Failure-Averse Active Learning for Physics-Constrained Systems

Cheolhei Lee, Xing Wang, Jianguo Wu, *Member, IEEE*, and Xiaowei Yue, *Senior Member, IEEE*

***Abstract*—Active learning is a subfield of machine learning that is devised for the design and modeling of systems with highly expensive sampling costs. Industrial and engineering systems are generally subject to physics constraints that may induce fatal failures when they are violated, while such constraints are frequently underestimated in active learning. In this paper, we develop a novel active learning method that avoids failures considering implicit physics constraints that govern the system. The proposed approach is driven by two tasks: safe variance reduction explores the safe region to reduce the variance of the target model, and safe region expansion aims to extend the explorable region. The integrated acquisition function is devised to conflate two tasks and judiciously optimize them. The proposed method is applied to the composite fuselage assembly process with consideration of material failure using the Tsai-Wu criterion, and it is able to achieve zero failure without the knowledge of explicit failure regions.**

***Note to Practitioners*—This paper is motivated by engineering systems with implicit physics constraints related to system failures. Implicit physics constraints refer to failure processes in which explicit analytic forms do not exist, so demanding numerical simulations or real experiments are required to check one's safety. The main objective of this paper is to develop an active learning strategy that safely learns the target process in the system by minimizing failures without preliminary reliability analysis. The proposed method mainly targets real systems whose failure conditions are not thoroughly investigated or uncertain. We applied the proposed method to the predictive modeling of composite fuselage deformation in the aircraft manufacturing process, and it achieved zero failure in sampling by considering the composite failure criterion.**

***Index Terms*—Active learning, physics-integrated machine learning, composite structures assembly.**

I. Introduction

ACTIVE learning is a subfield of machine learning that maximizes information acquisition to reduce the labelling cost in supervised learning [1]. Contrary to passive learning such as factorial, maximum entropy, Latin Hypercube design (LHD) [2], active learning optimizes the acquisition function that quantifies the potential importance of unlabelled data, and interactively queries the most informative design point to the oracle. Generally, acquisition functions refer to the up-to-date model or labelled data, while there are various strategies that can be adopted according to the preference in information criteria and characteristics of systems. The engineering domain is one of the beneficiaries of active learning due to the high complexity of systems and the expensive evaluation cost [3], [4], [5].

However, active learning in engineering applications has been mostly utilized without considering coexisting or inherent constraints that may have different processes thereof. It is very crucial to consider such constraints in engineering systems, since most of them are subject to physics constraints that may induce fatal and irreversible failures. For example, design of the automatic shape control system in composite aircraft manufacturing need to consider potential material failures such as crack, buckling, and delamination caused by intolerable inputs [6]. The inverse partial differential equation (PDE) problem is another example that is used to calibrate parameters in physics models based on observations. It usually involves physics constraints comprised of different PDEs, and the constraints must be satisfied in order to build first-principle models. In both cases, the application of active learning without considering physics constraints may induce fatal failures or biased models, so constraints must be considered in engineering problems.

To consider physics constraints in active learning, it is straightforward to define a safe region where design points satisfy the safe conditions (with high probability at least), and conduct active learning within the safe region. However, it is not always possible in practice, since physics constraints cannot be explicitly attained due to the complex nature of the system. Examples include the crash damage analysis of commercial vehicles composed of different materials, and physicochemical interactions in corrosion of alloys. In these cases, fundamental physics laws and equations cannot be directly applied or are insufficient to accommodate the complexities of mechanisms. Physics-based numerical methods

Manuscript received 30 April 2022; revised 18 July 2022 and 6 September 2022; accepted 7 October 2022. This article was recommended for publication by Associate Editor K. Paynabar and Editor J. Li upon evaluation of the reviewers' comments. The work of Cheolhei Lee and Xiaowei Yue was supported in part by the U.S. National Science Foundation under Grant CMMI-2035038 and in part by the Grainger Frontiers of Engineering Grant Award from the National Academy of Engineering. The work of Xing Wang was supported by the U.S. National Science Foundation under Grant 2153329. *(Corresponding author: Xiaowei Yue.)*

Cheolhei Lee and Xiaowei Yue are with the Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: cheolheil@vt.edu; xwy@vt.edu).

Xing Wang is with the Actuarial Science and Mathematics Department, Illinois State University, Normal, IL 61790 USA (e-mail: xwang70@ilstu.edu).

Jianguo Wu is with the Department of Industrial Engineering and Management, Peking University, Beijing 100080, China (e-mail: j.wu@pku.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TASE.2022.3213827.

Digital Object Identifier 10.1109/TASE.2022.3213827

such as finite difference methods and finite element methods (FEM) [7] are well-established and convincing to analyze large classes of complex structures including failures, while they are too computationally demanding to identify the entire safe region. Moreover, their deterministic solutions are vulnerable to various uncertainty sources such as material properties, geometries, and loads.

In order to circumvent the aforementioned limitations of physics-based approaches, machine learning models have been widely used in the engineering domain due to their flexibility, inexpensive prediction, and capability of uncertainty quantification (UQ). Physics information can be highly advantageous for machine learning in several aspects such as generalization and physical consistency [8]. Especially, Gaussian processes (GPs) have shown remarkable performance in stochastic analysis of structural reliability that aims to evaluate the probability of system failure [9]. A common reliability analysis approach employs the GP surrogate model of performance function associated with the system failure, and uses the acquisition function (e.g., [10] and [11]) that leads to sampling near the boundary of safe and failure regions. The boundary is called the limit-state, which is the margin of acceptable structural design. However, the reliability analysis is mainly interested in the response surface associated with failure, and it is time-consuming due to the requirement of a large number of samples to estimate the underlying distribution at the limit-state. Hence, it can be data-inefficient to implement reliability analysis prior to the estimation of safe region. So the development of flexible active learning that takes account of target and failure processes simultaneously is promising for systems with implicit physics constraints.

The principle of active learning is exploitation of knowledge from observations, and exploration by tackling the knowledge such as the design point with maximum entropy or the most disagreeable point in the set of hypotheses. However, if implicit constraints exist in the design space, active learning can be very challenging, since the most informative design point may be located in the failure region. Conversely, if active learning is too conservative to avoid failures, the resulted model will be vulnerable in the unexplored safe region. Consequently, active learning for physics-constrained systems should simultaneously take into account the following objectives:

1) maximizing the information acquisition for the target model;
2) expanding the explorable safe region by focusing on constraint functions,

and they are must be achieved safely. Definitely, two objectives are at odds since they are associated with different functions, so the active learning strategy must be judiciously controlled.

In this paper, we propose an active learning methodology for systems that are constrained by implicit failure processes. The overview of the proposed method is illustrated in Fig. 1. The target process in the system which we aim to learn is subject to failure processes with implicit physics constraints. Both processes can be evaluated via physics-based simulations or experiments, which are expensive to observe. In order to alleviate the sampling cost, we build the predictive model for the target function and constraints by imposing GP priors on them, and initializing with a proper design (e.g., space-filling). The objective of our active learning is to train the predictive model of the target function data-efficiently and safely by minimizing the cost from undesirable failures incurred by implicit physics constraints. The active learning strategy is built upon two sub-strategies: (i) safe variance reduction; and (ii) safe region expansion. To minimize the predictive model variance with respect to the implicit safe region, safe variance reduction explores the estimated safe region induced by the constraint model. Concurrently, safe region expansion evaluates unobserved samples with respect to their closeness to the safe region boundary to improve the estimated safe region. Two sub-strategies are threaded under the multi-objective optimization (MOO) framework so that informativeness in both the target and the safety can be simultaneously considered. Our contributions in this paper are as follows.

1) We develop the safe variance reduction strategy to improve the predictive performance of the target model under the regime of implicit physics constraints.
2) The safe region expansion strategy is devised to expand the explorable safe region for further improvement of the target model, which concurrently dedicates to avoiding failures.
3) A new acquisition function is proposed that flexibly integrates two heterogeneous strategies.

This paper is organized as follows. In Section II, we review literature related to active learning applications in which physics involved and machine learning with implicit constraints. In Section III, we elucidate our active learning strategy considering implicit constraints to avoid failures. Section IV illustrates how the proposed strategy works under the regime of implicit constraints with the simulation study. The real-world application to predictive modeling of composite fuselage deformation considering structural failures is presented in Section V. Lastly, a summary of this paper is provided in Section VI.

II. Literature Review

In this section, we discuss related literature dichotomizing into (i) active learning for engineering systems; and (ii) sequential sampling with implicit constraints. While the literature of two topics may not be related, the integration of two approaches can be the cornerstone of our approach. In the application of active learning to engineering systems, we focus on that how physics in the engineering system influences active learning. In sequential sampling with implicit constraints, we do not restrict the constraints therein being related to physics, yet focus on the ways of considering constraints in sequential sampling, which includes active learning and Bayesian optimization.

