Intrinsic-Motivated Sensor Management: Exploring with Physical Surprise

Jingyi Yuan  
jyuan46@asu.edu  
Arizona State University  
Tempe, Arizona, USA

Yang Weng  
Yang.Weng@asu.edu  
Arizona State University  
Tempe, Arizona, USA

Erik Blasch  
erik.blasch@gmail.com  
Air Force Research Lab  
Arlington, Virginia, USA

ABSTRACT

In modern complex physical systems, advanced sensing technologies extend the sensor coverage but also increase the difficulties of improving system monitoring capabilities based on real-time data availability. Traditional model-based methods of sensor management are limited to specific systems/settings, which can be challenged when system knowledge is intractable. Fortunately, the large amount of data collected in real-time allows machine learning methods to be a complement. Especially, reinforcement learning-based control is recognized for its capability to dynamically interact with systems. However, the direct implementation of learning methods easily overfits and results in inaccurate physics modeling for sensor management. Although physical regularization is a popular direction to bridge the gap, learning-based sensor control still suffers from convergence failure under highly complex and uncertain scenarios. This paper develops physics-embedded and self-supervised reinforcement learning for sensor management using an intrinsic reward. Specifically, the intrinsic-motivated sensor management (IMSM) constructs the local surprise information from the physical latent features, which captures hidden states in observations, and thus intrinsically motivates the agent to speed-up exploration. We show that the designs can not only relieve the lack of consistency with underlying physics/physical dynamics, but also adapt the global objective of maximizing monitoring capabilities to local environment changes. We demonstrate its effectiveness by experiments on physical system sensor control. The proposed model is implemented for the sensor management of unmanned vehicles and sensor rescheduling in complex/settled power systems, with or without observability constraints. Numerical results show that our model provides consistently higher threat detection accuracy and better observability recovery, as compared to existing methods.

CCS CONCEPTS

- Computer systems organization → Sensor networks;  
- Theory of computation → Sequential decision making;  
- Hardware → Smart grid.

KEYWORDS

Sensor management, learning-based control, neural networks, reinforcement learning, intrinsic motivation, unmanned vehicles, power systems

ACM Reference Format:


1 INTRODUCTION

As sensing technologies develop, Internet of everything (IoE) significantly extends the sensor coverage to the edges of physical systems [26]. However, these system areas are with relatively limited and compromised unequal sensors due to low-cost, low-power sensor technologies, calling for methods to optimize monitoring capabilities based on data availability and traceability of physical laws for maximum information gain [29, 30, 43]. To boost the relevancy of data sources, sensor reconfiguration is the basic step to maximize monitoring performance or sustain cyber attacks in operations [21, 25, 47]. For example, [38] formulates a linear programming model for reconfiguring the communication channel of phasor measurement units (PMUs) in power grids. Unlike the deterministic model, [8] proposes a stochastic model to capture uncertainties in the environment for unmanned vehicle path-planning. For well-understood physical or engineering-based systems of moderate scale, detailed physical models or/and linearized dynamics are derived for control. However, in real-world applications, the detailed system models may be incomplete or intractable to construct, and linear models may poorly represent system dynamics [7, 14]. For example, when unmanned vehicles are collecting information with mobile sensors in an unknown field, there can be potential threats that require exact quantification for monitoring [46]. To solve it, previous works assume a prior system basis, which may be unavailable [8, 38].

Since measurements are collected in real-time, emerging model-free methods can discover system dynamics and update reconfiguration policies by leveraging the information from data [17, 23]. In particular, reinforcement learning (RL) has shown good performance in many domains due to its capability to characterize the control rule based on the dynamic interaction with the systems [31, 33]. When applying RL to sensor management of physical systems, however, the performance may deteriorate due to: 1) lack of model consistency with the underlying physics/physical dynamics, and 2) slow convergence with sparse reward signals. Specifically,
the former limitation becomes a major concern when deep neural networks are used in RL for (dynamics) model approximation. Though the numerical accuracy may be high for in-sample range, the black-box approximation is far from physics for extrapolation. Moreover, due to the same issue of incomplete system information, the policy learning process may also suffer from environment unobservability, leading to propagated bias in exploration. Although the embedding of domain knowledge can regularize the learning model to avoid overfitting, the sparse incentive is unavoidable and could result in non-convergence and the stop of policy update. This is because, RL algorithms update policy with reward signals from environment; but, the application of the sensor management problem usually provides a reward signal at the end. For example, an agent in the energy management case will only receive a reward (environmental incentive) if it succeeds in completely maintaining the safety operation during a long mission. Similarly, the agent for sensor management can be rewarded only when the receiver obtains all necessary feedback for localizing threats. During the long mission of an unmanned vehicle, i.e., unmanned aerial vehicle (UAV) and unmanned surface vehicle (USV), and when the mobile sensors are waiting for feedback, the control agents have almost zero incentive to explore and become smarter. Thus, the global objective of sensor control may not be adaptive to local environment changes for continuous exploration [20, 36].

