, i.e., the controller may send the system beyond w^{max}. We seek to detect this condition and respond.

Assume there is an input that is always safe to the muscle, for example $u = \bar{u} = 0$, such that $w(t+1) \in \mathcal{S} \; \forall w(t) \in \mathcal{S}$ if $w(t+1) = f(w(t), \bar{u})$. We next denote the redundant sensing information about the muscle's state as $\boldsymbol{\xi} \in \mathbb{R}^P$, such as electrical resistance and hysteresis calculations.

We define a *safety margin* ρ that uses $\boldsymbol{\xi}$ in addition to w_t, as $\rho(w, \boldsymbol{\xi}) : \mathbb{R} \times \mathbb{R}^P \mapsto \mathbb{R}$. The proposed fault-aware probabilistically-safe controller is:

$$u = v^*(\mathbf{x}, \boldsymbol{\xi}) = \begin{cases} \bar{u} & \text{if } \rho(w, \boldsymbol{\xi}) < \rho_L \\ u^*(\mathbf{x}) & \text{else.} \end{cases} \tag{4}$$

with $\rho_L \in \mathbb{R}$ a lower limit on safety that triggers a fault. We seek a $\rho(\cdot)$ and ρ_L such that $w(t+1) \in \mathcal{S}$, $w(t+1) = f(w(t), v^*(\mathbf{x}_t, \boldsymbol{\xi}_t))$. Propose the safety margin as the difference between the nominal and redundant estimates,

$$\rho_t := \rho(w_t, \boldsymbol{\xi}_t) := w_t - \hat{w}(\boldsymbol{\xi}_t) \tag{5}$$

where $\hat{w}$ is an estimate of the *best case possible actuator state* per the redundant sensors, to be synthesized later. Intuitively, when $\rho_t - \rho_L < 0$, the nominal sensor reading is lower than

the most generous possible value per the redundant sensors, and shutoff occurs.

C. Gaussian Process Regression for State Estimation

Our framework predicts the nominal sensor outputs using a Gaussian Process Regressor. With measurement information $\boldsymbol{\xi} \in \mathbb{R}^P$, the GP assumes an underlying function $f_\xi(\cdot) : \mathbb{R}^P \mapsto \mathbb{R}$ that deterministically defines the sensor's output $w(t)$ from $\boldsymbol{\xi}$, and that additive noise corrupts this measurement: $w = f_\xi(\boldsymbol{\xi}) + \varepsilon_\xi$. Here, $\varepsilon_\xi \sim \mathcal{N}(0, \sigma_n^2)$, and so w is a random variable with some mean and variance at time t, i.e., we are re-estimating $w_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$ from $\boldsymbol{\xi}$.

To do so, GPs employ the *kernel trick* [19]. The kernel function $k(\cdot,\cdot) : \mathbb{R}^{P\times P} \mapsto \mathbb{R}$ models the covariance between the test points. Based on empirical evaluations, we select the squared exponential (SE) kernel plus white noise for more conservativeness:

$$k(\cdot,\cdot) = \sigma_f^2 \exp\left(-\frac{1}{2}(\boldsymbol{\xi}_a - \boldsymbol{\xi}_b)^\top \Lambda^{-1}(\boldsymbol{\xi}_a - \boldsymbol{\xi}_b)\right) + \sigma_s^2. \quad (6)$$

with the hyperparameters of signal variance σ_f^2, noise σ_s, and characteristic length scales $\Lambda = \text{diag}[\ell_1, \ldots, \ell_N]$ which determine the relative influence of each redundant sensor state on the estimator output. The SE kernel is known for its smoothness assumption and has seen past successes in modeling artificial muscle actuators [30], [31], and white noise reflects past experience with imperfect soft robot test setups [18]. As is standard [19], [32], hyperparameters are tuned via log-likelihood maximization.