A. Active Learning for Engineering Systems

In the engineering domain, active learning is vastly utilized for surrogate modeling of expensive-to-evaluate systems, and the limit-state estimation in structural reliability analysis. For surrogate modeling, active learning aims to sample design

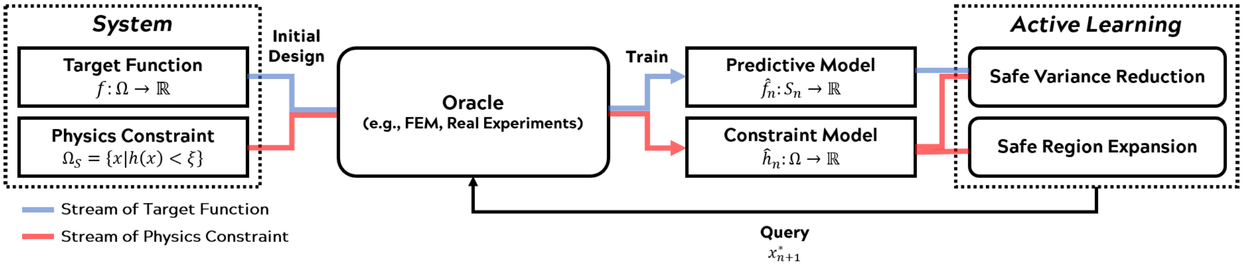
System
Target Function
$f:\Omega \to \mathbb{R}$
Physics Constraint
$\Omega_S = \{x|h(x) < \xi\}$
Initial Design
Oracle
(e.g., FEM, Real Experiments)
Train
Predictive Model
$\hat{f}_n: S_n \to \mathbb{R}$
Constraint Model
$\hat{h}_n: \Omega \to \mathbb{R}$
Active Learning
Safe Variance Reduction
Safe Region Expansion
Stream of Target Function
Stream of Physics Constraint
Query
x^*_{n+1}

Fig. 1. Overview of the proposed methodology.

points that can minimize the generalization error of the model. One of the most applied areas is computational physics that often involves surrogate models of systems of which physics-based models are costly or absent. For response surface method (RSM), Alaeddini et al. [12] proposed an active learning strategy adopting the variance reduction of Laplacian regularized parameters. Since RSM is often too restrictive to approximate complex phenomena, deep neural networks (DNNs) and GPs are substantially used for surrogate modeling of PDE-based models. For DNNs, Costabal et al. [13] proposed a physics-informed neural network for cardiac activation mapping. The neural network model is guided by the loss function that involves the physics equation, one of the most widely used approaches to infuse physics into machine learning. In their active learning, they chose the design point with the largest uncertainty. Lye et al. [14] proposed active learning for surrogate modeling of PDE solutions that chooses the next query point minimizing the cost function with the sequentially updated DNN model. In order to ensure the feasibility, they confined the explorable settings with the feasible region known a priori. Pestourie et al. [15] employed the ensemble neural network model for photonic-device model, and used quantified uncertainty of the network model for active learning. However, the aforementioned works do not consider implicit constraints, so they may induce failures if there exist failure conditions in the design space.

For GP models, Yang et al. [16], [17] proposed a physics-informed GP for the stochastic PDE simulator. The GP model is informed by replicated observations of the PDE simulator, and predictive variance is referred to for active learning. Chen et al. [18] developed the GP that incorporates linear and nonlinear PDE information. They involved the active learning strategy to determine PDE points for their GP model by employing the integrated mean-squared error (IMSE) criterion [2]. Yue et al. [4] proposed the variance-based and the Fisher information criteria for GPs considering uncertainty, and applied them to modeling of composite fuselage deformation. Likewise, existing active learning strategies for GPs in engineering applications are also unconstrained, so their approaches may lead to infeasible design points in constrained systems.

In structural reliability analysis, active learning confines its interest to the limit-state, which is a hyperplane of the input space, in order to estimate the probability of failure in the system. Generally, their acquisition functions are designed to give more weights on sampling from the vicinity of the limit-state. Echard et al. [11] proposed an acquisition function to estimate the limit-state with GP models, and suggested the framework that encompasses the limit-state estimation and the Monte-Carlo simulation for the conditional density estimation. Bichon et al. [10] proposed an acquisition function called expected feasibility function that quantifies the closeness of a design point to the limit-state considering quantified uncertainty with the GP. More explicitly, the vicinity of the limit-state was defined with predictive uncertainty of the GP for every candidate design point. Bect et al. [19] proposed the stepwise uncertainty reduction approach that employs the IMSE criterion associated with the estimated safe region. Wang et al. [20] proposed the maximum confidence enhancement method whose acquisition function aggregates the distance to the limit-state, input density, and the predictive uncertainty by multiplying them. The components are the same to [10], while the formulation of the acquisition function is different. Sadoughi et al. [21] proposed a dynamically adjustable acquisition function, which uses smoothly weighted closeness, unlike the aforementioned approaches. However, surrogate-based structural reliability analysis is specifically devised for the estimation of constraints, so it does not involve other non-failure processes such as the target process in our problem. Furthermore, it does not constrain sampling from the safe region as far as the point is close to the limit-state, since their objective is to estimate the limit state with simulations.

B. Sequential Sampling With Implicit Constraints

Constraints are critical in optimization problems, while the feasible set may be unknown a priori in practice. It is important to consider such implicit constraints in machine learning as well, since it is one way to incorporate background domain knowledge and make the model more consistent with the reality. There are mainly two distinct manners in considering implicit constraints in sequential sampling, and the most appropriate scenario to engineering problems is that any evaluation in the infeasible design space should be avoided. We refer to this type as safe exploration, since any design point violating the constraint invokes undesirable system failures. Schreiter et al. [22] proposed active learning for GPs whose motivation is closest to ours. They used the nuisance function of GP classification to discriminate the safe region and the unsafe region, and used the lower confidence interval to ensure safety during the maximum entropy based active learning. However, their approach only focuses on entropy of the target function, so it may not efficiently expand the explorable safe region. Consequently, it may lead to inactive learning due to insufficiently revealed safe region, and poor performance of the target model over the unexplored safe region. Furthermore, the nuisance function of GP classification needs some attention in its application. First, the nuisance function refers to binary labels, so it may distort the numerical information from the constraint observations that are usually related to closeness to the boundary of the safe region. Second, when the dataset contains only either safe or unsafe samples, it fails to provide promising discrimination. Turchetta et al. [23] suggested safe exploration for interactive machine learning that can be adopted for active learning. They conduct safe region expansion when the most informative data is not located in the current safe region. However, their safe region expansion aims to validate the safety of unobserved data, so it is inefficient to expand the explorable space to the true safe region which should be maximized to reduce the risk of the model.

Another way to address implicit constraints is that the feasibility can be disregarded in the middle of the trajectory in sequential sampling. This type is usually found in constrained Bayesian optimization, which is a sequential design strategy for global optimization of black-box functions with constraints. First of all, it should be noted that active learning and Bayesian optimization have different aims. Active learning focuses on summarizing unobserved samples to improve the model's quality over the input space, while Bayesian optimization aims to find the optimal location over the input space by evaluating samples directly associated with improvement in the objective function. Although they have different aims and derivations of their acquisition functions, constrained Bayesian optimization problems have some common features in consideration of constraints in their queries. Many constrained Bayesian optimization methods [24], [25], [26], [27], [28] follow the framework of Schonlau et al. [29] that multiplies the probability of feasibility to the expected improvement (EI) function. Gramacy and Lee [30] deemed infeasible samples also can be informative, so they proposed the integrated expected conditional improvement that is also weighted by expected feasibility. AlBahar et al. [31] proposed a physics-constrained Bayesian optimization with multi-layer deep GPs, and used it for the optimal actuators placement. Hernández et al. [32] proposed the predictive entropy search with constraints that refers to the expected entropy reduction at the minimum associated with observations from the objective function and constraints. It automatically focuses on objective and constraints by merging them into the integrated entropy. Basudhar et al. [33] used the probabilistic support vector classifier to discriminate the safe region. Then, the probability of constraint satisfaction is used to weigh the EI function as well. However, the multiplication of the feasibility probability only aids in leading their solutions to be feasible, so it is not much informative to expand the feasible region. Meanwhile, Sui et al. [34] proposed a constrained Bayesian optimization algorithm which devoted to both optimization and safe region expansion. They expanded the safe region in the first phase, and then implemented typical Bayesian optimization within the disclosed safe region. However, their safe expansion may be subject to slow convergence to the ground-truth unless the Lipschitz constants of constraint functions are known beforehand.

III. Methodology

An efficient safe exploration under implicit constraints can be accomplished when active learning devotes to both target approximation improvement and explorable region expansion. However, existing methods in Section II-B are lack of explorable region expansion, and indeliberate in avoidance of failure sampling. In this section, we propose our failure-averse active learning method for physics-constrained systems. We begin with specifying our problem whose constraints can be evaluated along with the target process. Then, we describe safe variance reduction and safe region expansion in detail, respectively. At last, they are combined in the integrated acquisition function under the MOO paradigm, and the practical implementation of overall algorithm is discussed.

A. Problem Statement and Gaussian Process Priors

Consider a system defined over a compact and connected design space $\mathbf{\Omega} \subseteq \mathbb{R}^D$. The system includes the target function $f : \mathbf{\Omega} \to \mathbb{R}$, we want to predict, and the constraint function $h : \mathbf{\Omega} \to \mathbb{R}$, related to the system failure and assumed to be independent of f. Let $\xi \in \mathbb{R}$ be a tolerable failure threshold associated with h, which should be defined conservatively considering intrinsic uncertainty of the process. For any design point $\mathbf{x} = [\, x_1 \, \cdots \, x_D \,]^\top \in \mathbf{\Omega}$, the system failure occurs when

$$h(\mathbf{x}) \geq \xi, \tag{1}$$

and safe otherwise. For example, f can be dimensional deformation of a solid structure given a force vector $\mathbf{x}$, and h can be the resulted von Mises stress in the structure. Both functions can be observed or evaluated with the costly real experiment and relevant physics-based models as

$$y = f(\mathbf{x}) + \epsilon_f, \quad z = h(\mathbf{x}) + \epsilon_h,$$

where $\epsilon_f \sim \mathcal{N}(0, \nu_f^2)$ and $\epsilon_h \sim \mathcal{N}(0, \nu_h^2)$ are observation noise. We assume that both f and h are continuous, and ν_f and ν_h are known, while they can be also estimated with GP priors on f and h to be described later with nugget effects. The safe region $\mathbf{\Omega}_S$, the subset of design space comprised of non-failure design settings, is unknown and difficult to obtain due to the prohibitive cost of the evaluation. We refer to the complementary of safe region as the failure region such that $\mathbf{\Omega}_F = \mathbf{\Omega} \setminus \mathbf{\Omega}_S$.

