Meta-Active Learning in Probabilistically Safe Optimization

Mariah Schrum, Mark J Connolly, Eric Cole, Mihir Ghetiya, Robert Gross, and Matthew C. Gombolay, *Member, IEEE*

***Abstract*—When a robotic system is faced with uncertainty, the system must take calculated risks to gain information as efficiently as possible while ensuring system safety. The need to safely and efficiently gain information in the face of uncertainty spans domains from healthcare to search and rescue. To efficiently learn when data is scarce or difficult to label, active learning acquisition functions intelligently select a data point that, if the label were known, would most improve the estimate of the unknown model. Unfortunately, prior work in active learning suffers from an inability to accurately quantify information-gain, generalize to new domains, and ensure safe operation. To overcome these limitations, we develop Safe MetAL, a probabilistically-safe, active learning algorithm which meta-learns an acquisition function for selecting sample efficient data points in safety critical domains. The key to our approach is a novel integration of meta-active learning and chance-constrained optimization. We (1) meta-learn an acquisition function based on sample history, (2) encode this acquisition function in a chance-constrained optimization framework, and (3) solve for an information-rich set of data points while enforcing probabilistic safety guarantees. We present state-of-the-art results in active learning of the model of a damaged UAV and in learning the optimal parameters for deep brain stimulation. Our approach achieves a 41% improvement in learning the optimal model and a 20% speedup in computation time compared to active and meta-learning approaches while ensuring safety of the system.**

***Index Terms*—Aerospace control, deep learning, learning systems.**

I. Introduction

ROBOTS need the ability to safely and efficiently learn to operate in new environments, as it is not possible for engineers to explicitly program responses for every contingency. Robotic vehicles that could safely and efficiently learn their own dynamics would be capable of adapting to novel damage without crashing or needing to halt operation [20], [21]. In healthcare, robotic devices, such as deep brain stimulation (DBS) for epilepsy therapy, could automatically learn the optimal waveforms to reduce harmful electrical activity in the brain without patient-specific, manual tuning by a physician [1]. *Active learning* techniques seek to address this problem by utilizing an acquisition function to predict the *expected informativeness* of a data point [28], which is defined as the change in model's testing accuracy when adding a new data point to the training set [17]. By accurately estimating expected informativeness of a data point, data points can be judiciously selected to improve model accuracy and reduce uncertainty.

Researchers have previously investigated active learning techniques for sample efficient learning [37], [38]. However, prior work in active learning suffers from three weaknesses: 1) an inability to accurately quantify expected informativeness [16], 2) a lack of generalizability [27], and 3) a lack of safety considerations [29]. Active learning approaches typically hand-engineer heuristics or acquisition functions to select the best action [15], [29]. However, these heuristics are only proxies for true informativeness of a data point and may not accurately quantify the actual informativeness of a data point when updating the model with this new training data. Additionally, heuristics that are well suited for one active learning domain may not be effective in another. The few meta-active learning approaches proposed in recent years rely only on hand-engineered features which reduce generalizability and require expert feature selection [16]. Furthermore, prior approaches do not consider applications in safety critical domains in which constraints must be placed on the acquisition function to prevent the model from sampling unsafe configurations [28].

Yet, efficient learning is not the only criteria that must be met when dealing with safety critical domains. For example, when learning the model of a damaged UAV, one must reason about the safety of the system in addition to expected informativeness of an action to prevent the UAV from entering into an unrecoverable configuration. If the dynamics model of the damaged UAV can be learned efficiently and safely, the UAV may be able to safely land or even complete its assigned task despite the damage.

To achieve the goal of safe and efficient adaption, we introduce (Safe Meta-Active Learning) Safe MetAL, a hybrid meta-learning and mathematical programming approach that enables efficient, safe, and computationally fast optimization of a latent robotic system. The key to our approach is that we safely and efficiently meta-learn an acquisition function based on a learned representation of sample history that accurately quantifies the expected informativeness for an unknown, latent model when taking a given action and experiencing the resultant state. By directly encoding this acquisition function in a chance-constrained mixed-integer linear program (MILP), we can simultaneously

Manuscript received 23 February 2022; accepted 27 June 2022. Date of publication 25 July 2022; date of current version 9 August 2022. This letter was recommended for publication by Associate Editor K. Yamane and Editor C. Gosselin upon evaluation of the reviewers' comments. This work was supported in part by the McCamish Parkinson's Disease Innovation Program at Georgia Tech and Emory University, in part by the Office of Naval Research (ONR) under Grant N00014-19-1-2076, and in part by the NSF Accessibility, Rehabilitation and Movement Science Fellowship under Grant 1545287

Mariah Schrum and Matthew C. Gombolay are with the Institute for Robotics and Intelligent Machines, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: mschrum3@gatech.edu; matthew.gombolay@cc.gatech.edu).

Mark J Connolly, Eric Cole, Mihir Ghetiya, and Robert Gross are with the Department of Bioengineering, Emory University, Atlanta, GA 30322 USA (e-mail: mark.connolly@emory.edu; ercole@emory.edu; mghetiya@uthsc.edu; Rgross@emory.edu).