To resolve these issues, we improve sensor reconfiguration with self-supervision over exploration. The idea comes from a psychology concept where human behaviors can be driven by intrinsic motivation [6]. It explains how humans can acquire knowledge actively without any obvious external rewards. For reinforcement learning, such an intrinsic motivation is formed as the internal reward signal of surprise information. It measures the exploration experience where the agent sees the ground-truth states that are different from its belief about the environment dynamics. Although the integration of such surprise information push the agent to move forward, previous works prefer approximated models to compute the dynamics. It shows good capability in dimension reduction and feature extraction for high-dimension inputs (i.e., images, video frames), but the lack of physical consistency is insufficient to guarantee a same good performance in sensor management. Therefore, we propose to construct the physical surprise measure via physics embedding in latent features, and call the design intrinsic-motivated sensor management (IMSM). By motivating an agent with physical surprise, we aim to balance between reaching the specific but sparse objective via model approximation and persisting in local improvement of predictability based on prior physics.

It is essential to note that there are various use cases to implement the proposed method, e.g., unmanned vehicle path-planning, threat detection, and power system observability restoration, etc [11, 38, 46]. For example, sensor management for submarine tracking is an important application in Anti-Submarine Warfare (ASW) to guarantee observability [22], which is required to quickly identify and accurately localize as many enemy submarines as possible via active sonar systems [12, 42]. Thus, we conduct experiments on different physical systems to demonstrate the ability of the proposed method to reconfigure sensors for maximum observability under realistic scenarios.

2 RELATED WORK

2.1 Sensor Control/Management

As for physical systems with moderate sensor coverage, an internal adjustment of information sources can significantly improve the relevance of information for monitoring [18]. Determining the optimal sensor configuration is essential for both the systems with mobile sensors and systems with location-fixed sensors. Typical cases of applications include the path planning of unmanned vehicles and the restoration of observability in power systems [8, 38]. To solve such a problem of resource management, previous works usually assume sufficient prior knowledge of system basis, which may be unavailable in real-world applications. Moreover, some methods have a quasi-static setup [13], which is difficult to accommodate uncertainties in environment during a period of exploration. To consider uncertainties in sensor control, past methods utilized Kalman filter type of methods because of their capability to obtain analytical solutions for real-time control. However, such a Markov process has many unrealistic assumptions of the environment dynamics, e.g., linear transition models, Gaussian distributed noises, and nearly constant velocity (NCV) model of target motion [5, 15, 32]. In particular, the vehicles (e.g., UAVs and USVs) usually have to operate in undetermined and unpredictable environments (e.g., UAVs in rain). Other methods seek to use the contextual information for adaptable sensor management [40]. The bias of environment dynamics makes a prompt policy update difficult, so we implement reinforcement learning algorithms to learn the policy for sensor management by maximizing the rewards signals.

2.2 Intrinsically Motivated Reinforcement Learning

2.2.1 Intrinsic Reward. Conventional reinforcement learning agents are motivated by a reward signal from the environment, which can be quite sparse in many real-world applications. For sensor management examples, the reward signal is only available when reaching the final operational objective. In contrast to the extrinsic rewards, psychologists define intrinsic motivation as being moved to do something for one’s own sake without external stimulus. An analogy to this concept, intrinsic reward is used to actively drive the agents to acquire knowledge from the environment and update the policy, where the learning is proceeded in a self-supervised manner. It augments the reward signal for local exploration guidance during training steps [6]. Specifically, the intrinsic motivation is embedded as measuring the predictability of the consequences of actions; and the goal is to improve the predictability [34, 35]. There are many metrics available to determine such prediction error, while cognitive experimental evidence shows that Bayesian surprise perfectly characterizes the extent where human gaze towards unexpected items in dynamic natural scenes [20]. The core design of encouraging exploration by Bayesian surprise is to leverage the information gain to measure and reduce the uncertainty on the environment dynamics [20, 39].