Predicting μ_t and σ_t applies the kernel to our known datapoints and a test point. We have a precollected set of K-many datapoints as a time series, $\{\Xi, \mathbf{w}\}$, where $\Xi = [\boldsymbol{\xi}(0), \ldots, \boldsymbol{\xi}(K)] \in \mathbb{R}^{P\times K}$ and $\mathbf{w} = [w(0), \ldots, w(K)]^\top \in \mathbb{R}^K$. Assuming we have found a set of hyperparameters for the kernel, the prediction (via [19]) at a new test point $\boldsymbol{\xi}_t$ is

$$\mu_t = \mathbf{k}_{\boldsymbol{\xi}_t}^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{w}, \quad (7)$$

$$\sigma_t^2 = 1 - \mathbf{k}_{\boldsymbol{\xi}_t}^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{k}_{\boldsymbol{\xi}_t}, \quad (8)$$

where $\mathbf{k}_{\boldsymbol{\xi}_t} \in \mathbb{R}^K$ and $\mathbf{K} \in \mathbb{R}^{K\times K}$ are

$$\mathbf{k}_* = \begin{bmatrix} k(\boldsymbol{\xi}_*, \boldsymbol{\xi}_1) \\ \vdots \\ k(\boldsymbol{\xi}_*, \boldsymbol{\xi}_K) \end{bmatrix}, \mathbf{K} = \begin{bmatrix} k(\boldsymbol{\xi}_1, \boldsymbol{\xi}_1) & k(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) & \ldots \\ k(\boldsymbol{\xi}_2, \boldsymbol{\xi}_1) & k(\boldsymbol{\xi}_2, \boldsymbol{\xi}_2) & \ldots \\ \vdots & \ddots & \vdots \end{bmatrix}.$$

D. Safe Control with the GP Fault Detector

We propose $\hat{w}(\boldsymbol{\xi})$ and ρ_L based on the GP model:

$$\hat{w}(\boldsymbol{\xi}) = \mu_t(\boldsymbol{\xi}) - m\sigma_t(\boldsymbol{\xi}), \qquad \rho_L = -n\sigma_w \quad (9)$$

where m is the number of standard deviations desired for probabilistic safety. These satisfy Theorem 1.

Theorem 1: Fault-Aware Safety. Consider applying the fault-aware, probabilistically-safe controller in (4), using the safety margin of (5), estimator with fault guard of (9), in a problem statement where Assumption 1 (noise) and Assumption 2 (monotonicity) hold. If in addition:

3) The function $w = f_\xi(\boldsymbol{\xi})$ is stationary, i.e. a fixed value of $\boldsymbol{\xi}$ uniquely defines a w,
4) Training data are of sufficient quality that true value of the function $w(t)$ lies in the interval $\mu_t - m\sigma_t < w(t) < \mu_t + m\sigma_t$ per eqns. (7)-(8),

then $u = v^* \Rightarrow w(t+1) \in \mathcal{S}$ for $w_t = w_t^F$ to the level of m-many standard deviations in the GP predictor. Therefore the set $\mathcal{S}$ is forward invariant (i.e. safe).

Proof: Assume $w_t = w_t^F$, so that the measurement is outside the statistical variance allowed of the true value $w_t < w(t) - n\sigma_w$ discussed in Assumption 1. Then,

$$\begin{aligned} w_t - \hat{w}(\boldsymbol{\xi}_t) &< w(t) - n\sigma_w - \hat{w}(\boldsymbol{\xi}_t) \\ \rho_t &< w(t) - n\sigma_w - \mu(\boldsymbol{\xi}_t) + m\sigma_t(\boldsymbol{\xi}) \\ \rho_t - \rho_L &< (w(t) - \mu_t(\boldsymbol{\xi}_t)) + m\sigma_t(\boldsymbol{\xi}) \end{aligned}$$

Consider when the true actuator state is at the lower border of the GP's statistical prediction, $w(t) = \mu_t - m\sigma_t$. Then, $\rho_t - \rho_L < m\sigma_t - m\sigma_t$ so $\rho_t - \rho_L < 0$, and consequently, $v^* = \bar{u}$, giving $w(t) \in \mathcal{S} \Rightarrow w(t+1) = f(w(t), v^*) \in \mathcal{S}$. ■

Remark 1: Although assumptions 1-2 are mild, assumptions 3-4 are strong and may not hold in practice. Smart artificial muscles often have many latent states [12], [32] not sensed. The primary use of Theorem 1 is in our controller tuning procedure for *arbitrarily low-quality sensing*. We chose to assess the safety margin experimentally.

E. Safety Under Outliers

Though proposed method can be tuned for conservativeness by increasing m in the fault guard, outlier measurements in sensors (common in smart materials) give false positives using the controller in eqn. (4). Instead, consider a detection window of N_D-many steps before the guard is activated:

$$v^*(\mathbf{x}, \boldsymbol{\xi}_{t-N_D:t}) = \begin{cases} \bar{u} & \text{if } \rho_t < \rho_L \ \forall t \in \{t - N_D : t\} \\ u^*(\mathbf{x}) & \text{else.} \end{cases} \quad (10)$$

This controller also meets a formal safety condition.