Digital Object Identifier 10.1109/LRA.2022.3193497

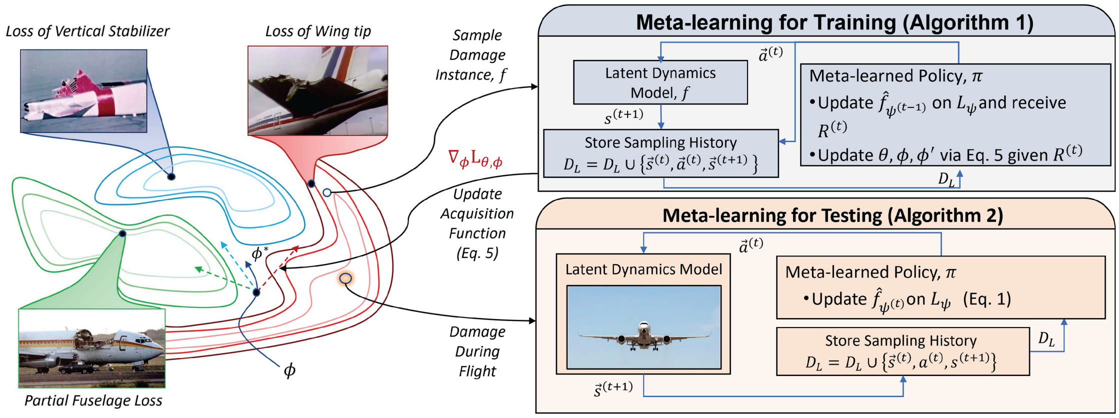

Loss of Vertical Stabilizer
Loss of Wing tip
Sample Damage Instance, f
Meta-learning for Training (Algorithm 1)
Latent Dynamics Model, f
$\vec{a}^{(t)}$
Meta-learned Policy, π
•Update $\hat{f}_{\psi^{(t-1)}}$ on L_ψ and receive $R^{(t)}$
•Update θ, ϕ, ϕ' via Eq. 5 given $R^{(t)}$
$s^{(t+1)}$
Store Sampling History $D_L = D_L \cup \{\vec{s}^{(t)}, \vec{a}^{(t)}, \vec{s}^{(t+1)}\}$
D_L
$\nabla_\phi L_{\theta,\phi}$
Update Acquisition Function (Eq. 5)
ϕ^*
Meta-learning for Testing (Algorithm 2)
$\vec{a}^{(t)}$
Latent Dynamics Model
Meta-learned Policy, π
•Update $\hat{f}_{\psi^{(t)}}$ on L_ψ (Eq. 1)
Store Sampling History $D_L = D_L \cup \{\vec{s}^{(t)}, a^{(t)}, s^{(t+1)}\}$
D_L
$\vec{s}^{(t+1)}$
Damage During Flight
ϕ
Partial Fuselage Loss

Fig. 1. Meta-learning framework, grounded in our UAV application. The red, blue and green curves represent the hypothetical manifolds within our distribution of damage scenarios. Our meta-learning algorithm samples from the distribution of damage conditions and learns a function, Q_ϕ, describing the expected informativeness of taking an action. We embed this Q_ϕ in a MILP to enforce safety-constraints, thereby ensuring safe flight while enabling efficient recovery from damage.

enforce safety guarantees [29] while taking an action which maximizes expected informativeness. This acquisition function is meta-learned offline over a distribution of tasks, allowing the policy to benefit from past experience and provide a more robust measure of the value of a labeled data point.

We demonstrate the advantage of Safe MetAL across two domains: 1) a high-dimensional damaged UAV domain and 2) a novel DBS domain, both safety critical environments in which sample efficiency is of utmost importance. Our approach outperforms previous Bayesian [1], [35], meta-learning [16], [35], and active learning approaches [15], [29] in terms of expected informativeness, safety, and computation time.

Contributions:

1) We present Safe MetAL, a meta-learning algorithm for learning a domain-specific acquisition function that accurately quantifies expected informativeness. Safe MetAL (1) meta-learns an acquisition function to quantify domain specific expected informativeness of a data point without the need for hand-derived features and ad hoc engineering, and (2) reasons explicitly about exploitation vs. exploration by trading off gaining information and probabilistically-safe control.
2) We formulate a novel bridge between deep learning and mathematical programming techniques in a way that is fully, end-to-end differentiable and trainable by embedding this meta-learned acquisition function within a chance-constrained optimization framework to achieve probabilistic guarantees.
3) We show that our approach generalizes across two disparate domains and sets a new state-of-the-art for increase in model accuracy (41%) compared to Bayesian [1], [35], active [15], [29] and meta-learning [16], [35] approaches and computational speed (+20%) versus two active and meta-learning baselines while also providing probabilistic guarantees.

II. Problem Set-Up

We describe our problem set-up via a motivating example: learning the dynamics model of a damaged UAV. In this example, our objective is to safely and efficiently learn the altered UAV dynamics, $\hat{f}_\psi$, and maintain controllability of the system despite damage. UAVs are susceptible to a range of failure scenarios that are difficult to predict and model, and, when damaged, UAVs have tight time constraints for recovery [3], [20], [21]. Specifically, we seek to determine the action the UAV should take next to provide maximum information about the nature of the damage given the UAV's previously experienced states and actions without going into unsafe configurations. To do so, we learn a function that describes the expected informativeness of taking any action, conditioned on the prior experience of the UAV and subject to safety considerations. We set up our problem in three steps: 1) active learning, 2) safety, and 3) meta-learning.

First, we define our unlabeled dataset, $D_U = \langle \vec{s}^{(i)}, \vec{a}^{(i)} \rangle_{i=1}^n$, as consisting of all possible state-actions pairs that the UAV could potentially experience and the labeled dataset, $D_L = \langle \vec{s}^{(i)}, \vec{a}^{(i)}, \vec{s}^{(i+1)} \rangle_{i=1}^m$, as the set of state transition triples experienced by the UAV in flight. $\vec{s}^{(t+1)}$ is the state that results from applying action $\vec{a}^{(t)}$ in state $\vec{s}^{(t)}$ at time t as governed by the latent dynamical model, f (Fig. 1).

Our Long Short-Term Memory (LSTM) neural network, with parameters θ, learns an encoding of sample history, $z^{(t)} = \mathcal{E}_\theta(\mathcal{S}^{(t)})$. This sample history through time, t, is defined as $\mathcal{S}^{(t)} = \langle \vec{s}^{(0)}, \vec{a}^{(0)}, \vec{s}^{(1)}, \ldots, \vec{a}^{(t-1)}, \vec{s}^{(t)} \rangle$ which we refer to as the *meta-state*. Our acquisition function, $Q_\phi : \mathcal{A} \times Z \rightarrow \mathbb{R}$, learns to map a candidate action, $\vec{a}$, to a measure of expected informativeness conditioned on the embedding of sample history, $\vec{z}$. This problem setup corresponds to a Partially Observable Markov Decision Process (POMDP), where the observations are our samples, $\vec{s}$ and the state describes the latent dynamics (i.e., the transition function, f) with actions, $\vec{a}$, discount factor, γ, and reward function, R, described below (2). We do not have access to the observation function, Ω. Similar to [22], we convert this POMDP to a Markov Decision Process (MDP) in which we use function approximation to (1) learn a compact representation, $z^{(t)}$, of the history of observations via $\mathcal{E}_\theta$ and leverage this representation to (2) train a history-dependent Q-function, Q_ϕ

We utilize expected informativeness (i.e., improvement in model accuracy due to the addition of new observations to the training set) as our reward signal for training the network. To determine the decrease in model error, as shown in (1), we create a dataset, D^{Test}, by sampling from the known dynamics model, which we have access to during training. The reward signal, $R^{(t)}$, which is defined in (2), is the decrease in model error