2.2.2 State Representation for Dynamics Model. Since the learning states of many tasks are indirectly observable, many approaches use different tools for state representation to embed intrinsic rewards. [36] uses an inverse dynamics model to encode features
controlled by the agent, and [37] instead embed the disagreement of an ensemble of forward dynamics models to stimulate the agent. Moreover, the auto-encoder is used in [41] to encode the states for dimension reduction and feature extraction. [16] builds the world model to capture the compact spatial-temporal representation of environment dynamics and train the agent inside the simulated latent space. Especially, it shows better performance to compute predictions with fixed feature representation [4]; however, the pre-defined features may not be consistent with the underlying physical system model and they are limited to uncertainties.

3 PROBLEM STATEMENT

Generally, solving the sensor management problem, one needs to control the degrees of freedom of sensor network in order to reach the final target of physical system operation. Therefore, sequential decisions need to be made based on the available measurements. Namely, we aim to find a policy to reconfigure sensors at each step. In this section, we formulate the problem and elaborate with two typical cases.

3.1 Sensor Management as a Partially Observable Markov Decision Process

In physical system operation, the sensor network collects measurement samples at time $t$, which are observations $o_t \in \Omega$. The system observations may implicitly reveal the physical system states $s_t \in S$ for control. Subsequently, the agent chooses the control action of the sensor network, which is defined as $a_t \in A$. The physical system transits to the next state $s_{t+1}$ and $p$ denotes the probability function of state transitions

$$p(s_{t+1}|s_t, a_t).$$

Since $s_t$ is indirectly observable, the environment dynamics is represented as

$$p(s_{t+1}|o_t, a_t).$$

The goal of the agent is to find a policy mapping from state space $S$ to action space $A$. In the setting of partial observability, the agent of sensor reconfiguration will condition the actions based on the observations stochastically, which is to learn the a mapping rule of control policy

$$\pi : \Omega \rightarrow A.$$ (3)

For conventional reinforcement learning, such a mapping rule is optimized by maximizing the reward $R$ from the extrinsic environment incentive $J_e$. For the self-supervision design, we will introduce the intrinsic motivation $J_i$ later in detail. Therefore, the decision-making process for sensor reconfiguration is formulated as a partially observable Markov decision process (POMDP) in the tuple

$$\mathcal{G} = \{S, \mathcal{A}, p, R, O, \gamma\},$$ (4)

where $O$ is a set of conditional observation probabilities and $\gamma \in [0, 1]$ denotes the discount factor. Fig. 1 shows the decision network. The mathematical definition is general and can be adapted to specific examples of sensor management.

3.2 Specification of Applicable Sensor Management Problems in Physical Systems

Example: Sensor Reconfiguration for Observability Restoration in Power System. In power system operation, available observations include noisy measurements of different system quantities like power injections, setpoints of power electronics, and battery state of charge, etc. While the sensor network has growing coverage to support downstream operations, it is vulnerable to cyber-attacks and potential outages. To restore the observability, the sensor network (e.g., phasor measurement units (PMUs)) can be reconfigured by updating the sensor connections, communication channels, and measurement types. The goal of the learning agent is to find a reconfiguration rule that restores the observability of system states as much as possible while making a small number of changes to the sensors.

Example: Mobile Sensor Control for Threat Detection: For sensor control in vehicles, especially unmanned vehicles, the problem aims to plan a path or track a target while exploring an unknown field [3, 45, 46]. Take unmanned aerial vehicle (UAV) as an example, the available observations from the sensor networks include noisy range and/or azimuth measurements. The system states include the target vehicle position and velocity Cartesian coordinate. The learning objective is to determine the management policy of sensor resources to reach the final goal of tracking the target, e.g., coordinate multi-static radar networks to maximize target coverage.