Theorem 2: Safety Under Outliers. Assume that Theorem 1 holds, and in addition:

5) There exists a constraint on the input, $u \le u^{max}$,
6) A backward-reachable set of N_D steps using $u = u^{max}$ is calculated as $\mathcal{S}_D = \{w | w < w_{N_D}^{max}\}$, where $w_{N_D}^{max} = f^{-N_D}(w_{max}, u_{max})$, using f from eqn. (1).

Then if w_t is in $\mathcal{S}_D$, applying $u = v^*$ in eqn. (10) $\Rightarrow w(t+k) \in \mathcal{S}\ \forall k \in \mathbb{R}^+$.

Proof: Assumption 2 gives monotone dynamics, so f^{-1} exists. Calculate $w_{N_D}^{max}$ as above. Consider t such that $w_t = w_t^F$, and the worst-case when $u_{t:t+N_D}^* = u^{max}$. If $w_t \in \mathcal{S}_D$, then $w_{t+N_D} < w^{max} \in \mathcal{S}$ by monotonicity. But then, $\rho_t \ldots \rho_{t+N_D} < \rho_L$, and by eqn. (10), $v_{t+N_D}^* = \bar{u}$. Theorem 1 ensures invariance of $\mathcal{S}$ inductively for $t > t+N_D$. ■

Remark 2: Informally, Theorem 2 states that as long as the safe controller keeps the system at least N_D-many steps of $u = u^{max}$ away from unsafety, then we may operate under a detection window of N_D. Future work will make this approach less tautological by analyzing outlier statistics.

IV. Application to Electrothermal Actuators

Our motivation for fault detection is a soft limb powered by shape memory alloy wires, which contract via temperature change from electrical heating. The temperature sensors for these wires (thermocouples) are fragile and commonly fracture under human contact (Fig. 1). However, the electrical connection is robust, and since the induced stress is well-modeled by the electrical resistance in the muscle [33], [34], resistance serves as a redundant sensor to temperature [18].

A. Robot Architecture and Hardware Framework

To obtain temperature and resistance measurements, we use the sensing and control framework from our prior work [2], [18]. The setup (Fig. 1-2) contains a 10cm-long soft limb cast from silicone rubber, with $M = 1$ active SMA coil (Dynalloy Flexinol 90° C, 0.020") along one edge. It is connected to a programmable power supply that applies a voltage, $u = V \in \mathbb{R}^+$, and measures electrical current I for resistance calculations as $R_t = V_t/I_t$. A computer vision system using AprilTags [35] captures fiducials from which one constant-curvature [6] bending angle is measured, $\theta(t)$. A type-K thermocouple is adhered to the rear of the SMA coil with thermally-conducting epoxy. Therefore, actuator state in one muscle is temperature, $w = T$, and the overall robot state includes the bending angle as $\mathbf{x} = [\theta, \dot{\theta}, w]^\top$.

B. Redundant Sensing Featurization

Applying Theorem 1 requires choosing the input vector $\boldsymbol{\xi}$. We propose a featurization motivated by hysteresis in SMAs [12], [36] and the availability of a nominal thermal model, $w_{t+1} = f(w_t, u_t)$. For hysteresis, we store a sliding window of N_r-many resistance measurements $\mathbf{R} = [R(t - N_r), \ldots, R(t)]^\top$, normalize as $\bar{\mathbf{R}} = (1/\sum \mathbf{R})\mathbf{R}$, and take its 25%, 50%, and 75% quartiles as $\bar{R}_{Q1}, \bar{R}_{Q2}, \bar{R}_{Q3}$. This choice of window and quartiles was motivated by observations within our datasets and prior work [18]. For other applications, domain knowledge must be used to adapt this featurization as needed, or first-principles models such as Prandtl-Ishlinskii [37] could be employed instead. The nominal model for Joule heating of a wire approximated as linear, $\hat{T}_{t+1}(T_t, u_t) = a_1 T_t + a_2 u_t + a_3$ where $a_{1\ldots3}$ are constants calibrated from the dataset Ξ. Prior work has shown good agreement with this model when temperature sensors are not faulty [2], [32]. Consequently,

$$\boldsymbol{\xi}_t = \begin{bmatrix} R_t & \bar{R}_{Q1,t} & \bar{R}_{Q2,t} & \bar{R}_{Q3,t} & \hat{T}_t(\mu_{t-1}, u_{t-1}) \end{bmatrix}^\top \in \mathbb{R}^5 \tag{11}$$

where the mean of the GP is $w = T$ and so represents an open-loop rollout from a non-faulty T_0. This differs from residual learning [38], as we predict w_t rather than errors $e = w_{t+1} - w_t$, allowing for hysteresis-dependence.