In our case, due to the high cost of evaluation of h and f, we prefer to sample the most informative set of design points that minimizes the generalization error associated with the target function. Suppose we have n samples from the system, denoted by $\mathcal{D}_n = \{(\mathbf{x}_i, y_i, z_i)\}_{i=1}^n$, and $\hat{f}_n$ and $\hat{h}_n$ are our predictors of f and h trained with $\mathcal{D}_n$, respectively. Then, the expected risk of $\hat{f}_n$ using L_p loss is

$$\mathcal{R}_f(\hat{f}_n) = \int_{\mathbf{\Omega}_S} L_p(f(\mathbf{x}), \ \hat{f}_n(\mathbf{x})) \, d\lambda(\mathbf{x}), \tag{2}$$

where $L_p(y, y') = |y - y'|^p$ is the loss function, and $\lambda(\mathbf{x})$ is a probability measure defined over $\mathbf{\Omega}$. In our problem, any violation of (1) may incur prohibitive cost of failure in the system, so any $\mathbf{x} \in \mathbf{\Omega}_F$ will not be considered for the system.

We assume that there exists a reproducing kernel Hilbert space for each of f and h, and they are bounded therein. It allows us to model both functions with GPs with corresponding kernels such that $k_f : \mathbf{\Omega}^2 \to \mathbb{R}$ and $k_h : \mathbf{\Omega}^2 \to \mathbb{R}$ [35]. In this paper, we consider the automatic relevance determination using radial basis function (RBF) kernel for $\mathbf{x}, \mathbf{x}' \in \Omega$ as

$$k_f(\mathbf{x}, \mathbf{x}') = \kappa_f^2 (\mathbf{x} - \mathbf{x}')^\top M_f^2 (\mathbf{x} - \mathbf{x}') + \nu_f^2 \delta(\mathbf{x}, \mathbf{x}'),$$

where κ_f is the nonnegative scale hyperparameter, M_f is the diagonal matrix of nonnegative length hyperparameters $\boldsymbol{\theta}_f = [\, \theta_{f,1}, \ldots, \theta_{f,D} \,]^\top$, and δ is the Kronecker delta function for the nugget effect. By defining k_h in the same manner, we can write f and h as

$$f(\mathbf{x}) \sim \mathcal{GP}\big(\mu_f(\mathbf{x}),\ k_f(\mathbf{x}, \mathbf{x}')\big),$$
$$h(\mathbf{x}) \sim \mathcal{GP}\big(\mu_h(\mathbf{x}),\ k_h(\mathbf{x}, \mathbf{x}')\big),$$

where μ_f and μ_h are mean functions, assumed to be zero without loss of generality.

Instead of employing a discriminative function for estimating the safe region, the GP regressor is more suitable for physics constraints since the output of h is numerically informative. More explicitly, as $h(\mathbf{x})$ is closer to the failure threshold, we may notice that $\mathbf{x}$ is closer to the safe boundary. Moreover, discriminative functions require observations from both safe and failure regions, while regressors are not subject to such imbalance or absence of one class. Therefore, we fit our GP regression model directly on observed outputs from h, and refer to the distance between the output and the failure threshold to infer the probability of safety.

Let us denote X_n as the $D \times n$ design matrix of $[\, \mathbf{x}_1 \cdots \mathbf{x}_n \,]$, and $\mathbf{y}_n$ and $\mathbf{z}_n$ as the vector of n observations from f and h, respectively. With GP priors on f and h, the hyperparameters $\Theta_f = \{\kappa_f, \boldsymbol{\theta}_f\}$ and $\Theta_h = \{\kappa_h, \boldsymbol{\theta}_h\}$ can be estimated by maximizing the log marginal likelihoods, which are

$$\ell(\mathbf{y}_n |\, X_n, \Theta_f) = -\frac{1}{2}\mathbf{y}^\top K_{f,n}^{-1}\mathbf{y} - \frac{1}{2}\log|K_{f,n}| - \frac{n}{2}\log 2\pi,$$
$$\ell(\mathbf{z}_n |\, X_n, \Theta_h) = -\frac{1}{2}\mathbf{z}^\top K_{h,n}^{-1}\mathbf{z} - \frac{1}{2}\log|K_{h,n}| - \frac{n}{2}\log 2\pi,$$

where $K_{f,n}$ and $K_{h,n}$ are covariance matrices comprised of every pair of $\mathbf{x}, \mathbf{x}' \in X_n$ given Θ_f and Θ_h, respectively. Once $\hat{f}_n$ and $\hat{h}_n$ are obtained with maximizing their log marginal likelihoods, the predictive mean and variance of $\hat{f}_n$ at an unobserved design point $\mathbf{x} \in \Omega$ can be derived as

$$\mathbb{E}[\hat{f}_n(\mathbf{x})] = \mathbf{k}_f(\mathbf{x}, X_n) K_{f,n}^{-1}\mathbf{y}_n,$$
$$\mathrm{Var}\big(\hat{f}_n(\mathbf{x})\big) = k_f(\mathbf{x}) - \mathbf{k}_f(\mathbf{x}, X_n) K_{f,n}^{-1}\mathbf{k}_f(\mathbf{x}, X_n)^\top,$$

and so does $\hat{h}_n$'s.

B. Safe Variance Reduction

Let us consider L_2 loss called the mean squared error (MSE), although it is not required in practice for our approach, and suppose f_* is an unbiased predictor of f with the minimum MSE (also called the best MSE predictor [2]) with respect to $\mathbf{\Omega}_S$ in the family of GP. Then, (2) can be decomposed as

$$\mathcal{R}_f(\hat{f}_n) = \mathcal{R}_f(f_*) + \int_{\Omega_S} \mathrm{Var}\big(\hat{f}_n(x)\big) d\lambda(\mathbf{x}), \tag{3}$$

which is the sum of the L_2 risk of f_* and the variance of $\hat{f}_n$. Since the L_2 risk of f_* is negligible due to its unbiasedness, (3) can be reduced by focusing on the variance reduction in $\hat{f}_n$. Let us denote the integrated variance of the predictor in (3) as

$$\mathbb{V}_{\Omega_S}(\hat{f}_n) = \int_{\Omega_S} \mathrm{Var}\big(\hat{f}_n(\mathbf{s})\big) d\lambda(\mathbf{s}),$$

where $\mathbf{s} \in \mathbf{\Omega}_S$. Then, the expected variance reduction over $\mathbf{\Omega}_S$ in $\hat{f}_n$ for an unobserved $\mathbf{x} \in \Omega$ is

$$\Delta\mathbb{V}_{\Omega_S}(\hat{f}_n|\mathbf{x}) = \mathbb{V}_{\Omega_S}(\hat{f}_n) - \int_{\Omega_S} \mathrm{Var}\big(\hat{f}_n(\mathbf{s}|\mathbf{x})\big) d\lambda(\mathbf{s}), \tag{4}$$
$$\mathrm{Var}\big(\hat{f}_n(\mathbf{s}|\mathbf{x})\big) = k_f(\mathbf{s}) - \mathbf{k}_f(\mathbf{s}, X_{n+1})^\top K_{f,n+1}^{-1}\mathbf{k}_f(\mathbf{s}, X_{n+1}), \tag{5}$$

where $X_{n+1} = [\, X_n\ \mathbf{x} \,]$, and $K_{f,n+1}$ is the covariance matrix of X_{n+1}. Eq. (4) is the IMSE criterion and always nonnegative (Proposition 1 and 2 in [36]), and it turns out that the IMSE criterion can be simplified by (5) as

$$\Delta\mathbb{V}_{\Omega_S}(\hat{f}_n|\mathbf{x}) = \int_{\Omega_S} \mathbf{k}_f(\mathbf{s}, X_{n+1})^\top K_{f,n+1}^{-1}\mathbf{k}_f(\mathbf{s}, X_{n+1}) d\lambda(\mathbf{s}), \tag{6}$$

which is to be maximized. For the GP, choosing the next point with the IMSE criterion is called active learning Cohn (ALC), and it is widely used along with active learning Mckay (ALM) which refers to the maximum entropy [37].