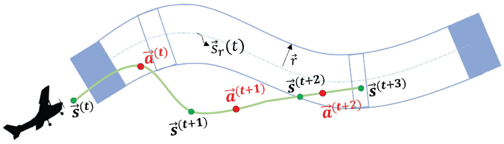


Fig. 2. This figure depicts a g volume of safety, i.e. convex constraints around reference trajectory, $\vec{s}_r(t)$. Action, $\vec{a}^{(t)}$, is an exploratory action, which may bring the system outside of the safe region. Given $\hat{f}_{\psi^{(t)}}$, Safe MetAL ensures the probability that $\vec{a}^{(t+2)}$ returns the system to a safe state is at least $1-\epsilon$.

when applying action, $\vec{a}^{(t)}$, in state, $\vec{s}^{(t)}$, and experiencing state, $\vec{s}^{(t+1)}$ (i.e., $D_L \cup \langle \vec{s}^{(t)}, \vec{a}^{(t)}, \vec{s}^{(t+1)} \rangle$). Intuitively, a large reward means that we have selected an action that greatly decreases the error of the dynamics model, $\hat{f}_\psi$. ψ is the parametrization of $\hat{f}_{\psi^{(t)}}$ at time, t.

$$L_\psi(D_L) = \frac{1}{|D_L|} \sum_{i=1}^{|D|} \left(\hat{f}_\psi \left(\vec{s}^{(i)}, \vec{a}^{(i)} \right) - \vec{s}^{(i+1)} \right)^2 \qquad (1)$$

$$R^{(t)} = \frac{\left(L_{\psi^t}(D^{Test}) - L_{\psi^{t-1}}(D^{Test}) \right)}{L_{\psi^t}(D^{Test})} \qquad (2)$$

Second, we need to incorporate *safety* when selecting the optimal action. We consider the system to be safe if there is a high probability of the system returning to a safe volume, which we discuss further in Section III. Therefore, we encode our acquisition function into a mixed-integer linear program (MILP) which allows us to impose safety constraints and choose the set of actions which maximize expected informativeness, while also ensuring safety.

In our formulation, chance-constraints allow us to model uncertainty and ensure the probability of failure remains under a certain threshold. Thus, by utilizing a chance-constrained MILP, we can efficiently arrive at a solution for non-convex optimization problems while also providing probabilistic guarantees [5]. We transform each piece-wise term in our acquisition function into a set of integer, linear constraints via the "big M" method [13]. We solve our chance-constrained MILP via linearization techniques discussed in [5], [29]. While limited prior work [31] has explored safety and chance constraints for learning and control, we go beyond this prior work by taking into account the effect that querying a label has on the underlying system's ability to remain in a safe configuration. In our damaged UAV and DBS domains, choosing a sequence of unsafe actions can lead to the UAV crashing or an ictal state in the brain. As depicted in Fig. 2, we assume a set of known safe states (e.g., level flight above the ground for a UAV) and we allow the system to deviate from a safe region temporally to gain information provided that the system has a sufficient probability of returning to a safe state. We elaborate on how these safe states are identified and the external validity in Section IV. To approximate the uncertainty of the states, we assume our model error comes from a Gaussian distribution with a known mean and variance calculated via the bootstrapping method described in [15].

Finally, we seek to enable our system to generalize beyond a single active learning task (e.g., damage to a specific part of the UAV) to a broader class of tasks (i.e., any type of damage). We aim to learn this acquisition function without hand-engineering features or heuristics. Therefore, we incorporate meta-learning to train our acquisition function, Q_ϕ, and embedding of previously experienced states and actions, $\mathcal{E}_\theta$. We train Q_ϕ over a *distribution* of optimization problems (e.g., loss of vertical stabilizer, wing damage etc.) to enable Q_ϕ to generalize to an unforseen damage scenario.

III. Safe Meta-Learning Architecture

Our architecture consists of three key components: 1) an LSTM-based representation of sample history, 2) a meta-learned acquisition function that accurately quantifies expected informativeness, and 3) safety constraints imposed via the linear program. An overview of our architecture is shown in Fig. 1, and is described below.

A. Policy

Our policy (Eq. 3) is determined by maximizing both the probability of the system remaining in a safe configuration and expected informativeness along the finite trajectory horizon, $[t, t+T)$. Therefore, our policy selects the set of actions, $\vec{a}^{(t:t+T)}$, which maximizes both safety and expected informativeness. We linearize our objective function following the linearization procedures introduced in [29].

$$\vec{a}^{(t:t+T)} = \pi \left(\mathcal{E}_\theta(\mathcal{S}^{(t)}) \right) \qquad (3)$$
$$= \underset{\vec{a}^{(t:t+T)},\epsilon}{\operatorname{argmax}} \, \lambda \left[Q_\phi \left(\vec{a}^{(t:t+T)}, \vec{z}^{(t)} \right) \right] + (1-\lambda) \left[1-\epsilon \right]$$

subject to

$$1-\epsilon \le \Pr \left\{ \left\| \vec{s}^{(t+T)} - \vec{s}_r(t) \right\|_1 \le \vec{r} \right\} \qquad (4)$$

$Q_\phi(\vec{a}^{(t:t+T)}, \vec{z}^{(t)})$ describes the expected informativeness along the trajectory when the set of actions, $\vec{a}^{(t:t+T)}$, is taken in the context of the sample history encoding, $\vec{z^{(t)}}$. π is the chance-constrained policy which selects an action that the UAV should take to maximize both expected informativeness and safety. The LSTM neural network, $\mathcal{E}_\theta$, maps the sample history, $\mathcal{S}^{(t)} = \langle \vec{s}^{(0:t)}, \vec{a}^{(0:t)} \rangle$, (i.e., previously experienced states and actions), to the encoding, $\vec{z}^{(t)}$. λ is a hyper-parameter that allows us to adjust the trade-off between safety and expected informativeness while still guaranteeing a minimum level of safety. Properly balancing λ, as in any multi-criteria optimization problem, requires domain expertise. The estimated probability of remaining in a safe configuration is $1-\epsilon$, where $\epsilon \in [0, \epsilon_{\max}]$ and $1-\epsilon_{\max}$ is the minimum acceptable safety level. We provide more details on safety in the following section.