4 PROPOSED METHOD

In reinforcement learning, the extrinsic reward is designed to prompt the achievement of long-term control goals. However, sensor management tasks provide the reward signal only after reaching the final target, which is too sparse for the agent to proceed step-by-step and learn reward-seeking strategies. Therefore, as a complement of the end-target $J_e$, we introduce the intrinsic reward $J_i$ for two purposes: 1) driving the agent to explore the environment and update the policy without external guidance, and 2) serving as a local and dense/frequent regularization of the control policy.

In order to achieve the goals, we need to design an appropriate measure of intrinsic reward in the sensor management problem.

Figure 1: Compare the decision network using the proposed physical surprise (right) with the standard POMDP (left).
Since the learning method is implemented on critical physical systems, the potential requirement is to disturb the system as little as possible while recovering the physical dynamics. To make each exploration step count, we integrate the intrinsic reward as the agent’s predictability of feedback after taking control actions. Namely, it is the surprise of the agent’s prediction, which compares the posterior and prior beliefs of the environment dynamics and system states. There are many metrics available to mathematically determine such prediction error, and [20] use the experimental evidence to show that Bayesian Surprise perfectly characterizes the extent where human gaze towards unexpected items in dynamic natural scenes [20]. To represent the Bayesian Surprise as intrinsic reward, we will first quantify the difference between the agent’s prediction and the ground truth, and then generate latent features to quantify the uncertainty in the environment and observations.

4.1 Bayesian Surprise to Integrate Intrinsic Guidance
To be specific with intrinsic information, the agent derives the prior belief of the next states based on its current observations and actions, resulting in the forecast of the environment dynamics

\[ p(s_{t+1}|o_t, a_t). \]

(5)

Then, after receiving the new observation, the realized environment dynamics or dynamics posterior is denoted as

\[ q(s_{t+1}|o_t, a_t, o_{t+1}). \]

(6)

The surprise information means how different they can be from each other. Since the state transitions are probability distributions in a stochastic process, measuring their difference is to quantify the divergence between the prior and posterior distributions of the agent’s belief on environment dynamics. In information theory, such a measure is perfectly characterized by Kullback-Leibler (KL) divergence \( D_{KL} \) [19, 20]. However, the difficulty lies in how to compute the system dynamics with uncertainty to be consistent with the underlying physics during training.

4.2 Virtual Physical Features to Represent System States
4.2.1 Fixed Feature Representation. In order to compute the forecast of the environment dynamics (5), a feature representation of states is needed because the systems states are not directly observable in POMDP. An intuitive way is to use a fixed feature transformation

\[ \phi_t = f(a_t) \]

(7)

to encode the system states, which improve the performance. However, the deterministic features of raw observations are limited to capturing the uncertainty in the environment simultaneously. Also, for physical systems, using arbitrary nonlinear feature for approximation may fail to capture the true dynamics.

4.2.2 Virtual Physical Feature Representation. Therefore, we intend to regularize the feature transformation with specific domain information. When analyzing complex physical system, e.g., power system, there is a tradition to create a virtual node, aggregating an arbitrary part of nodes for network reduction and compact representation [1, 9]. Motivated by this method, we propose to use virtual states in feature representation for system dynamics model. Although the true system states \( s_t \), which could be some unmeasurable physical quantities, are unobservable, we introduce the virtual state variable \( z_t \). In the system dynamics model, \( z_t \) is updated to capture the state transition with uncertainties. Simultaneously, since the feature transformation could improve state representation, we adopt it to extract information from raw observations but in the form of

\[ \phi_t = g(z_t). \]

(8)

In particular, we customize \( \phi_t \) to reveal physics based on domain knowledge, as shown on the left of Fig. 2. For example, the nodal voltages are usually unmeasured in power systems, but there are governing functions over the measurable power injections and voltages [49]. Thus, the functional types are embedded in \( \phi_t \) to reveal true physics in the model. Therefore, we can model the state transition with respect to the virtual state variable, and denoted the dynamics as

\[ p(z_{t+1}|o_t, a_t). \]

(9)

The formula is to compute the prior beliefs over the next states based on observations and actions. After realizing the control actions in the system, the posterior becomes

\[ q(z_{t+1}|o_t, a_t, o_{t+1}), \]

(10)

which is approximated during learning, i.e., by a parameterized neural network. Accordingly, the decision network is updated as the right-hand side of Fig. 2. The comparison of the prior beliefs and the posterior is shown as the right bottom in Fig. 2.
With the above computation, the intrinsic reward can be formed as the KL divergence between the system dynamics and the posterior approximation:

\[
J_i = D_{KL}[q(z_t+1|a_t, o_t, o_{t+1})||p(z_{t+1}|a_t, o_t)].
\]  

(11)

The formula measures the agent’s predictability using the regularization of virtual physical features on learning system dynamics.