Unfortunately, (4) cannot be used directly, since the safe region is unknown a priori. Thus, we stick to our predictor $\hat{h}_n$ to estimate the safe region as

$$S_n = \{\mathbf{x} \in \Omega |\ \hat{\mu}_n^h(\mathbf{x}) + \beta_n \hat{\sigma}_n^h(\mathbf{x}) < \xi\}, \tag{7}$$

where $\hat{\mu}_n^h(\mathbf{x})$ and $\hat{\sigma}_n^h(\mathbf{x})$ are the mean and standard deviation of $\hat{h}_n(\mathbf{x})$, and $\Phi(\beta_n) = \Pr(\hat{h}_n(\mathbf{x}) < \xi) = 1 - \gamma_n$ for which $\gamma_n \in (0, 1)$. That is, β_n is related to the failure probability of $\mathbf{x} \in S_n$. By constraining our choice of next design point $\mathbf{x} \in S_n$, (4) can be written as

$$\Delta\mathbb{V}_{\Omega_S}(\hat{f}_n|\mathbf{x}) = \mathbb{V}_{\Omega_S \setminus S_n}(\hat{f}_n) + \Delta\mathbb{V}_{S_n}(\hat{f}_n|\mathbf{x}). \tag{8}$$

The first term of (8) indicates the irreducible variance induced by discrepancy between $\mathbf{\Omega}_S$ and S_n, while the second term is the reducible variance in the estimated safe region. It implies that a consequence of adopting S_n instead of $\mathbf{\Omega}_S$ with extremely low γ is underestimating the expected variance reduction of $\mathbf{x}$ over $\mathbf{\Omega}_S$. In order to extend the purview of variance reduction by $\mathbf{x}$ in (8), we may consider another safe region, called the progressive safe region, which has a more generous safety level than S_n as

$$S_n^+ = \{\mathbf{x} \in \Omega |\ \hat{\mu}_n^h(\mathbf{x}) + \beta_n^+ \hat{\sigma}_n^h(\mathbf{x}) < \xi\},$$

where $\Phi(\beta_n^+) = \Pr(\hat{h}_n(\mathbf{x}) < \xi) > 1 - \gamma_n^+$ of which $\gamma_n^+ < \gamma_n$. Straightforwardly, we have $S_n \subset S_n^+ \subseteq \Omega$. By considering S_n^+ as the reference set for the integrand (5), i.e., $\mathbf{s} \in S_n^+$, we can reduce the irreducible variance in the first term. Consequently, we have the following acquisition function

$$J_f(\mathbf{x}) = \Delta\mathbb{V}_{S_n^+}(\hat{f}_n|\mathbf{x}), \tag{9}$$

where $\mathbf{x} \in S_n$.

Eq. (8) shows that the choice of γ_n for S_n affects the learnability of active learning and the safety. More explicitly, S_n needs to be conservative to prevent failure by setting γ_n small enough, while too conservative setting of S_n will increase the irreducible variance term in (8) and reduce the explorable region. Therefore, a promising choice of γ_n should consider the capacity to afford failures, and the following proposition can be prescribed.

Proposition 1 (Failure Probability): For N-sampling budget and any $\zeta \in (0, 1)$, choosing design points $\mathbf{x}_i$'s from the safe region S_i for $i \in \{n+1, \ldots, n+N\}$ has the failure probability as

$$Pr\left(\bigcup_i \left(\hat{h}(\mathbf{x}_i) \geq \xi \,\middle|\, \mathbf{x}_i \in S_{i-1}\right)\right) \leq \zeta,$$

where S_i of which $\beta = \Phi^{-1}(1 - \zeta/N)$ for $\forall\ i$.

The proof is provided in supplementary material. For the progressive safe region, increasing γ_n^+ will reduce the unconsidered safe region $\mathbf{\Omega}_S \setminus S_n^+$, while it can also simultaneously increase the variance in $\mathbf{\Omega}_F \setminus S_n$, which is meaningless. Therefore, S_n^+ also need not to be defined too generous.

Even though J_f is designed carefully with appropriate safe regions, we can minimize the irreducible variance by minimizing the discrepancy between S_n and $\mathbf{\Omega}_S$. Generally speaking, J_f only focuses on reducing the variance of $\hat{f}_n$, and does not care about reducing the discrepancy. In order to efficiently expand the explorable region and improve the estimation accuracy of safe region, we need to incorporate the information from h as well as f in the information criterion. In the following section, we illustrate safe region expansion that focuses on the estimation of safe region boundary.

C. Safe Region Expansion

Safe region expansion is required to reduce the error induced by the mismatch of S_n and $\mathbf{\Omega}_S$, and to furnish higher confidence in exploration. It is metaphorically similar to that we can win when we know much more about the opposite. To expand the safe region without failure, we need to exploit the numerically informative output of h to approach the boundary of safe region from inside thereof, and expansion can be maximized when the design point is closest to the boundary [38]. Based on [39], we incorporate uncertainty of $\hat{h}_n$ and closeness to the boundary with the following criterion:

$$I(\mathbf{x}) = \begin{cases} \eta_n(\mathbf{x})^2 - (\hat{h}_n(\mathbf{x}) - \xi)^2 & \hat{h}_n(\mathbf{x}) \in (\xi - \eta_n(\mathbf{x}), \xi) \\ 0 & \text{Otherwise} \end{cases},$$

where $\eta_n(\mathbf{x}) = \alpha\hat{\sigma}_n^h(\mathbf{x})$ of which $\alpha > 0$. $I(\mathbf{x})$ attains its maximum when $\hat{h}_n(\mathbf{x}) = \xi$, which is the case of $\mathbf{x} \in \partial\mathbf{\Omega}_S$. Otherwise, it gets additional scores when $\hat{h}_n(\mathbf{x})$ does not exceed the threshold within an acceptable interval. The role of α is to magnify the effect of uncertainty in $I(\mathbf{x})$. Let the expected value of $I(\mathbf{x})$ with respect to $\hat{h}_n$ be $J_h(\mathbf{x})$, which is the acquisition function for safe region expansion, and expressed as

$$J_h(\mathbf{x}) = \eta_n(\mathbf{x})^2\left(\hat{\Phi}_n^h(\xi) - \hat{\Phi}_n^h(\xi - \eta_n(\mathbf{x}))\right) - \int_{\xi-\eta_n(\mathbf{x})}^{\xi} (h - \xi)^2\hat{\phi}_n^h(h)dh, \tag{10}$$

where $\hat{\Phi}_n^h$ and $\hat{\phi}_n^h$ are the CDF and the PDF of $\hat{h}_n(\mathbf{x})$. Eq. (10) is composed of two terms: the first term is related to uncertainty, and the second term is related to closeness to the boundary. Consequently, maximizing (10) leads to sampling near the boundary with high uncertainty if such points exist, and the most uncertain point otherwise.

Obviously, interests of J_f and J_h are inherently different, since they are associated with different mechanisms, f and h. Also, they are formulated for different purposes. It implies that the safe approximation of target function and the safe region expansion have trade-off, thus we need to compromise between both criteria to determine the most informative design point. We discuss the framework for addressing the balance between two criteria in the next section.

D. Harmonizing Acquisition Functions

In this section, we integrate two acquisition functions to optimize (maximize) them judiciously to achieve safe active learning. Two distinct acquisition functions are proposed to accomplish different objectives, and optimization of two criteria is a MOO problem. Conceptually, we may think of a point $\mathbf{x} \in S_n$ that achieves the maximum of each criterion simultaneously, called as the utopia point. However, MOO typically has no single optimal solution contrary to usual single-objective optimization. Therefore, the Pareto optimality concept is mostly referred to define the optimality in this regime. Let $\mathbf{J}(\mathbf{x}) = \left[\, J_f(\mathbf{x}) \; J_h(\mathbf{x}) \,\right]^\top$, then Pareto optimality and its weaker version are defined as follows.

Definition 1 (Pareto Optimal): A point, $\mathbf{x}_* \in \mathbf{\Omega}$, is Pareto optimal if and only if there does not exist another point, $\mathbf{x} \in \mathbf{\Omega}$, such that $\mathbf{J}(\mathbf{x}) \leq \mathbf{J}(\mathbf{x}_*)$, and $J_i(\mathbf{x}) < J_i(\mathbf{x}_*)$ for at least one of $i \in \{f,\ h\}$.

Definition 2 (Weakly Pareto Optimal): A point, $\mathbf{x}_* \in \mathbf{\Omega}$, is weakly Pareto optimal if and only if there does not exist another point, $\mathbf{x} \in \mathbf{\Omega}$, such that $\mathbf{J}(\mathbf{x}) < \mathbf{J}(\mathbf{x}_*)$.

Note that inequalities in the definitions associated with vectors stand for element-wise inequality. Obviously, every Pareto optimal point is weakly Pareto optimal, while the reverse is not true.

It is common to scalarize the vector-valued objective functions in MOO, and the formulation of problem is critical for Pareto optimality of the solution. In this paper, we use the weighted sum, which is widely used, to scalarize our criteria with the integrated acquisition function:

$$J(\mathbf{x}) = \left((1 - w)J_f(\mathbf{x})^p + wJ_h(\mathbf{x})^p\right)^{1/p}, \quad w \in [0, 1], \tag{11}$$

where $p \in \mathbb{N}$. The weight parameter w in (11) exactly conveys the preference of decision maker between two objectives. For example, if one is more interested in the safe region expansion, the decision maker will weigh more on J_h, and decrease w when the estimated safe region seems acceptable. Otherwise, some may begin with small w to see if the safe region expansion is necessary. It turns out that the integrated acquisition function guarantees the Pareto optimality of its solution given w as shown in the following proposition.

Proposition 2 (Sufficient Pareto Optimality, [40]): For any $w \in (0, 1)$, a solution that maximizes the integrated criterion is Pareto optimal associated with w. When $w = 0$ or $w = 1$, a solution of the integrated criterion is weakly Pareto optimal associated with w.