B. Definition of Safety

Next, we detail our safety constraints which are enforced via our MILP. We define a volume of safety, as depicted in Fig. 2, around a desired reference trajectory, $\vec{s}_r(t)$, and enforce the constraint that the system must be able to return with probability $1-\epsilon$ to a state, $\vec{s}_{t+T}$, at time $t+T$, such that $\vec{s}_{t+T}$ is within this volume of safety. Intuitively, this means that the UAV will be able to take an action outside of the volume of safety to gain information but must return to a safe state at time $t+T$. Mathematically, we define this safety constraint in (4). $\vec{s}_r$ defines a safe state (e.g., straight and level flight for a UAV), and $\vec{r}$ is the radius encompassing all known safe system states. The radius of the volume is a hyperparameter defined by the user and requires domain expertise to determine. The closer the user wants the system to remain to the nominal safe state, the smaller the radius should be. $1 - \epsilon$ is the probability of remaining in the safe region. This volume can be converted into linear constraints, thus creating a convex optimization problem [29]. Leveraging such a pre-defined safety envelope is consistent with prior work in safe robotics and chance-constrained optimization [9], [23]–[25], [40]. By formulating our safety constraints in this way, we can guarantee a minimum probability of safety while simultaneously optimizing for additional safety and expected informativeness. In other words, the MILP will select an action that meets the minimum safety requirements, and if possible, will select an even safer action than the minimum specified level of safety, all other things being equal.

C. Meta-Learning

To infer the acquisition function, we meta-learn over a distribution of related tasks, which, in our motivating example, consist of various damage modes of the UAV (e.g., wing damage, actuator damage, etc.) as shown in Fig. 1. By meta-learning over this distribution, we can construct an acquisition function that accurately defines the expected informativeness of an action when learning the unknown UAV dynamics model.

The acquisition function, Q_ϕ, is trained via Deep Q-Learning [11] with target network, $Q_{\phi'}$, which has been shown in prior work to improve training stability [33]. The learned acquisition function, Q_ϕ, is utilized by our MILP policy, which selects the optimal actions, $\vec{a}^{(t:T)}$, subject to safety-constraints. The reward, $R^{(t)}$, for taking a set of actions in a given state is defined as the decrease in the MSE error of the model, $\hat{f}_{\psi^{(t)}}$, achieved by adding training data, $\langle \vec{s}^{(t)}, \vec{a}^{(t)}, \vec{s}^{(t+1)} \rangle$, to D_L, as described in (1) and (2). The Q-function is trained on a set of optimization problems drawn from a distribution of similar black-box functions to minimize the Bellman Residual (5).

$$\mathcal{L}_{\theta,\phi} = \Big(R^{(t)} + \gamma Q_{\phi'}\left(\pi\left(\mathcal{E}_\theta(\mathcal{S}^{(t+1)})\right), \vec{z}^{(t+1)}\right) - Q_\phi\left(\vec{a}^{(t)}, \vec{z}^{(t)}\right)\Big)^2 \tag{5}$$

This Bellman loss of the Q-function is backpropagated through the Q-function in the MILP and through the LSTM encoder, $\mathcal{E}_\theta$. The dynamics model, $\hat{f}_{\psi^{(t)}}$, is retrained with each new set of state-action pairs.

D. Algorithm

Algorithm 1 describes our training procedure. For each episode, we sample from the distribution of altered dynamics and limit each episode to the number of time steps, M, tuned to collect enough data to accurately learn the dynamics. At each iteration, we select $\vec{a}^{(t)}$ (line 6) via our MILP objective described in (3) and execute the action to observe the resultant state, $\vec{s}^{(t+1)}$ (line 7–8). Our dynamics model, $\hat{f}_{\psi^{(t)}}$, is retrained by minimizing the MSE, as shown in (1). After observing the reward (2), we update our Q-function (line 11–12) via a sampled batch of transitions.

Algorithm 2 describes how we perform our online, safe, active learning. Intuitively, our algorithm initializes a new dynamics model (line 2) to represent the unknown or altered dynamics, and we iteratively sample information rich, safe actions via our MILP policy (line 5), update $\hat{f}_{\psi^{(t)}}$, (line 9) and repeat. We assume at test time that the unknown model comes from the same distribution as the training models.

Algorithm 1: Meta-Learning for Training.

```
 1: Randomly initialize $Q_\phi$ and $Q_{\phi'}$ with weights $\phi = \phi'$
 2: Initialize replay buffer, D
 3: for episode=1 to Ndo
 4:   Initialize $\hat{f}_{\psi^{(0)}}$ based on meta-learning distribution
 5:   for t=1 to M do
 6:     Select $\vec{a}^{(t)}$ from (3)
 7:     Execute $\vec{a}^{(t)} + \mathcal{N}$ with exploration noise, $\mathcal{N}$
 8:     Observe state, $\vec{s}^{(t+1)}$
 9:     $D_L \leftarrow D_L \cup \langle \vec{s}^{(t)}, \vec{a}^{(t)}, \vec{s}^{(t+1)} \rangle$
10:     $\psi^{(t)} \leftarrow argmin_\psi L_\psi(D_L)$; observe $R^{(t)}$
11:     Update $Q_\phi$ and $\mathcal{E}_\theta$ via (5)
12:     $Q_{\phi'} \leftarrow \tau Q_\phi + (1-\tau) Q_{\phi'}$
13:   end for
14: end for
```

Algorithm 2: Meta-Learning for Testing.

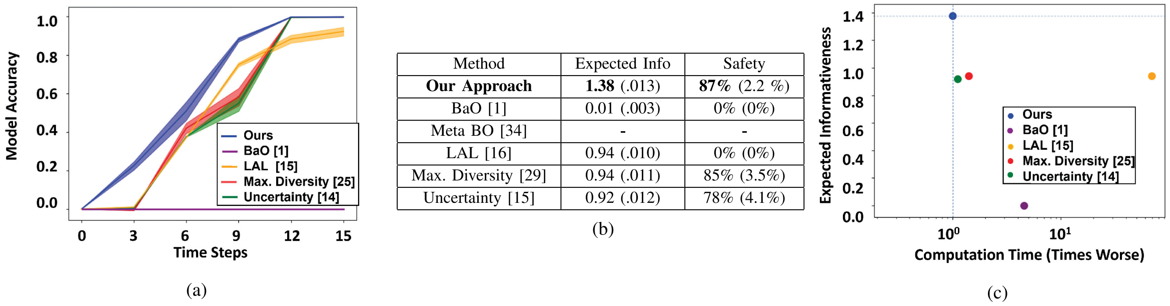
```
 1: Draw test example from distribution
 2: Initialize $\hat{f}_{\psi^{(0)}}$ based on meta-learning distribution
 3: $D_L \leftarrow \emptyset$
 4: for t=1 to M do
 5:   Select $\vec{a}^{(t)}$ according to (3)
 6:   Execute $\vec{a}^{(t)}$
 7:   Observe state $\vec{s}^{(t+1)}$
 8:   $D_L \leftarrow D_L \cup \langle \vec{s}^{(t)}, \vec{a}^{(t)}, \vec{s}^{(t+1)} \rangle$
 9:   $\psi^{(t)} \leftarrow argmin_\psi L_\psi(D_L)$
10: end for
```

IV. Experimental Evaluation

We compare Safe MetAL against several baseline approaches in two experimental domains described below.