### 4.3 Integrated Rewards for Gradual Exploration in Sensor Management

In general, the optimal policy \( \pi^* \) of control is obtained by jointly maximizing the long-term extrinsic reward \( J_e \) and the intrinsic reward \( J_i \):

\[
\max c_1 \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t J_e[a_t] + c_2 \mathbb{E}_z q(z_t) \{ \log p(\phi_{t+1}|s_{t+1}) \} - D_{KL}[q(z_{t+1}|a_t, o_t, o_{t+1})||p(s_{t+1}|a_t, o_t)] \right]
\]

(12)

where \( \gamma \) is the discount factor of reward, and \( c_1 \) and \( c_2 \) are positive constants to scale and balance the global and local rewards for exploration. Specifically, scalars \( \in [0, 1] \) are used for \( c_1 \) and \( c_2 \) in this paper. They are selected from a group of candidates by cross-validation. Based on the mathematical definition, Fig. 3 illustrates the general idea of how two types of rewards jointly guide the exploration process. With the above designs, the control agent improves the estimation of environmental dynamics as well as regularizes the local information transition for policy improvement of sensor control. Thus, we call the proposed method as intrinsic-motivated sensor management (IMSM).

With the general formula of IMSM, we illustrate the adaption to application examples. For the power system example, the long-term objective \( J_e \) is to maximize the coverage of the observable region, which indicates the recovery of monitoring capability. Mathematically, it is defined as the positive measure of knowledge acquisition, e.g., the count of recovered physical parameters by re-dispatching sensors, and the negative summation of reconfiguration costs, e.g., cost of deploying sensors [38]. In contrast, the intrinsic reward \( J_i \) is defined as the prediction accuracy of how a specific sensor configuration leads to change of power system measurements. By combining both reward functions, the learning agent tends to explore the efficient trigger actions with minimum perturbations, which speeds up the exploration.

For the threat detection example, the long-term goal of \( J_e \) is mathematically defined as the maximized information gain using Rényi information divergence [2, 44]. To speed up the exploration, the intrinsic reward \( J_i \) is defined as the prediction accuracy of how specific sensor control actions lead to the temporal observation coverage of the target. \( J_e \) and \( J_i \) together ensure maximum coverage of both spatial and temporal sensing for target (i.e., threat) detection, for which one example is used to illustrate the mechanism in Fig. 4. An optimal policy should generate a decision process that maximizes an expected reward, e.g., the negative mean square error of tracking or the probability/rate of successful threat detection.

### 5 EXPERIMENTS

#### 5.1 Setups

For experiments, we focus on addressing the problem of sensor management for physical systems under uncertain operation scenarios. The setup of testing cases, experiment results, and discussion are included in the following.

**Simulation Setup.** We conduct experiments on two physical and engineering-based systems for sensor management: the threat detection for mobile vehicles and the observability restoration in power system.

For the first experiment, when unmanned vehicles are collecting information with mobile sensors in an unknown field, there can be potential threats that require exact quantification for monitoring. The environment is customized based on the Grid World by adding the threat field onto the grid [24]. Different types of threats are considered, such as line threat, point threat, and Gaussian Distributed threat, which are generally illustrated as the red line and circle in Fig. 4. We use the threats only to generate the environment and they are assumed unknown to the agent during learning.

For the power system experiment, we select IEEE 123-bus network and a utility feeder [27, 28]. For the IEEE 123-bus network, we design the sensor coverage scenarios to be 100% and 60% of the total nodes. Subsequently, for the large-scale utility distribution grid, it has 2721 nodes in total (2354 passive nodes can be reduced to simplify the model), and the sensor data is collected from the local utility, which is the power injection data (load data, PV data, and injection data from smart meters) recorded every 15 minutes for one year. Since the measurements are incomplete and the scenarios are limited, we prepare the data set by conducting traditional power flow simulations on the given feeders using MATPOWER [50, 51] and OpenDSS [10]. Different contingencies and attacks are considered to simulate the scenarios that require restoration. For the available measurement, slight Gaussian noises

\[ \mathcal{N}(0, 0.02) \]

and 2% outliers (5-10 times larger than regular values) are added as usual in the case of power system.