However, J_f and J_h may be different in their scales, so the weight parameter w cannot be determined straightforwardly. It is common to transform component objective functions in the formulation of MOO, so we may normalize each criterion as

$$\overline{J_i}(\mathbf{x}) = \frac{J_i(\mathbf{x}) - \min J_i}{\max J_i - \min J_i},$$

where minimum and maximum of J_i for $i \in \{f, h\}$ stand for the minimum and maximum over S_n. To normalize objective functions, we need their maxima and minima, so we provide two scaling options in this paper. The first scaling method is for global searching in a dense-grid over the design space. By discretizing the design space into a dense-grid, we may evaluate all criteria over the grid. It can provide the heuristically global optimal solution, and make scaling more consistent. Another scaling method is to incorporate the lower and upper bounds of criteria. We already have that both criteria are nonnegative, so they are lower bounded by zero. For the upper bounds, J_f is upper bounded by

$$\mathbb{V}_{S_n^+}(\hat{f}_n) = \int_{S_n^+} \mathrm{Var}\big(\hat{f}_n(\mathbf{s})\big) d\lambda(\mathbf{s}),$$

since the expected variance reduction in (9) is nonnegative. Meanwhile, J_h is upper bounded by

$$\sup_{\mathbf{x} \in S_n} \eta_n(\mathbf{x})^2 = \sup_{\mathbf{x} \in S_n} \alpha^2 \hat{\sigma}^h(\mathbf{x})^2$$

from its original formulation. In this way, we can scale both criteria by their tractable bounds that can be obtained prior to the evaluation of every candidate.

The upper bounds of J_f and J_h can be referred to the asymptotic convergence of the integrated acquisition function as described in Proposition 3, of which proof is given in supplementary material.

Proposition 3 (Asymptotic Convergence): Suppose a non-empty S_n. As $n \to \infty$, $S_n \to S_* \subseteq \Omega_S$, and also $J(\mathbf{x}|S_n) \to 0$, for every $\mathbf{x} \in S_n$, and any $w \in [0, 1]$.

Proposition 3 shows that the integrated criterion J leads our estimators to the best estimators of f and h over a conservative estimation of Ω_S, regardless of the choice of w.

Let us refer to our active learning as PhysCAL (**Phys**ics-**C**onstrained **A**ctive **L**earning), and its pseudocode is provided in Algorithm 1. In practice, integrals in (9) and (10) require numerical methods such as averaging integrand uniformly sampled within integration limits. The computational cost of the algorithm is mostly dominated by the inverse of $K_{f,n+1}$ in (5), which originally takes $\mathcal{O}(n^3)$. Thus, we adopt the rank one Cholesky update in [41] to alleviate the cost to $\mathcal{O}(n^2)$ and to improve the numerical stability. If active learning has no finite candidate pool, we may need a grid or uniform space-filling designs over Ω to realize S_n and S_n^+ and solve line 5 in the algorithm. Note that PhysCAL can be terminated by not only the sampling budget, but also the prediction accuracy of the target model when the sampling budget is implicit or early stopping is reasonable. In order to do so, a separated testing dataset or cross-validation is required.

Algorithm 1 Active Learning for Physics-Constrained Systems

```
1: **Prerequisite**: $N$(Sampling budget), $\mathcal{D}$, $\beta$, $\beta^+$, $\alpha$, $w$
2: Train $\hat{f}$, $\hat{h}$ with $\mathcal{D}$
3: **while** $N > 0$ **do**
4:     Evaluate $S_n$, $S_n^+$ over $\Omega$
5:     $\mathbf{x}_* = \arg\max_{\mathbf{x} \in S_n} \overline{J}(\mathbf{x})$
6:     Observe $y_*$, $z_*$ at $\mathbf{x}_*$
7:     $N = N - 1$
8:     $\mathcal{D} = \mathcal{D} \cup \{\mathbf{x}_*, y_*, z_*\}$
9:     Update $\hat{f}$, $\hat{h}$ with $\mathcal{D}$
10:    Update $\beta$, $\beta^+$, $\alpha, w$ (Optional)
11: **end while**
```

IV. Simulation Study

In this section, we apply our active learning to the approximation of a constrained 2-D simulation function. The response surface of the target function is defined over $\Omega = [-0.5, 0.5]^2$, which is

$$f(\mathbf{x}) = |x_0 x_1|,$$

where $\mathbf{x} = [\, x_0 \; x_1 \,]^\top \in \Omega$, and the constraint function is

$$h(\mathbf{x}) = (\cos(2\pi x_0) - \cos(2\pi x_1))^2 - 0.8 \exp(|x_0 x_1|),$$

which is assumed to be implicit. Both functions and the failure region are illustrated in Fig. 2a. We set the safe region as $\Omega_S = \{\mathbf{x} \in \Omega |\, h(\mathbf{x}) < 0.7\}$, thereby the failure region ratio to the design space being approximately 0.28. Assuming we have no prior knowledge of the safe design settings, 10 initial samples are obtained over the design space using the maximin LHD, which yields 2-3 samples from the failure region in 10 replications. Observations from f and h are corrupted by Gaussian noise, and additional 20 samples are obtained with active learning. Parameters of each method were fixed in this study, so we omit the subscription n in the parameters. For PhysCAL, we set $\gamma = 0.001/20$, which is less than the defective percentage of six sigma in statistical process control, and $\gamma^+ = 0.01$ to extend the considered region in safe variance reduction to the progressive safe region with the failure probability of 0.99. As the benchmark method, safe exploration for GP in [22] (referred as SEGP) is considered. For the other parameter settings, $\alpha = 2$ according to [39] and $w = \{0.0, 0.1, \ldots, 0.9, 1.0\}$ are used.

During the simulation, setting the safety level in SEGP as high as ours was impossible in some replications due to the insufficient number of failure samples to make a nontrivial explorable region, which was also mentioned in

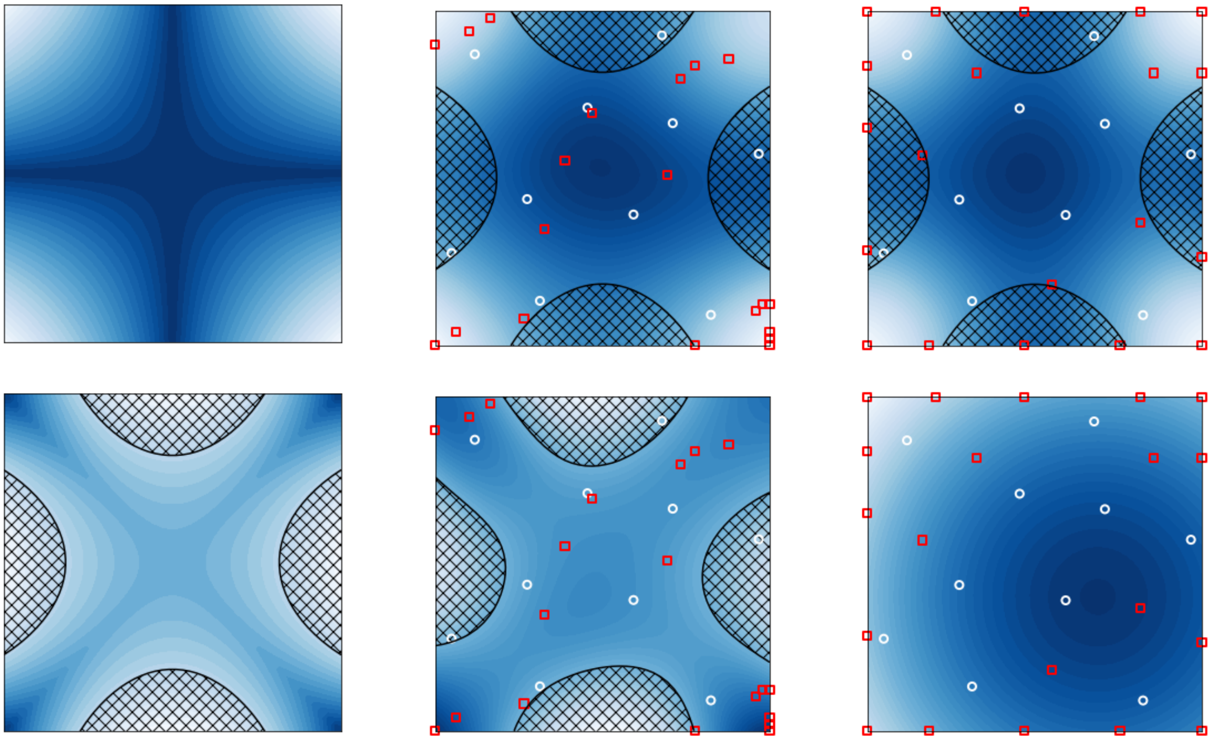

Fig. 2. Simulation Study Result (a) Ground truths of target (top) and constraint (bottom) functions. (b) PhysCAL's (proposed method) estimation of target (top) and constraint (bottom). (c) Benchmark method's (Schreiter et al. [22]) estimation of target (top) and constraint (bottom). Hatched regions in top figures are the true failure region, and the bottom of (b) is the estimated failure region. Note that the bottom of (c) has no estimated failure region. White circles are initial design points, and red squares are sampled with active learning.

their work. Furthermore, we observed that the benchmark method failed to estimate a meaningful failure region as shown at the bottom of Fig. 2c. These issues can be explained as follows. First, the dataset is imbalanced due to the larger safe region, so it resulted in unsatisfactory classification, which underestimates the failure region. Second, the nuisance function of SEGP is adapted for binary classification via the Laplace approximation [35], so the predictive variance induced by the nuisance function is inadequate for the safe region estimation as ours. More explicitly, encoding the failure process observation into a binary class discards numerical information from the original output (i.e., the response of the failure process). Consequently, the GP classifier determines the failure only based on the spatial input in disregarding of output's numerical information, thereby inducing improper predictive variance.