C. Nominal Controller and Fault-Free Safe Controller

We adopt our prior work for the no-fault safety-verified controller, $u^*(\mathbf{x})$, needed in eqn. (4). Any nominal controller for the robot is $u_{nom}(\mathbf{x})$, for example, we use PID feedback on the pose of the robot $\theta - \bar{\theta}(t)$ per [2]. We dynamically saturate the control input as $u^* = \min(u_{nom}(\mathbf{x}_t), \gamma u_{sat}(\mathbf{x}_t))$, where $\gamma \in (0, 1]$ is a conservativeness tuning parameter. The saturating limit inverts the nominal model as $u_{sat}(\mathbf{x}_t) = (1/a_2)(T^{max}_{adj} - a_1 T_t - a_3)$ with the adjusted setpoint of $T^{max}_{adj} = (1/\gamma - a_1((1-\gamma)/\gamma))T^{max} - a_3((1-\gamma)/\gamma)$ giving stability around the equilibrium of T^{max}, c.f. Sec. III-A, $w^{max} = T^{max}$. Prior work proves invariance of the set $\mathcal{S} = \{T | T < T^{max}\}$ under the action of this u^* assuming noise-free measurements of actuator temperature.

In this soft robot prototype, inaccuracies in the redundant sensor are much more significant than nominal sensor noise: $\sigma_w << \sigma_t$. So, we take $n\sigma_w = 0 \Rightarrow \rho_L = 0$ below.

V. Experimental Results

We validate our approach with both simulated and real-world failures of temperature measurement on an SMA actuator. We consider the following events during tests:

Definition 1: Active Fault, when $\dot{\rho}_t < 0$, i.e., the fault time occurs as the robot is becoming less safe,

Definition 2: Pre-fault, when $\dot{\rho}_t > 0$, i.e., $v^* = u^*$ as the robot is already retreating from the safe boundary.

In our application, active faults occur as the muscle is heating actively whereas pre-faults occur during cooling or non-operation below the safety boundary ($u = 0$).

A. Model Training and Validation

We use the dataset Ξ from [18], containing 10k sec. of motions of a prototype with one SMA. Training and testing of the GP, eqn. (7)-(8), was performed with a 70-30 split.

We validate our model on three new prototypes, which necessarily have manufacturing differences, over a total of ∼450 sec. of data with both heating and cooling. Predictions of the GP using the older Ξ on the new dataset give a root mean squared error (RMSE) 8.79°C using the GP, closely matching prior work [32]. Fig. 3 shows this high performance as a test under a representative fault guard ($3\sigma_t$).

B. Tuning via Simulated Faults

The proposed fault response controller has multiple parameters to tune, in particular, the m standard deviations for the safety margin, and the N_D detection window per eqn. (10). To do so, we execute a series of simulated faults on the dataset from Sec. V-A and compare against Theorem 2.

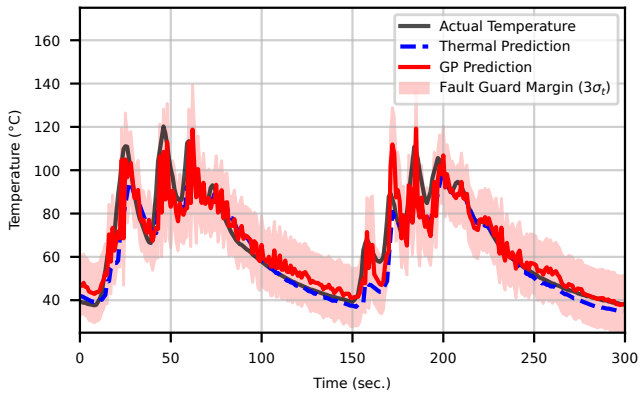

Fig. 3. Our estimator generalizes to new prototypes of this artificial muscle. Model prediction with representative bounds of $\mathbf{3}\boldsymbol{\sigma_t}$, red, and thermal model, blue, are trained only on data from [18].