**Model Implementation.** As previously shown in Fig. 2, embedding the proposed intrinsic reward is to compute the difference between the predicted and realized dynamics, which are represented by parameterized models. In this paper, we adopt neural networks (NNs) to approximate the models and optimize the representation. The basic NN structure is multi-layer perceptron (MLP) with ReLU (Rectified Linear Unit) activation function. For feature representation, both physical features and nonlinear features are used, and the parameters can be optimized during training, according to [48]. For all the NN training in experiments, we use Adam optimizer with a learning rate hyperparameter set

\[
\{0.001, 0.0005, 0.0002, 0.00005\}
\]

and momentum parameters

\[
\beta_1 = 0.5 \\
\beta_2 = 0.999
\]

(13)

(14)

to train 200 epochs for each experiment. The policy value loss coefficient is fixed to 0.5 and the entropy coefficient to 0.001. All the experiments are completed with a computer equipped with Inter(R) Core(TM) i7-9700k CPU and Nvidia Gerforce RTX 2080Ti GPU.

**Baseline Methods.** We compare the proposed methods with two types of baselines: 1) the traditional methods on sensor reconfiguration using predefined system basis or biased dynamics assumption, where [8] is for threat detection case and [38] is for observability restoration case. They are denoted as *Baseline 1* in the table and plots; 2) the previous intrinsic-motivated reinforcement learning method [36], which computes the intrinsic reward as the mean square error (MSE) between the model predictions based on compact feature and the true features. And, the raw observations are mapped to feature space using a nonlinear function. It is combined with the dynamics model to optimize in an inverse-dynamics manner and the model architecture is called intrinsic curiosity module (ICM). We denote it as *Baseline 2: ICM*.

To compare the performance of sensor management, we use the following evaluation metrics.

- The success rate of sensor management, which is indicated by two values: the count of successfully reaching the final target over the total number of trials, and the count of algorithm convergence over the total number of trials. We use both...
Intrinsic-Motivated Sensor Management: Exploring with Physical Surprise

KDD '22, August 14–18, 2022, Washington, DC, USA

Figure 5: Comparing the proposed methods with baselines on sensor reconfiguration for observability restoration in power systems. Different testing systems and sensor coverage scenarios are considered. The rates are calculated based on trails under different initializations.

<table>
<thead>
<tr>
<th>Physical System</th>
<th>Grid Size (No. of Grid Points)</th>
<th>Metrics</th>
<th>Baseline 1 (×100%)</th>
<th>Baseline 2: ICM (×100%)</th>
<th>Proposed Physical Surprise (×100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid World with Threats</td>
<td>20^2 = 400</td>
<td>Convergence Rate</td>
<td>92.3%</td>
<td>95.3%</td>
<td>92.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Target-Reaching Rate</td>
<td>84.8%</td>
<td>80.1%</td>
<td>85.5%</td>
</tr>
<tr>
<td></td>
<td>30^2 = 900</td>
<td>Convergence Rate</td>
<td>85.5%</td>
<td>82.9%</td>
<td>83.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Target-Reaching Rate</td>
<td>70.5%</td>
<td>73.5%</td>
<td>83.0%</td>
</tr>
<tr>
<td></td>
<td>40^2 = 1600</td>
<td>Convergence Rate</td>
<td>84.0%</td>
<td>82.0%</td>
<td>82.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Target-Reaching Rate</td>
<td>70.0%</td>
<td>70.0%</td>
<td>76.0%</td>
</tr>
</tbody>
</table>

Figure 6: Comparing the proposed methods with baselines on sensor reconfiguration for observability restoration in power systems. Different testing systems and sensor coverage scenarios are considered. The rates are calculated based on trails under different initializations.