As a result, PhysCAL achieved the prediction error (MSE) of 0.0023 with 0.5 additional number of failures on average, and there were five zero-failure cases of ten replications. Meanwhile, the benchmark method achieved a better prediction accuracy with 0.001, while the averaged number of additional failures was 6.7 among 20 queries with no case of zero-failure. It is not surprising that the benchmark method did better in target process prediction, since exploration was unrestricted by underestimating the failure region. Fig. 2c shows that the classifier learned that the center region is safe (with dark blue), while it could not discriminate the failure region due to the aforementioned reasons. Consequently, the entropy-based strategy in SEGP led to evenly distributed sampling as their property [2]. Meanwhile, our method estimated the failure region much better and explored more safely as shown at the bottom of Fig. 2b. Hence, in the case that a single failure is very crucial, our approach will be more suitable.

Fig. 3 shows the performance of PhysCAL with different weight parameter settings. We can observe that a low weight parameter may improve the prediction accuracy by focusing more on variance reduction, while it does not necessarily make our approach safe due to lack of knowledge in failure region. Meanwhile, setting w too high also induced the increased number of failures and low predictive accuracy due to indifference to variance reduction in the target approximation. In this simulation, $w = 0.4$ was the best choice among considered values with promising predictive accuracy and the least number of failures.

V. Case Study

In this section, the proposed method is applied to predictive modeling of composite fuselage deformation in the aerospace manufacturing process. Composite materials such as carbon fiber reinforced polymers are extensively applied to various domains (including aerospace, automotive, construction and energy) due to their versatility, high strength-to-weight ratio, and corrosion-resistance. However, composite materials are

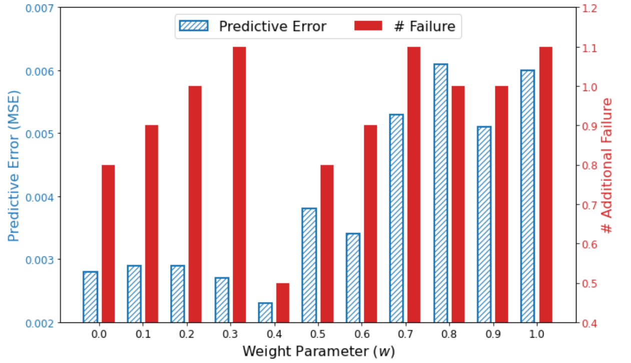

Fig. 3. Performance of PhysCAL associated with different weight parameters in the simulation study.

nonlinear and anisotropic due to their structural natures [42], so flexible models and adaptive design of experiments are required to predict their deformation precisely. Moreover, since they are also subject to structural failures, manufacturers should avoid unsafe load settings in the fabrication of composite structures.

Aerospace manufacturing considered in this case study is subject to dimensional deviations at the joint rim of composite fuselage sections due to the multi-batch manufacturing system. Therefore, to assemble fuselage sections, the shape control procedure is required to reshape them homogeneously. In the shape control procedure, the fuselage section is placed on the supporting fixture, and ten equispaced hydraulic actuators are introduced to reshape the fuselage as shown in the left figure of Fig. 4. A promising approach for the optimal shape control procedure is to employ a highly precise predictive model of composite fuselage's deformation instead of time-consuming physics-based model. However, the construction of such model validated with real experiment is very challenging due to the expensive cost of sampling, and the risk of structural failure because of unsafe loads. Especially, since composite failures in the shape control process may result in disposal of nonconforming parts or delayed delivery due to the restoration, up to date applications have restricted viable actuator forces in a conservative manner to prevent structural failures [43], [44]. It may lead to a suboptimal shape control, so our objective is to extend the feasible actuator forces that may include structural failure settings, thereby providing higher degrees of freedom to the shape control.

There are different types of failures in composite materials such as fiber, matrix, and ply failures. Likewise, a number of composite material failure criteria (e.g., Tsai-Wu, Tsai-hill, Hoffman, Hashin) are devised for different modes of failures [45]. In this paper, we considered the Tsai-Wu criterion, which is one of the most widely used interactive failure criterion. Note that it is possible to consider more than one criterion simultaneously by taking the most parsimonious criterion, or considering the intersection of safe regions defined by multiple constraint GP models. Briefly, Tsai-Wu criterion considers interactions between different stress components in addition to the principal stresses (in a homogeneous element). Using the principal material coordinate system on the cubic element of composite material, consider three directions: 1 is the fiber direction; and 2 and 3 are directions perpendicular to 1, respectively. Let σ_i^T and σ_i^C be the tensile failure stress and the compressive failure stress in $i \in \{1, 2, 3\}$ direction, and τ_{12}^F be the shear failure stress in the 12 plain. The Tsai-Wu criterion is defined as

$$\left(\frac{1}{\sigma_1^T}-\frac{1}{\sigma_1^C}\right)\sigma_1+\left(\frac{1}{\sigma_2^T}-\frac{1}{\sigma_2^C}\right)\sigma_2+\frac{\sigma_1^2}{\sigma_1^T\sigma_1^C}+\frac{\sigma_2^2}{\sigma_2^T\sigma_2^C}+\left(\frac{\tau_{12}}{\tau_{12}^F}\right)^2-\frac{\sigma_1\sigma_2}{\sigma_1^T\sigma_1^C\sigma_2^T\sigma_2^C}\geq 1,$$

where the left-hand side is the nonnegative criterion value, and the failure occurs when it exceeds one.

The Tsai-Wu criterion value induced by the shape adjustment solved by the FEM is shown in the right of Fig. 4. We can observe that failures are occurred at the bottom of the fuselage, since fixtures that sustain the fuselage are restricting its deformation. Not only limited to the Tsai-Wu criterion, physics constraints have many assumptions such as homogeneity, absence of higher-order interactions, etc., although they are convincing apparatuses to consider the structural reliability. Hence, they are typically utilized with the safety-of-margin (the reciprocal of the acceptable failure criterion) or UQ to prevent unexpected failures. Likewise, the safe shape control system should consider the failure criterion not only its value, but also the additional safety measures.

A. Experiment Settings

A well-calibrated FEM simulator of the procedure is referred to as our oracle considering the risk of real experiment. The target function's input is the vector of unidirectional forces (in lbf) of ten actuators, and the output is Y, Z-directional deformation (in microinch) of fuselage at one of 91 critical points around the rim. The maximum magnitude of actuator force is 1,000 lbf, which may cause failures in the structure (see Fig. 4). As our additional safety measure, the margin of safety with Tsai-Wu criterion is set at 1.25 (i.e., the acceptable criterion is 0.8).

For the initial design, the maximin LHD is used to generate 20 observations for both deformation and failure criterion, and additional 20 samples are queried by different methods: random, ALM, ALC, SEGP, and PhysCAL. We adopted the pool-based scenario in this case study by providing the 400 size of candidate pool that uniformly spreads out the design space. Considering the variability in the initial design, we have generated ten initial dataset independently, and replicated the experiment. It is noteworthy that we do not cease active learning, even though we encounter a failure in the construction of the predictive model for this case. In practice, the composite failure is definitely an undesirable event, while it is feasible as far as the design point is within the input space (i.e., the maximum actuator forces). Therefore, failure events are also included in the training dataset, and we compare the number of additional failures in learning to evaluate that how well each method avoids failures.

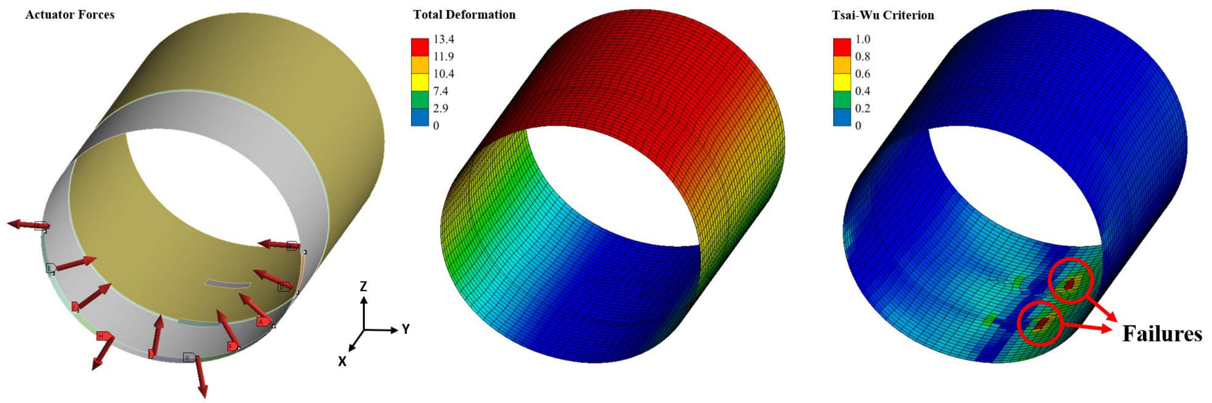

Fig. 4. Shape control of composite fuselage in the FEM. (left) Actuator input. (center) Resulted deformation. (right) Resulted Tsai-Wu criterion.