A. High-Dimensional UAV Domain

Safe control of damaged UAVs is a difficult problem in robotics due to the tight time constraints and non-linear dynamics. We test our algorithm's ability to learn the non-linear dynamics of a UAV before the UAV enters an unrecoverable configuration (e.g., crashing). Because active learning algorithms can be ineffective in high-dimensional domains, our aviation domain also serves to stress test our algorithm's ability to quickly learn a high-dimensional dynamics model given tight time constraints. We base our simulation on theoretical damage models from prior work describing the full equations of motion [26], [36], [39] within the Flightgear virtual environment. The objective of this domain is to learn the altered dynamics that results from the damage and to maintain safe flight. The UAV takes an information rich action potentially resulting in a deviation outside of the d-dimensional volume of safety, guaranteeing that the UAV



Method	Expected Info	Safety
Our Approach	**1.38** (.013)	**87%** (2.2 %)
BaO [1]	0.01 (.003)	0% (0%)
Meta BO [34]	-	-
LAL [16]	0.94 (.010)	0% (0%)
Max. Diversity [29]	0.94 (.011)	85% (3.5%)
Uncertainty [15]	0.92 (.012)	78% (4.1%)

Fig. 3. This figure depicts our empirical validation, generated via Monte Carlo simulation, in our robotic UAV domain benchmarking algorithm accuracy per time step (Fig. 3(a)), overall expected informativeness over the time horizon (Fig. 3(b)), and vs. computation time (Fig. 3(c)). Error is calculated in batches of three time steps, enabling the robot to deviate from the safe region temporarily to gain information. The results shown in Fig. 3(a) comply with safety results reported in Fig. 3(b).

returns to a safe state with probability $1-\epsilon$ via action $\vec{a}^{(t+1)}$ at the end of the planning horizon.

B. Deep Brain Stimulation (DBS)

DBS is a cutting-edge approach for treating seizure conditions that cannot be controlled via pharmacological methods. Currently, surgeons employ trial-and-error to find control settings that reduce seizures. However, there is no clear mapping from parameter values to reduction in seizures that applies to all patients, as the optimal parameter settings can depend on placement of the device, the individual anatomy, and other confounding factors. Further, a latent subset of parameters can cause negative side-effects. In keeping with [1], we create simulation environments based on data from six rats where, at each DBS parameter setting, the cognitive function of a rat is measured by a "memory score." Data from each rat is then dissimulated into many digital twins of the rat, creating a population pool over which we can meta-learn. To create these digital twins, we employ a validated *in silico* procedure in which we bootstrap Gaussian Process models trained on *in vivo* data of DBS in rats to create a virtual experimental domain. The task is to determine the DBS parameters (i.e., signal amplitude) in the simulation environments that maximize each rat's memory score (i.e., rat's ability to recall the location of objects) without causing unwanted side effects (e.g., memory deficits or seizures) which occur when the memory score drops below zero. The reward signal utilized by our meta-learner is the percent decrease in error between the predicted and actual optimal parameters. This domain and the established *in silico* evaluation procedure are described further in [1].

V. Results

A. Baseline Comparisons

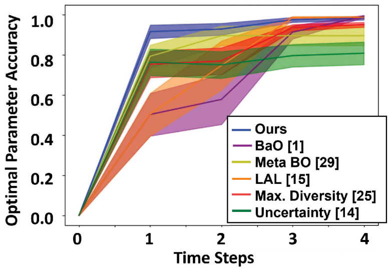
To demonstrate that meta-learning is a vital component of our framework and produces results superior to prior work, we benchmark against active learning functions, Epistemic Uncertainty [15] and Maximizing Diversity [29]. These active learning functions are linearized and embedded in our safety constrained framework therefore providing a head-to-head comparison between our meta-learned acquisition function and these active learning heuristics. We additionally benchmark against several Bayesian and meta-learning approaches. We empirically validate that Safe MetAL outperforms baselines in the DBS and UAV domains in terms of its ability to safely and actively learn latent parameters.

- *Epistemic Uncertainty [15]* - Selects the action which maximizes the uncertainy of the model, while also imposing safety constraints via a chance-constrained linear program.
- *Maximizing Diversity [29]* - Selects actions which maximize the difference between previous states and actions, subject to safety constraints via a chance-constrained linear program.
- *Bayesian Optimization (BaO) [1]* - Developed in previous work for the DBS domain (Section IV) and is based upon a Gaussian Process model which attempts to efficiently determine the optimal parameters.
- *Meta Bayesian Optimization (Meta BO) [34]* - Meta-learns a Gaussian process prior offline.
- *Learning Active Learning (LAL) [16]* - Meta-learns an acquisition function leveraging hand-engineered features.

B. Active Learning

Results from both the UAV and the DBS domains empirically validate that our algorithm more efficiently learns the optimal parameters (Fig. 4) and system dynamics (Fig. 3) in both domains compared to baseline approaches. The Bayesian baseline, BaO struggles to learn in the UAV domain, and we find that Meta BO is computationally intractable due to the complexity of the task. Again, Safe MetAL outperforms both active learning heuristics, achieving a 46% improvement over Maximizing Diversity and a 49% improvement over Uncertainty. Safe MetAL achieves a 47% higher expected informativeness versus LAL.

