

Falsification of Multiple Requirements for Cyber-Physical Systems Using Online Generative Adversarial Networks and Multi-Armed Bandits

Jarkko Peltomäki and Ivan Porres
Faculty of Science and Engineering,
Åbo Akademi University
Turku, Finland
name.surname@abo.fi

Abstract—We consider the problem of falsifying safety requirements of Cyber-Physical Systems expressed in signal temporal logic (STL). This problem can be turned into an optimization problem via STL robustness functions. In this paper, our focus is in falsifying systems with multiple requirements. We propose to solve such conjunctive requirements using online generative adversarial networks (GANs) as test generators. Our main contribution is an algorithm which falsifies a conjunctive requirement $\varphi_1 \wedge \dots \wedge \varphi_n$ by using a GAN for each requirement φ_i separately. Using ideas from multi-armed bandit algorithms, our algorithm only trains a single GAN at every step, which saves resources. Our experiments indicate that, in addition to saving resources, this multi-armed bandit algorithm can falsify requirements with fewer number of executions on the system under test when compared to (i) an algorithm training a single GAN for the complete conjunctive requirement and (ii) an algorithm always training n GANs at each step.

I. INTRODUCTION

In this paper, we study the problem of black-box falsification of safety requirements of Cyber-Physical Systems (CPS). This is a validation method to increase the confidence that a system works as expected before it is taken into production. This is especially important with safety requirements of CPS, where a fault can lead to severe damage or injuries.

We focus on Cyber-Physical Systems where inputs and outputs are given as real-valued signals. We define the safety requirements of such systems as properties expressed in signal temporal logic (STL) [12]. An example of such requirement is

$$(\Box_{[0,30]} \text{RPM} < 3000) \rightarrow (\Box_{[0,4]} \text{SPEED} < 35).$$

Informally, this states that if the RPM signal is below 3000 during the first 30 units of time, then SPEED signal should be below 35 for the first 4 units of time. We test the system against these requirements by generating concrete input signals aiming to obtain an output signal that falsifies the given STL properties. By doing this, the input signals become a witness that the system under study does not fulfill the provided requirements.

Black-box falsification methods only consider inputs and observable outputs and do not require access to the internals

of the system under study nor to its design specifications or source code. Black-box methods can be used at any point of the development process and only require that we are able to evaluate inputs and output pairs by simulation or actual execution. An example of a simulation environment is MATLAB. It is a popular design and simulation toolset for Cyber-Physical Systems and it is widely used in academia and industry. Therefore many falsification tools, including the implementations of the algorithms presented here, can use it as a simulation environment.

The challenge in black-box safety validation of CPS lies in how to achieve the falsification of the proposed requirements with a limited testing budget. Corso et al. have recently published a survey of algorithms tackling this problem [3]. This survey presents four main strategies studied in the literature: optimization, path planning, reinforcement learning, and importance sampling.

Optimization-based falsification uses requirement robustness functions [2] to drive the search of falsifying inputs. Intuitively a robustness function indicates how close a given input is falsifying a requirement. A robustness function returns a value ≤ 0 for inputs that falsify the requirements, and it should yield progressively lower values when its inputs approach a falsification region. Given this, minimizing a robustness function will provide us with inputs that falsify the requirements, if they exist.

Different authors have proposed the use of existing optimization meta-heuristics such as genetic algorithms or simulated annealing to drive the search for falsification inputs [3] and also new algorithms that combine model learning with global and local search [8]. It is characteristic to these earlier algorithms that they have no knowledge of how the robustness functions are derived from requirements. More recently Zhang et al. [19] and Mathesen et al. [13] have proposed falsification algorithms that use information about the structure of the requirements to drive the falsification search. The work by Zhang et al. accepts requirements of the form $\Box_I(\varphi_1 \wedge \varphi_2)$ or $\Box_I(\varphi_1 \vee \varphi_2)$ while the more recent work of Mathesen can deal with more general requirements of the form $\varphi_1 \wedge \dots \wedge \varphi_n$ [13].

In this paper, we continue this line of research and propose an algorithm for the robustness-based falsification of CPS that uses the online GAN framework to generate falsifying inputs. We have used the online GAN framework in the past for the problem of performance testing [15]. As the first contribution of this article, we show here how the online GAN framework can also be applied for the falsification of safety requirements with a competitive performance.