TABLE I
RESULT OF CASE STUDY

CASE	**Random**		**ALM**		**ALC**		**SEGP** [22]		**PhysCAL**	
	MAE	# Add. Fail	MAE	# Add. Fail	MAE	# Add. Fail	MAE	# Add. Fail	MAE	# Add. Fail
1	6.303	1	2.573	4	2.572	4	N/A		2.640	0
2	10.135	0	2.572	4	2.572	3	N/A		2.639	0
3	7.975	1	1.628	5	1.635	4	N/A		3.530	0
4	2.368	0	2.862	4	2.143	4	N/A		3.667	0
5	2.427	1	2.161	4	1.939	5	N/A		2.465	0
6	4.570	0	2.572	4	2.572	3	N/A		2.639	0
7	6.163	0	2.344	4	2.631	4	N/A		2.387	0
8	6.704	1	1.427	4	1.272	4	1.892	0	3.487	0
9	3.326	0	2.009	4	1.900	3	2.631	1	2.286	0
10	4.602	0	1.970	4	1.583	4	2.369	2	2.585	1
Mean	5.383	0.4	2.244	4.1	2.046	3.8	2.297	1.0	2.832	0.1
(Std.)	(2.345)	(0.5)	(0.470)	(0.3)	(0.441)	(0.6)	(0.305)	(0.8)	(0.518)	(0.3)

For PhysCAL, we also considered different weight parameters as the simulation study, and set other parameters as $\alpha = 2$, $\gamma = 0.001/20$, and $\gamma^+ = 0.1$. In SEGP, we reduced the safety level of which from the PhysCAL's until that SEGP induced a nonempty explorable space. For the model evaluation, we used 100 safe samples as the testing dataset that is independently generated with the candidate pool, and the mean absolute error (MAE) is used as the metric.

B. Result

The result is summarized in Table I. First, we can observe that PhysCAL outperforms other methods in the number of additional failures. It achieves nine zero-failures from ten cases. Also, we can observe that ALM, ALC, and SEGP incurred more failures than the random. The reason is that the design space is almost dominated by the safe region, while the failure region may be more interesting than elsewhere. Interestingly, SEGP is inapplicable in this case when the initial dataset does not contain failure samples, since the method uses the binary classifier. It implies that employing a regressor as the constraint model is more advantageous when the prior information has no failure.

In terms of prediction accuracy, PhysCAL performs much better than random sampling, and comparable to other active learning approaches considering the scale of metric. We can conjecture that other methods are able to observe from the failure region that may be informative, so their accuracy is the consequence of unsafe exploration. Furthermore, PhysCAL is more flexible than other methods, since we may update the weight parameter in PhysCAL during data acquisition to focus more on the variance reduction as well as other methods. Therefore, PhysCAL is more promising for this case considering the risk of failure in the system.

C. Weight Parameter

Different weight parameters are considered for PhysCAL in the case study, and the performance of different weight parameters is provided in Fig. 5. Likewise, only focusing in one acquisition function is not optimal in this case, and the performance is better with $w = 0.7 \sim 0.9$, which is higher than the simulation study. The possible reasons are as follows. First, the failure region in this case is much smaller than the safe region, thus quite aggressive exploration may be acceptable (i.e., increasing w). Second, the actuator force is positively correlated with both von Mises stress, which is linearly correlated with the Tsai-Wu criterion, and deformation [6]. Consequently, the uncertainty term of safe region expansion could be informative to model variance reduction.

D. Margin of Safety

In order to observe the effect of margin of safety, we increased the value from 1.25 to 1.5, which reduces the acceptable Tsai-Wu criterion to 0.66. Although we may consider margin of safety higher than 1.5, such high value is

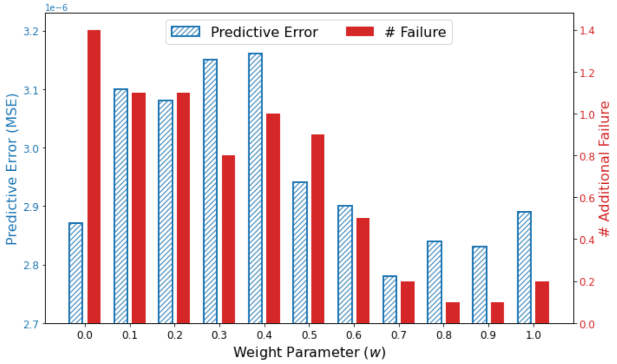

Fig. 5. Performance of PhysCAL associated with different weight parameters in the case study.

irrelevant in practice. Increasing the margin of safety yields the increment of failure region, so it results with the increased numbers of failures. In average of 10 replications, PhysCAL got 0.5 additional failures (six zero-failures), which is the least among considered methods, and achieved the predictive accuracy of 3.74 microinches. Random sampling got 0.6 additional failures, and achieved the predictive error of 5.16 microinches. Meanwhile, ALM and ALC achieved 8.3 and 7.9 failures, respectively. Likewise, SEGP was applicable only for the last three cases, and induced 3.3 failures from those cases.

VI. Summary

For physics-constrained systems that are expensive-to-evaluate, failure-averse active learning is proposed in this paper. In order to achieve safe active learning under the regime of implicit physics constraints, GP priors are imposed on the target function and physics constraints, and two acquisition functions are developed for safe variance reduction and safe region expansion. For safe variance reduction, two safe regions with different safety levels are employed in the IMSE criterion, thereby maximizing the safety and variance reduction over the underlying safe region. For the safe region expansion, the acquisition function is devised to sample near the safe region boundary considering uncertainty. Two acquisition functions are endowed with different objectives, so the MOO framework with Pareto optimality is applied to integrate them into the flexible global criterion. The integrated acquisition function is sufficient for the Pareto optimality of the design point to be queried, and can be flexibly adjusted by decision maker's preference considering the trade-off between two acquisition functions. Also, it is shown that the integrated acquisition function asymptotically leads to the best estimation of system.

In the simulation study, the proposed approach showed promising performance with the achievement of zero-failure, while the benchmark method failed to avoid failures in its learning process. Furthermore, with different parameters settings, we empirically observed that safe variance reduction and safe region expansion should be involved simultaneously for better predictive accuracy and higher safety. Our method has also shown remarkable performance in the predictive modeling of composite fuselage deformation considering its structural failure with Tsai-Wu criterion. It achieved zero-failure in most cases, while other benchmark methods induced more failures or inferior predictive accuracy. Our proposed method is adaptive, since it can incorporate domain knowledge and decision maker's preference with amenable parameters. Therefore, it is also applicable to other domains that are subject to implicit constraints.

VII. Code

The code for physics-constrained active learning can be found from https://github.com/cheolheil/ALIEN.

References

[1] B. Settles, *Active Learning* (Synthesis Lectures on Artificial Intelligence and Machine Learning Series). CA, USA: Morgan & Claypool, 2012.

[2] T. J. Santner, B. J. Williams, W. I. Notz, and B. J. Williams, *The Design and Analysis of Computer Experiments*, vol. 1. Berlin, Germany: Springer, 2003.

[3] C. Chen and G. X. Gu, "Generative deep neural networks for inverse materials design using backpropagation and active learning," *Adv. Sci.*, vol. 7, no. 5, Mar. 2020, Art. no. 1902607.

[4] X. Yue, Y. Wen, J. H. Hunt, and J. Shi, "Active learning for Gaussian process considering uncertainties with application to shape control of composite fuselage," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 1, pp. 36–46, Jan. 2021.

[5] X. Zhang, L. Wang, and J. D. Sørensen, "REIF: A novel active-learning function toward adaptive Kriging surrogate models for structural reliability analysis," *Rel. Eng. Syst. Saf.*, vol. 185, pp. 440–454, May 2019.

[6] Y. Wen, X. Yue, J. H. Hunt, and J. Shi, "Feasibility analysis of composite fuselage shape control via finite element analysis," *J. Manuf. Syst.*, vol. 46, pp. 272–281, Jan. 2018.

[7] S. Moaveni, *Finite Element Analysis Theory and Application With ANSYS, 3/e*. London, U.K.: Pearson Education India, 2011.

[8] J. Willard, X. Jia, S. Xu, M. Steinbach, and V. Kumar, "Integrating scientific knowledge with machine learning for engineering and environmental systems," 2020, *arXiv:2003.04919*.

[9] S.-K. Choi, R. Grandhi, and R. A. Canfield, *Reliability-Based Structural Design*. Berlin, Germany: Springer, 2006.

[10] B. J. Bichon, M. S. Eldred, L. P. Swiler, S. Mahadevan, and J. M. McFarland, "Efficient global reliability analysis for nonlinear implicit performance functions," *AIAA J.*, vol. 46, no. 10, pp. 2459–2468, Oct. 2008.

[11] B. Echard, N. Gayton, and M. Lemaire, "AK-MCS: An active learning reliability method combining Kriging and Monte Carlo simulation," *Struct. Saf.*, vol. 33, no. 2, pp. 145–154, Mar. 2011.

[12] A. Alaeddini, E. Craft, R. Meka, and S. Martinez, "Sequential Laplacian regularized V-optimal design of experiments for response surface modeling of expensive tests: An application in wind tunnel testing," *IISE Trans.*, vol. 51, no. 5, pp. 559–576, May 2019.

[13] F. S. Costabal, Y. Yang, P. Perdikaris, D. E. Hurtado, and E. Kuhl, "Physics-informed neural networks for cardiac activation mapping," *Frontiers Phys.*, vol. 8, p. 42, Feb. 2020.