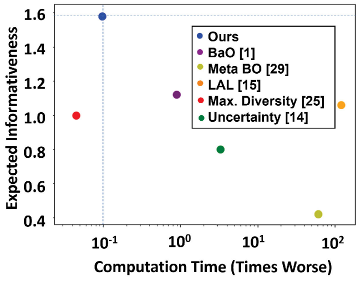
We find similarly positive results in the DBS domain. In this domain, Safe MetAL selects an action that results in 58% higher expected informativeness and a 267% higher expected informativeness on average compared to our two Bayesian baselines, BaO and Meta BO respectively. Compared to our active learning baselines, Maximizing Diversity and Uncertainty, Safe MetAL performs 41% and 98% better in terms of average expected informativeness respectively. This large increase in expected informativeness that Safe MetAL is able to achieve compared to hand-engineered heuristics, suggests that the meta-learning aspect of Safe MetAL is vital for synthesizing a precise, task-specific acquisition function. Lastly, we show that Safe MetAL outperforms by 47% our meta-learning baseline, LAL, which meta-learns over hand-engineered features. These results demonstrate that our meta-learned embedding is more capable



(a)

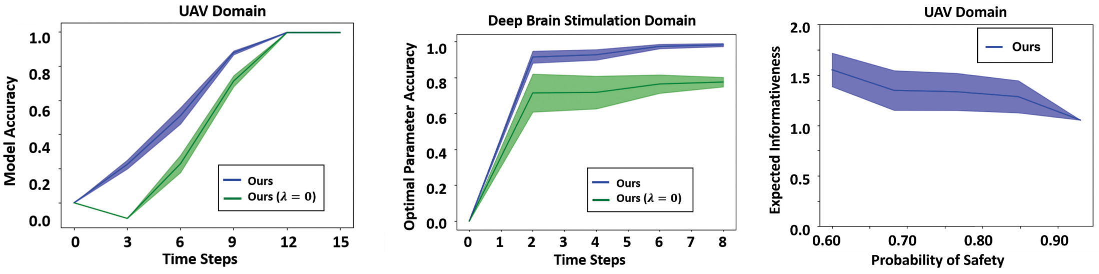
Method	Expected Info	Safety
Our Approach	**1.58** (.022)	**97%** (3.2 %)
BaO [1]	1.00 (.105)	86% (3.0%)
Meta BO [34]	0.43 (.183)	83% (11%)
LAL [16]	1.07 (.098)	88% (2.3%)
Max. Diversity [29]	1.12 (.153)	90% (1.9%)
Uncertainty [15]	0.80 (.241)	**97%** (1.1%)

(b)



(c)

Fig. 4. This figure depicts our empirical validation in the DBS domain, benchmarking algorithm accuracy per time step (Fig. 4(a)), overall (Fig. 4(b)), and vs. computation time (Fig. 4(c)). The optimal parameter accuracy is defined as $1 - \frac{\vec{a}^* - \hat{\vec{a}}}{\vec{a}^*}$ where $\vec{a}^*$ is the optimal stimulation parameter and $\hat{\vec{a}}$ is the predicted parameter. In Fig. 4(b) we also report the ground truth safety of our algorithm compared to baselines. The results shown in Fig. 4(a) comply with the safety results reported in Fig. 4(b).



(a) (b) (c)

Fig. 5. This figure shows the results of our ablation analysis and the trade-off between expected informativeness and safety. In Fig. 5(a) (UAV domain) and 5(b) (DBS domain), we set $\lambda = 0$, meaning there is no active learning and only safety is maximized. In both domains, our meta-learned acquisition function is an important component to achieve efficient learning. Fig. 5(c), shows an ablation study, demonstrating the trade-off between expected informativeness and safety in the UAV domain when we vary λ. We show that we can tune λ to achieve the desired tradeoff between expected informativeness and safe operation.

of extracting salient information than the hand-engineered features in LAL. To further verify that the meta-learning aspect of Safe MetAL is necessary for achieving high expected informativeness, we perform an ablation study as shown in Figs 5(a) and 5(b).

C. Safety

Because Safe MetAL is able to more quickly learn the optimal parameter settings, it is also able to ensure safe operation to a greater degree than the baselines in both domains. To empirically validate the safety of each algorithm, we perform a Monte Carlo simulation and determine the percentage of the time that the UAV is able to return to the safe region. We find that Safe MetAL achieves an 87% probability the UAV will return to the safe region (Fig. 3(a)). As shown in Fig. 3(b), Safe MetAL outperforms the baselines simultaneously in safety and expected informativeness in the high-dimensional UAV domain.

In the DBS domain, Safe MetAL achieves a 6.3% higher guarantee of safety compared to Maximizing Diversity [29] in Fig. 4. Safe MetAL achieves a 98% greater expected informativeness compared to Uncertainty [15] and achieves an equivalent safety guarantee. In Fig. 5(c), we show the trade-off between the probability of safety as determined by the MILP and the expected informativeness of an action as a result of adjusting λ. This flexibility allows for greater emphasis on safety in more safety critical domains, whereas in less safety critical domains, these constraints can be relaxed in favor of higher expected informativeness.

D. Computation Time

The computation time of active learning algorithms can be of critical importance especially in highly unstable systems such as a damaged UAV. Across both domains, Safe MetAL not only achieves a more efficient reduction in model error and improvement in expected informativeness, but we are also faster than all baselines in the high-dimensional UAV domain (Fig. 3(c)). We demonstrate that we are 20% and 29% faster than the baseline acquisition functions, Uncertainty [15] and Maximizing Diversity [29] respectively. We also demonstrate that we are more than 69 times faster than our meta-learning baseline [16] and 78% faster than our Bayesian baseline, BaO [1].

In the DBS environment (Fig. 4(c)), BaO has a slight advantage in computation time, but Safe MetAL trades the time for 58% greater expected informativeness. Additionally, Safe MetAL is 68x faster than LAL and 61x faster than Meta BO, our two meta-learning benchmarks.

VI. Related Work

A. Active Learning

Active learning acquisition functions provide heuristics to select the candidate unlabeled training data sample that, if the label were known, would provide the most information to the model being learned [6], [7], [14], [15]. In [15], the sample is selected that the learner is least certain about. In work by [1], the authors utilize Expected Improvement (EI) heuristic to balance exploration versus exploitation to determine the optimal stimulation parameters in DBS. Prior literature has also investigated on-the-fly active learning and meta-active learning [2], [16]. [16] describes the algorithm Learning Active Learning (LAL). The authors present a meta-learning method for learning an acquisition function in which a regressor is trained to predict the reduction in model error of candidate samples via hand engineered features. Volpp *et al.* [34] alternatively considers a Gaussian Process based method to meta-train an acquisition function on a distribution of tasks. Work by Geifman et al. [12] actively learns the neural network architecture that is most appropriate for a given task, e.g. active learning. Pang et al. [27] additionally proposed a method to learn an acquisition function that generalizes to a variety of classification tasks. Yet, this work has only been demonstrated for classification.