<table>
<thead>
<tr>
<th>Physical System</th>
<th>Observability</th>
<th>Metrics</th>
<th>Baseline 1 (×100%)</th>
<th>Baseline 2: ICM (×100%)</th>
<th>Proposed Physical Surprise (×100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>123-bus Network</td>
<td>100% Sensor Coverage</td>
<td>Convergence Rate</td>
<td>86.8%</td>
<td>85.3%</td>
<td>91.5%</td>
</tr>
<tr>
<td></td>
<td>Target-Reaching Rate</td>
<td>86.8%</td>
<td>79.6%</td>
<td>88.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60% Sensor Coverage</td>
<td>Convergence Rate</td>
<td>55.3%</td>
<td>82.9%</td>
<td>84.7%</td>
</tr>
<tr>
<td></td>
<td>Target-Reaching Rate</td>
<td>55.3%</td>
<td>73.4%</td>
<td>80.5%</td>
<td></td>
</tr>
<tr>
<td>Utility Feeder</td>
<td>N/A</td>
<td>Convergence Rate</td>
<td>70.0%</td>
<td>76.5%</td>
<td>74.0%</td>
</tr>
<tr>
<td></td>
<td>Target-Reaching Rate</td>
<td>70.0%</td>
<td>65.5%</td>
<td>72.5%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 7: The average rewards (y-axis) of the proposed method (Physical Surprise as Intrinsic Reward) compared with RL-based sensor management without intrinsic guidance (W/O Intrinsic Reward) and the previous intrinsic-motivated RL model (ICM). The shaded areas represent the standard deviations due to different initialization for each experiment. The four testing cases are: (a) observability restoration of 123-bus power system network with 60% sensor coverage, (b) observability restoration of the power system of Utility feeder, (c) threat detection of grid with the size 30^2 = 900, and (d) threat detection of grid with the size 40^2 = 1600.
5.2 Numerical Results and Comparison with Baselines

This section demonstrates the performances of sensor reconfiguration by experiments on both mobile sensor networks (i.e., in UAVs) and power systems. For threat detection, Fig. 5 presents the success rate for three cases, for which the grid size is increased to enlarge exploration space. Note that Baseline 1 requires a known system basis while the others do not. The comparison shows that the proposed IMSM can reach a similar convergence rate as Baseline 1 even without the basis. While the intrinsic reward in ICM also improves the convergence rate, embedding the physical surprise shows a higher rate of successfully detecting the threats among independent trials. In particular, we compare the exploration using intrinsic reward with and without physical feature representation in Fig. 8. The visualization is the experimental realization of the diagram in Fig. 4, where lighter colors than the background can distinguish the Gaussian and line threats. The closer the threat center (the most delicate color), the more accurate threat detection. For the observability restoration, we increase the difficulty by changing the sensor coverage of power system, which also simulates the real-world condition. When the system is fully observable with 100% sensor coverage, all the methods can reach high success rate to restore observation after attacks. This is still a POMDP since not all the physical quantities are measured. However, when the sensor coverage is limited to 60% of the whole system, the performance of Baseline 1 deteriorates significantly as the incomplete system basis is insufficient to model dynamics. The ICM model shows a stable performance due to the DNN approximation in state representation. Moreover, our IMSM outperforms ICM, which should owe to embedding the physical features instead of random features.

Besides, to analyze the effects of different incentives for sensor management, Fig. 7 plot the average reward values during exploration for four representative cases. The blue curve plots indicate the extrinsic reward only, which serve as the reference for intrinsic motivation. It can be observed that, both agents of ICM and the proposed IMSM receive positive guidance to explore the environment and update policy. ICM converges faster than IMSM for most of the time. But, the performance is unstable as shown in Fig. 5 and Fig. 6. The baseline ICM tends to result in a local optimal solution more often than using physical surprise as intrinsic reward.

6 CONCLUSION

In this paper, we study the problem of sensor management in complex physical systems, even with limited sensor coverage. We show that both the biased system assumption and the end-target objective would limit the capability to efficiently manage sensing resources for monitoring and operation. To resolve the problem, we develop physics-embedded and self-supervised reinforcement learning for sensor management using an intrinsic reward. Specifically, we construct the local surprise information from the physical latent features, which captures hidden states in observations, and thus intrinsically motivate the agent to speed-up exploration. We show that the designs can not only relieve the lack of consistency with underlying physical dynamics but also adapt the global objective of maximizing monitoring capabilities to local environment changes. The proposed self-supervised learning method can be adapted to different sensor management tasks in physical systems.

REFERENCES