Since the most conservative scenario of thermal change for SMA is a pure dissipation model, we take $u = 0$ as the simulated fault. If the system can detect failure under this condition, it is expected to alert faster in actual experiments. So, we generate 100 rollouts where t_F is uniformly sampled from $(0, t_{max})$ and set $u_{t_F:t_{max}} = 0$. We then iterate over each rollout with increasing m (safety margin), and for each m, also iterate with increasing window size N_D until 100% precision is obtained (no false alerts). This produces the minimum number of consecutive steps needed to trip faults for a given conservativeness level. We record the time-to-detection $(t - t_F)$, and plot the averages in Fig. 4, color-coding according to N_D. Intuitively, a tradeoff exists between delay in detection (color) and safety level. We selecte the lowest average detection time as a compromise: $m = 5.5$, $N_D = 2$. We ensure not just 100% precision, but observationally, also 100% recall: all simulated faults are detected.

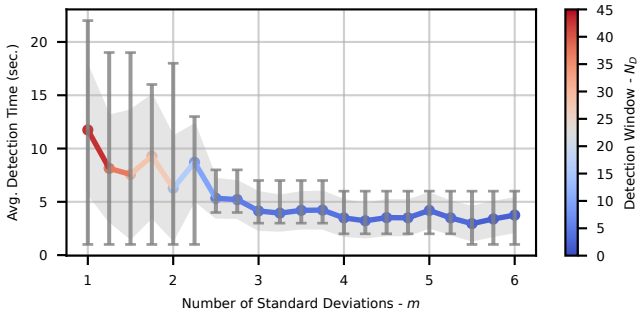

Fig. 4. Safety margin ($m\sigma_t$) vs. detection time delay in simulated fault tests, choosing the rolling detection window N_D to ensure 100% precision. To tune eqn. (10), take the minimum detection delay. Minimum/maximum detection time delay and standard deviations are shown in grey bars and shaded areas correspondingly.

C. Hardware Experiments

Finally, we verify the system's fault detection and response performance in real-time tests of five prototype muscles with thermocouple affixed (see Supplementary Video). In each experiment, the nominal controller tracks a predefined trajectory of bending angles $\theta(t)$ which includes infeasible poses: some θ_t which would require muscle temperatures above T^{max}. We apply the safety-verified controller from [2], which prevents overheating so long as no faults occur.

During each test, we manually fracture the thermocouple from the SMA muscle (Fig. 1). We test both active faults ($u > 0$, heating) as well as pre-faults ($u = 0$, cooling). An example with the fault detection but no fault response ($v^* = u^*$) is shown in Fig. 5, illustrating how the muscle will continue to heat when the thermocouple itself is detached and starts to cool down (red v.s. black line). The other four experiments used the fault-aware controller in eqn. (10), and fractured the thermocouple at $\sim 100°$ C. Theorem 2 held in all cases: the muscles remained below $T^{max} = 130°$ C.

D. Comparison Against Baseline

To evaluate the performance improvement of our method, we implement a comparison baseline that replaces eqn. (9) with a linear fit for $\hat{w}(\boldsymbol{\xi})$ rather than the GP:

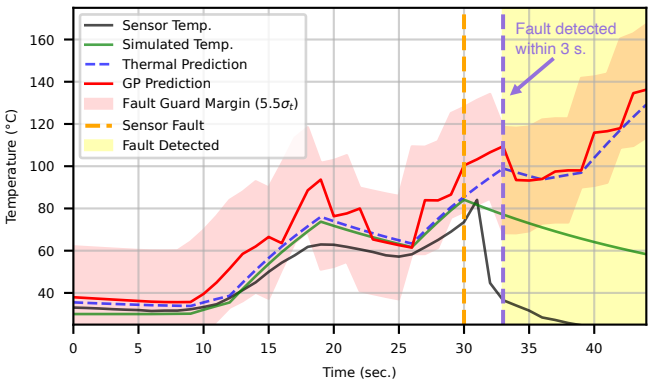

Fig. 5. Real-time plot from an SMA temperature sensing failure test illustrating the GP model's fault detection. The orange dashed line marks sensor failure (at $t = 30$s), while the light yellow region indicates the detected interval (begins at $t = 33$s), reflecting a 3s detection delay.

$$\hat{w}_B(\boldsymbol{\xi}) = \sum_{i=1}^{P} c_i \xi_i + c_0, \quad (w - \hat{w}_B) < \rho_L \Rightarrow u = \bar{u} \tag{12}$$

We fit the constants c using the Ξ dataset. The detection threshold was manually optimized for detection time versus number of false positives: $\rho_L = 55°$ C.