[14] K. O. Lye, S. Mishra, D. Ray, and P. Chandrashekar, "Iterative surrogate model optimization (ISMO): An active learning algorithm for PDE constrained optimization with deep neural networks," *Comput. Methods Appl. Mech. Eng.*, vol. 374, Feb. 2021, Art. no. 113575.

[15] R. Pestourie, Y. Mroueh, T. V. Nguyen, P. Das, and S. G. Johnson, "Active learning of deep surrogates for PDEs: Application to metasurface design," *Npj Comput. Mater.*, vol. 6, no. 1, pp. 1–7, Dec. 2020.

[16] X. Yang, G. Tartakovsky, and A. Tartakovsky, "Physics-information-aided Kriging: Constructing covariance functions using stochastic simulation models," 2018, *arXiv:1809.03461*.

[17] X. Yang, D. Barajas-Solano, G. Tartakovsky, and A. M. Tartakovsky, "Physics-informed CoKriging: A Gaussian-process-regression-based multifidelity method for data-model convergence," *J. Comput. Phys.*, vol. 395, pp. 410–431, Oct. 2019.

[18] J. Chen, Z. Chen, C. Zhang, and C. F. J. Wu, "APIK: Active physics-informed Kriging model with partial differential equations," 2020, *arXiv:2012.11798*.

[19] J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez, "Sequential design of computer experiments for the estimation of a probability of failure," *Statist. Comput.*, vol. 22, no. 3, pp. 773–793, May 2012.

[20] Z. Q. Wang and P. F. Wang, "A maximum confidence enhancement based sequential sampling scheme for simulation-based design," *J. Mech. Des.*, vol. 136, no. 2, Feb. 2014, Art. no. 021006.

[21] M. Sadoughi, C. Hu, C. A. MacKenzie, A. T. Eshghi, and S. Lee, "Sequential exploration-exploitation with dynamic trade-off for efficient reliability analysis of complex engineered systems," *Structural Multidisciplinary Optim.*, vol. 57, no. 1, pp. 235–250, Jan. 2018.

[22] J. Schreiter, D. Nguyen-Tuong, M. Eberts, B. Bischoff, H. Markert, and M. Toussaint, "Safe exploration for active learning with Gaussian processes," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2015, pp. 133–149.

[23] M. Turchetta, F. Berkenkamp, and A. Krause, "Safe exploration for interactive machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2020, pp. 2868–2878.

[24] M. A. Gelbart, J. Snoek, and R. P. Adams, "Bayesian optimization with unknown constraints," 2014, *arXiv:1403.5607*.

[25] J. R. Gardner, M. J. Kusner, Z. E. Xu, K. Q. Weinberger, and J. P. Cunningham, "Bayesian optimization with inequality constraints," in *Proc. ICML*, 2014, pp. 937–945.

[26] R. Lam and K. Willcox, "Lookahead Bayesian optimization with inequality constraints," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[27] B. Letham, B. Karrer, G. Ottoni, and E. Bakshy, "Constrained Bayesian optimization with noisy experiments," *Bayesian Anal.*, vol. 14, no. 2, pp. 495–519, Jun. 2019.

[28] D. Eriksson and M. Poloczek, "Scalable constrained Bayesian optimization," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 730–738.

[29] M. Schonlau, W. J. Welch, and D. R. Jones, "Global versus local search in constrained optimization of computer models," *Lect. Notes-Monograph Ser.*, vol. 34, pp. 11–25, Jan. 1998.

[30] R. B. Gramacy and H. K. H. Lee, "Optimization under unknown constraints," in *Bayesian Statistics 9*, J. Bernardo et al., Eds., 2011, pp. 229–256.

[31] A. AlBahar, I. Kim, X. Wang, and X. Yue, "Physics-constrained Bayesian optimization for optimal actuators placement in composite structures assembly," *IEEE Trans. Autom. Sci. Eng.*, early access, Aug. 24, 2022, doi: 10.1109/TASE.2022.3200376.

[32] J. M. Hernández-Lobato, M. Gelbart, M. Hoffman, R. Adams, and Z. Ghahramani, "Predictive entropy search for Bayesian optimization with unknown constraints," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1699–1707.

[33] A. Basudhar, C. Dribusch, S. Lacaze, and S. Missoum, "Constrained efficient global optimization with support vector machines," *Struct. Multidisciplinary Optim.*, vol. 46, no. 2, pp. 201–221, Aug. 2012.

[34] Y. Sui et al., "Stagewise safe Bayesian optimization with Gaussian processes," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4781–4789.

[35] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[36] X. Chen and Q. Zhou, "Sequential design strategies for mean response surface metamodeling via stochastic Kriging with adaptive exploration and exploitation," *Eur. J. Oper. Res.*, vol. 262, no. 2, pp. 575–585, Oct. 2017.

[37] S. Seo, M. Wallat, T. Graepel, and K. Obermayer, "Gaussian process regression: Active data selection and test point rejection," in *Mustererkennung*. Berlin, Germany: Springer, 2000, pp. 27–34.

[38] R. Castro, R. Willett, and R. Nowak, "Faster rates in regression via active learning," in *Proc. NIPS*, vol. 18, 2005, pp. 179–186.

[39] P. Ranjan, D. Bingham, and G. Michailidis, "Sequential experiment design for contour estimation from complex computer codes," *Technometrics*, vol. 50, no. 4, pp. 527–541, Nov. 2008.

[40] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Struct. Multidisciplinary Optim.*, vol. 26, no. 6, pp. 369–395, Apr. 2004.

[41] C. Lee, K. Wang, J. Wu, W. Cai, and X. Yue, "Partitioned active learning for heterogeneous systems," 2021, *arXiv:2105.08547*.

[42] M. W. Hyer and S. R. White, *Stress Analysis of Fiber-Reinforced Composite Materials*. PA, USA: DEStech Publications, 2009.

[43] X. Yue, Y. Wen, J. H. Hunt, and J. Shi, "Surrogate model-based control considering uncertainties for composite fuselage assembly," *J. Manuf. Sci. Eng.*, vol. 140, no. 4, pp. 1–13, Apr. 2018.

[44] C. Lee, J. Wu, W. Wang, and X. Yue, "Neural network Gaussian process considering input uncertainty for composite structure assembly," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 3, pp. 1267–1277, Jun. 2022.

[45] A. C. Orifici, I. Herszberg, and R. S. Thomson, "Review of methodologies for composite material modelling incorporating failure," *Compos. Struct.*, vol. 86, nos. 1–3, pp. 194–210, 2008.

[46] P. Koepernik and F. Pfaff, "Consistency of Gaussian process regression in metric spaces," *J. Mach. Learn. Res.*, vol. 22, no. 244, pp. 1–27, 2021.

Cheolhei Lee received the B.E. degree in mechanical engineering from the Republic of Korea Naval Academy in 2011, and the M.S. degree in industrial engineering from the Virginia Tech, USA, in 2020, where he is currently pursuing the Ph.D. degree with the Grado Department of Industrial and Systems Engineering.

His research interests are data analytics, pattern recognition, and uncertainty quantification for system informatics and control. He was a Recipient of the ISA Scholarship Award, FTC Student Scholarship from ASQ, the IISE DAIS Best Student Paper Award, and the NAMRC/MSEC NSF Student Support Award.

Xing Wang received the B.S. degree in mathematics from Tsinghua University in 2013, the M.S. degree in statistics from the Georgia Institute of Technology in 2014, and the Ph.D. degree in risk management and insurance from the J. Mack Robinson College of Business, Georgia State University.

She served as an Instructor for the course risk modeling during the Ph.D. degree. She is currently an Assistant Professor with the Department of Mathematics, Illinois State University. Her research interests include risk measure, extreme value theory, and statistical inference. Her work focused on statistical inference about risk measures under extreme scenarios. She was a James C. Hickman Scholar. She was a recipient of the FTC Early Career Award and the ASA Early Career Travel Award.

Jianguo Wu (Member, IEEE) received the B.S. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2009, the M.S. degree in mechanical engineering from Purdue University in 2011, and the M.S. degree in statistics and the Ph.D. degree in industrial and systems engineering from the University of Wisconsin-Madison in 2014 and 2015, respectively.

He was an Assistant Professor at the Department of IMSE, UTEP, TX, USA, from 2015 to 2017. He is currently an Assistant Professor with the Department of Industrial Engineering and Management, Peking University, Beijing, China. His research interests focused on data-driven modeling, monitoring, and analysis of advanced manufacturing processes and complex systems for quality control and reliability improvement. He was a recipient of the STARS Award from the University of Texas Systems, and the Distinguished Young Scholars from China. He is a member of INFORMS, IISE, and SME.

Xiaowei Yue (Senior Member, IEEE) received the B.S. degree in mechanical engineering from the Beijing Institute of Technology, Beijing, China, in 2011, the M.S. degree in power engineering and engineering thermophysics from Tsinghua University, Beijing, in 2013, and the M.S. degree in statistics and the Ph.D. degree in industrial engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2016 and 2018, respectively.

He is currently an Assistant Professor with the Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, USA. His research interests focus on machine learning for advanced manufacturing. He is a Senior Member of ASQ and IISE and a member of ASME and SME. He was a recipient of the SME Outstanding Young Manufacturing Engineer Award, the IISE Manufacturing and Design Young Investigator Award, twelve best paper awards, and two best dissertation awards. He received the Grainger Frontiers of Engineering Grant Award from the National Academy of Engineering (NAE). He serves as an Associate Editor for the *Journal of Intelligent Manufacturing* and the *IISE Transactions*.