B. Meta-Learning for Dynamics

Prior work has attempted to address the problem of learning altered dynamics via meta-learning [8]. Belkhale et al. [4] investigated a meta-learning approach to learn the altered dynamics of a UAV carrying a payload; the authors train a neural network on prior data to predict environmental and task factors to inform how to adapt to new payloads. Finn et al. [10] presented a meta-learning approach to quickly learning a control policy. In this approach, a distribution over prior model parameters that are most conducive to learning the new dynamics is meta-learned offline. While this approach provides fast policies for learning new dynamics, it does not explicitly reason about sample efficiency or safety.

C. Safe Learning

Prior work has investigated safe learning in the context of Bayesian optimization and safe reinforcement learning. For example, Sui et al. [30] developed the SafeOpt which balances exploration and exploitation to learn an unknown function; however, this approach makes significant limiting assumptions about the underlying nature of the task. Turchetta et al. [32] safely explore an MDP by defining an unknown safety constraint updated during exploration, and Zimmer et al. [41] utilize a Gaussian process for safely learning time series data. Additionally, Nakka et al. introduced Info-SNOC which utilizes chance-constraints to safely learn unknown dynamics [19]. However, these approaches do not incorporate knowledge from prior data to increase sample efficiency, limiting their ability to choose the optimal action. Schrum and Gombolay [29] attempt to overcome this problem by employing a novel acquisition function, Maximizing Diversity, to quickly learn altered dynamics in a chance constrained framework. Yet, the hand engineered acquisition function limits the capabilities of this approach.

VII. Discussion

We present a novel architecture, SafeMetAL, which, unlike previous hand-engineered approaches, leverages sample history to meta-learn a domain-specific acquisition function for safe and efficient control of an unknown system. Through our empirical investigation, we demonstrate that our meta-learned acquisition function operating within a chance-constrained optimization framework outperforms prior work in active learning, meta-learning, and Bayesian optimization [1], [15], [16], [29], [35]. Our approach simultaneously increases expected informativeness while decreasing computation time. Safe MetAL achieves a 41% increase in expected informativeness while decreasing computation time by 20% versus active learning and Bayesian baselines in the DBS domain and is more than 60x faster versus meta-learning baselines in the UAV domain. We find that MetaBO is ill-suited for the UAV domain due to its high dimensionality.

Furthermore, our chance-constrained framework combined with higher sample efficiency results in greater probability of safe operation compared to prior work. The safety results for both LAL and BAO in our UAV domain are very poor due to the fact that both lack built-in safety constraints. Taking a single action in the unstable UAV domain that does not comply with any safety guarantees results in the UAV moving out of the safe region and into an unrecoverable configuration.

We additionally demonstrate state-of-the-art performance in a healthcare domain, demonstrating that our approach generalizes across diverse systems. We are able to outperform all active learning and meta-learning baselines in expected informativeness and safety. We thus demonstrate Safe MetAL's ability to learn the dynamics of a high-dimensional and safety critical UAV as well as the optimal parameter setting for control of a biological system (i.e., the brain) via DBS.

To the best of our knowledge, Safe MetAL is the first architecture to meta-learn an acquisition function for active learning embedded within a chance-constrained program for probabilistically safe control. Further our approach sets a new state of the art over prior work ([1], [15], [16], [29], [35]) for active learning across two, disparate domains. Our novel, deep learning architecture, offers a unique ability to learn an LSTM-based embedding of sample history while utilizing the power of deep Q-learning to learn a task-specific acquisition function. Safe MetAL's is able to optimize both for safety and expected informativeness by embedding our learned acquisition function in a chance constrained optimization framework. With this novel formulation, we demonstrate that Safe MetAL maintains a high probability of safety while also maximizing the expected informativeness based on a learned representation of sample history.

VIII. Limitations and Future Work

Safe MetAL assumes that the safety region is defined by an unchanging volume of safety and that uncertainty over our states is Gaussian. Additonally, Safe MetAL requires data to meta-learn an acquisition function. However, our results demonstrate that Safe MetAL enables greater expected informativeness and safety when sufficient training data is available. The distribution of scenarios from which we meta-learn over can be determined either by a domain expert or autonomously by a fleet of robots. First, a domain expert could posit various failure modes (e.g., partial wing damage, actuator failure, etc.) and distributions of cases describing possible dynamics for those modes (e.g., dynamics for partial wing loss of 25%, 50%, etc.). These finite set of cases could be artificially expanded through data augmentation, e.g. adding noise to each mode, similar to domain randomization in Sim2Real transfer [18]. Alternatively, a fleet of robots could collect and train on data on all novel situations experienced by any robot. Finally, we hypothesize that Safe MetAL's performance depends on the representativeness of the training data, which we will explore further in future work.

IX. Conclusion

In this paper, we demonstrate Safe MetAL a state-of-the art meta-learning approach for active learning for control. In our approach we 1) accurately quantify domain specific expected informativeness, 2) learn from sample history to improve generalizability and 3) include safety constraints to probabilistically ensure safe sample selection. We demonstrate that our approach achieves a 41% increase in expected informativeness, a 20% speedup in computation time and ensures a high degree of safety across both domains.

References

[1] O. Ashmaig, M. Connolly, Robert E. Gross, and B. Mahmoudi, “Bayesian optimization of asynchronous distributed microelectrode theta stimulation and spatial memory,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 2683–2686.

[2] P. Bachman, A. Sordoni, and A. Trischler, “Learning algorithms for active learning,” in *Proc. Mach. Learn. Res.*, 2016, pp. 301–310.

[3] N. Bakshi, “Model reference adaptive control of quadrotor UAVs: A neural network perspective,” in *Adaptive Robust Control Syst.*, 2018, Art. no. 135.

[4] S. Belkhale, R. Li, G. Kahn, R. McAllister, R. Calandra, and S. Levine, “Model-based meta-reinforcement learning for flight with suspended payloads,” *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 1471–1478, Apr. 2021.

[5] L. Blackmore, M. Ono, and Brian C. Williams, “Chance-constrained optimal path planning with obstacles,” *IEEE Trans. Robot.*, vol. 27, no. 6, pp 1080–1094, Dec. 2011.

[6] R. Burbidge, Jem J. Rowland, and Ross D. King, “Active learning for regression based on query by committee,” in *Intelligent Data Engineering and Automated Learning*, 2007, pp. 209–218.