Table I compares the detection time (Δt) and timesteps with a false positive (FP) between our method and this baseline. Our detector successfully and promptly captures sensor failures under both active and pre-faults, outperforming the baseline model in terms of response time and the number of false positive prediction timesteps for most active fault experiments. For pre-fault time, our method shows more consistent detection delays than the baseline, and achieves zero false positive predictions. We expect that pre-faults take significantly longer to detect (compare SMA 1 vs. 2), as detections only occur when the muscle is heating, and arbitrary time elapses between heating cycles. The detector trips a fault with the same delay in repeated pre-fault tests of SMA 2.

TABLE I
FAULT DETECTOR RESPONSE TIME AND FALSE POSITIVE RESULTS

SMA ID	Active Fault				Pre-fault			
	Res. steps (Δt)		FP steps		Res. steps (Δt)		FP steps	
	Ours	linear	Ours	linear	Ours	linear	Ours	linear
1	3 ↓	−−	1 ↑	0	6 ↑	5	0 ↓	1
2	9 ↓	12	0 ↓	1	8 ↓	19	0 –	0
3	4 ↓	11	0 ↓	1	9 ↓	−−	0 ↓	1
4	4 –	4	3 ↓	7	8 ↑	2	0 ↓	2
5	3 ↓	15	0 ↓	1	9 –	9	0 ↓	1
Avg.	4.6	10.5	0.8	2.0	8.0	8.75	0	1.0

Arrows indicate when our method outperforms the baseline: (↓) denotes a lower value (better). '−−' indicates no detection. Faults coded as: ⇒ $u > 0$, ⇒ $u = 0$.

VI. CONCLUSION

This work presents, to the authors' knowledge, the first generalizable approach to detecting and responding to sensor faults in soft artificial muscles that formally maintains safety conditions under feedback control. Using redundant self-sensing, a learned fault detector was able to successfully respond with both 100% precision and 100% recall on a simulated experiment, and met the safety-critical state criterion in five hardware experiments on an electrothermal artificial

muscle. This approach offers a formal engineering tool to enhance the trustworthiness of soft robots working alongside humans, extending the informal claim of 'safety' in soft robots into a control synthesis method.

Future work will address the limitations in the proposed method. The strong assumptions underlying Theorem 1 and Theorem 2 could be relaxed by alternative formulations of ρ and $\hat{w}$ that are tailored to specific smart materials, or are amenable to statistical testing beforehand, or use alternative learning methods. Sources of error could be addressed to improve Assumption 3 regarding $\boldsymbol{\xi}$ and w, in particular, an increase in sensing frequency and noise reduction.

The method here may also be extended to more complex settings, such as coupled actuator-robot models. Non-catastrophic faults that arise from degradation could be addressed by incorporating adaptation, for example, online learning into the GP while pruning the original dataset for outliers [31]. Similarly, intermittent faults could be addressed by actively adapting the detection window, i.e., recalculating $m\sigma_t$ by continually performing the Fig. 4 tuning process online. In addition, recovery strategies could be improved to have more intelligent maneuvers than simple shutdown.

REFERENCES

[1] H. Abidi and M. Cianchetti, "On Intrinsic Safety of Soft Robots," *Front. Robot. AI*, vol. 4, 2017.

[2] A. P. Sabelhaus, Z. J. Patterson, A. T. Wertz, and C. Majidi, "Safe Supervisory Control of Soft Robot Actuators," *Soft Robot.*, vol. 11, no. 4, pp. 561–572, Aug. 2024.

[3] K. Wong, M. Stölzle, W. Xiao, C. D. Santina, D. Rus, and G. Zardini, "Contact-Aware Safety in Soft Robots Using High-Order Control Barrier and Lyapunov Functions," May 2025, arXiv:2505.03841.

[4] A. K. Dickson, J. C. P. Garcia, M. L. Anderson, R. Jing, S. Alizadeh-Shabdiz, A. X. Wang, C. DeLorey, Z. J. Patterson, and A. P. Sabelhaus, "Safe Autonomous Environmental Contact for Soft Robots Using Control Barrier Functions," *IEEE Robot. Automat. Lett. (Early Access)*, pp. 1–8, 2025.

[5] Z. J. Patterson, W. Xiao, E. Sologuren, and D. Rus, "Safe Control for Soft-Rigid Robots with Self-Contact Using Control Barrier Functions," in *Proc. IEEE Int. Conf. Soft Robot.*, Apr. 2024, pp. 151–156.

[6] C. Della Santina, C. Duriez, and D. Rus, "Model-Based Control of Soft Robots: A Survey of the State of the Art and Open Challenges," *IEEE Control Syst. Mag.*, vol. 43, no. 3, pp. 30–65, Jun. 2023.

[7] C. Della Santina, M. Bianchi, G. Grioli, F. Angelini, M. Catalano, M. Garabini, and A. Bicchi, "Controlling Soft Robots: Balancing Feedback and Feedforward Elements," *IEEE Robot. Automat. Mag.*, vol. 24, no. 3, pp. 75–83, Sep. 2017.

[8] M. L. Anderson, R. Jing, J. C. Pacheco Garcia, I. Yang, S. Alizadeh-Shabdiz, C. DeLorey, and A. P. Sabelhaus, "Maximizing Consistent High-Force Output for Shape Memory Alloy Artificial Muscles in Soft Robots," in *Proc. IEEE Int. Conf. Soft Robot.*, Apr. 2024, pp. 213–219.

[9] S. Terryn, J. Brancart, D. Lefeber, G. Van Assche, and B. Vanderborght, "Self-healing soft pneumatic robots," *Sci. Robot.*, vol. 2, no. 9, p. eaan4268, Aug. 2017.

[10] D. K. Soother, J. Daudpoto, and B. S. Chowdhry, "Challenges for practical applications of shape memory alloy actuators," *Mater. Res. Express*, vol. 7, no. 7, p. 073001, Jul. 2020.

[11] S. Zaidi, M. Maselli, C. Laschi, and M. Cianchetti, "Actuation Technologies for Soft Robot Grippers and Manipulators: A Review," *Curr. Robot. Rep.*, vol. 2, no. 3, pp. 355–369, Sep. 2021.

[12] A. P. Sabelhaus, R. K. Mehta, A. T. Wertz, and C. Majidi, "In-Situ Sensing and Dynamics Predictions for Electrothermally-Actuated Soft Robot Limbs," *Front. Robot. AI*, vol. 9, 2022.

[13] M. Pontin, S. Miyashita, and D. D. Damian, "Development and Characterization of a Soft Valve for Automatic Fault Isolation in Inflatable Soft Robots," in *Proc. IEEE Int. Conf. Soft Robot.*, Apr. 2022, pp. 62–67.

[14] L. Balasubramanian, T. Wray, and D. D. Damian, "Fault Tolerant Control in Shape-Changing Internal Robots," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2020, pp. 5502–5508.

[15] H. Gu, H. Wang, F. Xu, Z. Liu, and W. Chen, "Active Fault Detection of Soft Manipulator in Visual Servoing," *IEEE Trans. Ind. Electron.*, vol. 68, no. 10, pp. 9778–9788, Oct. 2021.

[16] X. Li, G. Jin, and M. Deng, "SMA Actuator-Based Nonlinear Fault Tolerant Control for A Flexible Arm with Active Compensating Unit," in *Proc. Annu. Conf. Soc. Instrum. Control Eng.*, Sep. 2023, pp. 1313–1318.

[17] S. Rabiei, S. S. Nalkenani, I. Sharifi, and H. A. Talebi, "Fault tolerant position control of soft bending actuator in the presence of actuator leakage," *Robot. Auton. Syst.*, vol. 188, no. C, Apr. 2025.

[18] R. Jing, M. L. Anderson, J. C. P. Garcia, and A. P. Sabelhaus, "Self-Sensing for Proprioception and Contact Detection in Soft Robots Using Shape Memory Alloy Artificial Muscles," Sep. 2024, arXiv:2409.17111.

[19] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3.

[20] M. M. Seron, X. W. Zhuo, J. A. De Doná, and J. J. Martínez, "Multisensor switching control strategy with fault tolerance guarantees," *Automatica*, vol. 44, no. 1, pp. 88–97, Jan. 2008.