[7] W. Cai, M. Zhang, and Y. Zhang, “Batch mode active learning for regression with expected model change,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp 1668–1681, Jul. 2017.

[8] I. Clavera, A. Nagabandi, Ronald S. Fearing, P. Abbeel, S. Levine, and C. Finn, “Learning to adapt : Meta-learning for model-based control,” 2009, *arXiv:1803.11347*.

[9] S. Dai, S. Schaffert, A. Jasour, A. Hofmann, and B. Williams, “Chance constrained motion planning for high-dimensional robots,” in *Proc. Int. Conf. Robot. Automat.*, 2019, pp 8805–8811.

[10] C. Finn, K. Xu, and S. Levine, “Probabilistic model-agnostic meta-learning,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2018, pp. 9537–9548.

[11] M. Ganger, E. Duryea, and W. Hu, “Double Sarsa and double expected Sarsa with shallow and deep learning,” *J. Data Anal. Inf. Process.*, vol. 4 no. 4, pp 159–176, 2016.

[12] Y. Geifman and R. El-Yaniv, “Deep active learning with a neural architecture search,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2018, pp. 5976–5986.

[13] I. Griva, Stephan G. Nash, and A. Sofer, *Soc. for Ind. Math.*. Philadelphia, PA, USA: SIAM, 2009.

[14] M. Hasenjager and H. Ritter, “Active learning with local models 1 introduction,” *Neural Process. Lett.*, vol. 7, pp. 107–117, 1998.

[15] T. Hastie, R. Tibshirani, and J. Friedman, “Model assessment and selection,” in *The Elements of Stat. Learn. Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2017, pp. 249–252.

[16] K. Konyushkova, S. Raphael, and P. Fua, “Learning active learning from data,” in *Proc. Conf. Neural Inf. process. Syst.*, 2017, pp. 1–11.

[17] J. C. D. MacKay, “Information-based objective functions for active data selection,” *Neural Computation*, vol. 4 no. 4, pp 590–604, 1992.

[18] B. Mehta, M. Diaz, F. Golemo, Christopher J. Pal, and L. Paull, “Active domain randomization,” in *Proc. Conf. Robot Learn.*, PMLR, 2020, pp. 1162–1176.

[19] Y. K. Nakka, A. Liu, G. Shi, A. Anandkumar, Y. Yue, and Soon Jo Chung, “Chance-constrained trajectory optimization for safe exploration and learning of nonlinear systems,” *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 389–396, Apr. 2021.

[20] National Transportation Safety Board, “In-Flight Separation of Vertical Stabilizer American Airlines Flight 587, Airbus Industrie A300-605R, N14053,” Belle Harbor, New York, 2001.

[21] National Transportation Safety Board, “In-flight Separation of Right Wing Flying Boat, Inc. (doing business as Chalk’s Ocean Airways) light 101 Grumman Turbo Mallard (G-73 T), N2969,” Port of Miami, Florida, USA, 2005.

[22] K. Nguyen, “Converting POMPDs into MDPs using history representation,” Tech. Rep., 2021.

[23] M. Ono, M. Pavone, Y. Kuwata, and J. Balaram, “Chance-constrained dynamic programming with application to risk-aware robotic space exploration,” *Auton. Robots*, vol. 39, no. 4, pp. 555–571, 2015.

[24] M. Ono and Brian C. Williams, “Iterative risk allocation: A new approach to robust model predictive control with a joint chance constraint,” in *Proc. Conf. Decis. Control*, 2008, pp. 3427–3432.

[25] M. Ono, Brian C. Williams, and L. Blackmore, “Probabilistic planning for continuous dynamic systems under bounded risk,” *J. Artif. Intell. Res.*, vol. 46, pp 511–577, 2013.

[26] J. A. Ouellette, “Flight dynamics and maneuver loads on a commercial aircraft with discrete source damage,” M.S. thesis, Aerosp. Eng. Virginia Polytech. Inst. State Univ., 2010.

[27] K. Pang, M. Dong, Y. Wu, and T. Hospedales, “Meta-learning transferable active learning policies by deep reinforcement learning,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1–8.

[28] P. Ren et al., “A survey of deep active learning,” *ACM Comput. Surv.*, vol. 5, pp. 1–40, 2021.

[29] M. L. Schrum and M. C. Gombolay, “When your robot breaks : Active learning during plant failure,” *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp 438–445, Apr. 2020.

[30] Y. Sui, A. Gotovos, J. W. Burdick, and A. Krause, “Safe exploration for optimization with gaussian processes,” in *Proc. Int. Conf. Mach. Learn.*, 2015.

[31] E. Timmons et al., “Information-driven and risk-bounded autonomy for scientist avatars,” in *Proc. ASCEND*, 2021, Art. no. 4023.

[32] M. Turchetta, F. Berkenkamp, and A. Krause, “Safe exploration in finite Markov decision processes with Gaussian processes,” *Adv. Neural Inf. Process. Syst.*, 2016.

[33] H. V. Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double q-learning,” in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, pp. 2094–2100.

[34] M. Volpp et al., “Meta-learning acquisition functions for transfer learning in Bayesian optimization,” 2019, *arXiv:1904.02642*.

[35] L. Wang, Evangelos A. Theodorou, and M. Egerstedt, “Safe learning of quadrotor dynamics using barrier certificates,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 2460–2465.

[36] E. J. Watkiss, “Flight dynamics of an unmanned aerial vehicle,” *Flight Dyn.*, Calhoun, p. 3, 1994.

[37] S. Yan, K. Chaudhuri, and T. Javidi, “Active learning from imperfect labelers,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29, pp. 2136–2144.

[38] C. Zhang and K. Chaudhuri, “Active learning from weak and strong labelers,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 703–711.

[39] Y. Zhang, C. C. de Visser, and Q. P. Chu, “Aircraft damage identification and classification for database-driven online flight-envelope prediction,” *J. Guidance, Control, Dyn.*, vol. 41, pp. 449–460, 2017.

[40] H. Zhu and J. Alonso-Mora, “Chance-constrained collision avoidance for MAVs in dynamic environments,” *IEEE Robot. Automat. Lett.*, vol. 4, no. 2, pp 776–783, Apr. 2019.

[41] C. Zimmer, M. Meister, and D. Nguyen-Tuong, “Safe active learning for time-series modeling with Gaussian processes,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp 2730–2739.