[21] F. Stoican, S. Olaru, and G. Bitsoris, "Controlled invariance-based fault detection for multisensory control systems," *IET Control Theory Appl.*, vol. 7, no. 4, pp. 606–611, 2013.

[22] A. Rahman Kodakkadan, M. Pourasghar, V. Puig, S. Olaru, C. Ocampo-Martinez, and V. Reppa, "Observer-based Sensor Fault Detectability: About Robust Positive Invariance Approach and Residual Sensitivity," *IFAC-PapersOnLine*, vol. 50, no. 1, Jul. 2017.

[23] S. R. Saufi, Z. A. B. Ahmad, M. S. Leong, and M. H. Lim, "Challenges and Opportunities of Deep Learning Models for Machinery Fault Detection and Diagnosis: A Review," *IEEE Access*, vol. 7, 2019.

[24] K. Garg, C. Dawson, K. Xu, M. Ornik, and C. Fan, "Model-Free Neural Fault Detection and Isolation for Safe Control," *IEEE Control Syst Lett.*, vol. 7, pp. 3169–3174, 2023.

[25] H. Zhang, L. Niu, A. Clark, and R. Poovendran, "Fault Tolerant Neural Control Barrier Functions for Robotic Systems under Sensor Faults and Attacks," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2024, pp. 9901–9907.

[26] H. Zhang, Z. Li, and A. Clark, "Safe Control for Nonlinear Systems Under Faults and Attacks Via Control Barrier Functions," *IEEE Trans. Autom. Control*, pp. 1–16, 2025.

[27] A. Zhang, T.-H. Wang, R. L. Truby, L. Chin, and D. Rus, "Machine Learning Best Practices for Soft Robot Proprioception," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2023, pp. 2564–2571.

[28] J. Chen, W. Hu, D. Cao, M. Zhang, Q. Huang, Z. Chen, and F. Blaabjerg, "Gaussian Process Kernel Transfer Enabled Method for Electric Machines Intelligent Faults Detection With Limited Samples," *IEEE Trans. on Energy Conv.*, vol. 36, no. 4, Dec. 2021.

[29] M. Stölzle and C. D. Santina, "Piston-Driven Pneumatically-Actuated Soft Robots: Modeling and Backstepping Control," *IEEE Control Syst Lett.*, vol. 6, pp. 1837–1842, 2022.

[30] D. Nguyen-Tuong and J. Peters, "Model learning for robot control: a survey," *Cogn. Process.*, vol. 12, no. 4, pp. 319–340, 2011.

[31] Z. Q. Tang, H. L. Heung, K. Y. Tong, and Z. Li, "A Probabilistic Model-Based Online Learning Optimal Control Algorithm for Soft Pneumatic Actuators," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1437–1444, Apr. 2020.

[32] A. P. Sabelhaus and C. Majidi, "Gaussian Process Dynamics Models for Soft Robots with Shape Memory Actuators," in *Proc. IEEE Int. Conf. Soft Robot.*, Apr. 2021, pp. 191–198.

[33] C.-C. Lan and C.-H. Fan, "An accurate self-sensing method for the control of shape memory alloy actuated flexures," *Sens. Actuators A: Phys.*, vol. 163, no. 1, pp. 323–332, Sep. 2010.

[34] H. Kim, Y. Han, D.-y. Lee, J.-I. Ha, and K.-J. Cho, "Sensorless displacement estimation of a shape memory alloy coil spring actuator using inductance," *Smart Mater. Struct.*, Dec. 2012.

[35] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2011, pp. 3400–3407.

[36] D. Hughes and J. T. Wen, "Preisach modeling of piezoceramic and shape memory alloy hysteresis," *Smart Mater. Struct.*, vol. 6, no. 3, pp. 287–300, Jun. 1997.

[37] M. Al Janaideh, J. Mao, S. Rakheja, W. Xie, and C.-Y. Su, "Generalized Prandtl-Ishlinskii hysteresis model: Hysteresis modeling and its inverse for compensation in smart actuators," in *Proc. IEEE Conf. Decis. Control*, Dec. 2008, pp. 5182–5187.

[38] L. Hewing, J. Kabzan, and M. N. Zeilinger, "Cautious Model Predictive Control Using Gaussian Process Regression," *IEEE Trans. Autom. Control*, pp. 1–8, 2019.