The second contribution presented in this article is an extension of our base algorithm that uses specific knowledge on how to falsify a conjunctive requirement $\varphi_1 \wedge \dots \wedge \varphi_n$ efficiently. We first show how to solve this problem by training a GAN for each requirement φ_i separately as proposed in [13]. We offer experimental evidence that such an algorithm outperforms an algorithm training a single GAN for the complete conjunctive requirement. The drawback of such an approach is that it requires the simultaneous training of n different GANs. This problem is also present in [13] where the creation of n different Gaussian process models is required at each step.

Inspired by the multi-armed bandit problem, we address this issue by proposing a third algorithm which, after a warm up period, trains a single GAN at a time which saves resources. Our experiments indicate that, in addition to saving resources, it is possible that this variant falsifies requirements with fewer number of executions on the system under test.

We proceed as follows. In Section II, we introduce the problem of falsification of requirements of CPS formally and discuss related work in more detail. Section III presents our initial algorithm for a single conjunctive requirement falsification based on the online GAN framework. This algorithm is evaluated in Section IV using the Automatic Transmission Controller model, a benchmark problem presented in [5]. We continue in Section V by extending our original algorithm to deal with conjunctive requirements more efficiently and propose two new algorithms. The new algorithms are evaluated in Section VI using a synthetic problem proposed in [13]. Finally we present our concluding remarks in the last section.

II. FALSIFICATION OF FORMAL REQUIREMENTS OF CPS

A. Problem Description

We assume that the safety requirements of the system under test (SUT) can be expressed in signal temporal logic (STL) [12] as formulas $\varphi_1, \dots, \varphi_n$. Falsifying a formula φ means exhibiting a test such that the behavior (signal or trajectory) of the SUT while executing the test violates φ . In our case, falsifying the safety requirements amounts to falsifying the conjunctive requirement $\varphi_1 \wedge \dots \wedge \varphi_n$.

More precisely, we represent the SUT as a model \mathcal{M} which takes as its input a test t and outputs a (possibly vector-valued) signal $\mathcal{M}(t)$ for which the truth value of $\varphi_1 \wedge \dots \wedge \varphi_n$ can be evaluated. In order to search for a falsifying test t , we turn the problem into an optimization problem via robustness functions. As described, e.g., in [2], an STL formula φ can be effectively transformed into a real-valued robustness function ρ_φ such that φ evaluates to true for a signal s if and only if $\rho_\varphi(s) \geq 0$. Moreover, the robustness function has the

following stability property: small changes to a signal s with robustness $\rho_\varphi(s)$ of high absolute value do not affect the truth value of φ whereas small changes when $|\rho_\varphi(s)|$ is small could change it. This property turns the problem of falsifying φ to that of minimizing ρ_φ . In other words, our task is to solve the optimization problem

$$\arg \min_t \rho_\psi(\mathcal{M}(t)) \quad (1)$$

where $\psi = \varphi_1 \wedge \dots \wedge \varphi_n$. The elementary properties of robustness functions allows us to write (1) as

$$\arg \min_t \min_{i=1, \dots, n} \rho_{\varphi_i}(\mathcal{M}(t)). \quad (2)$$

We remark that the success of this plan depends on how complicated the formulas φ_i are. It is reasonable to assume, for example, that the function predicates appearing in the formulas refer to locally Lipschitz continuous functions. Observe also that continuous signals need to be discretized appropriately.

As pointed out in [13], it is often the case that evaluating $\mathcal{M}(t)$ is slow, but computing $\rho_\varphi(s)$ is fast. Therefore it makes sense to assume that we have all values $\rho_{\varphi_i}(\mathcal{M}(t))$, $i = 1, \dots, n$, available as soon as a test t has been executed on the SUT. With this assumption, we gain knowledge as we can use all values $\rho_{\varphi_i}(\mathcal{M}(t))$, $i = 1, \dots, n$, instead of just observing their minimum. The assumption also helps with the scale problem [13], [19]. Indeed, the value of ρ_{φ_i} could have wildly different scale than ρ_{φ_j} , so ρ_{φ_i} could effectively mask any information in ρ_{φ_j} if we only get to observe the minimum. It is worth remarking that this problem can persist even after appropriate scaling.

B. Previous Work

Solving (2) can be approached in different ways. Common optimization methods, like the cross-entropy method [17], have been used; see the introduction of [13] for more references to prior works.

We are interested in the recent paper [13] of Mathesen et al. where Bayesian optimization is used. The main points of their minBO algorithm are as follows. Let us write $\rho_i(t)$ for $\rho_{\varphi_i}(\mathcal{M}(t))$ for brevity. First sample tests randomly and execute them on the SUT to obtain a training data

$$(t_j, \rho_1(t_j), \dots, \rho_n(t_j)), \quad j = 1, \dots, N,$$

and best test t^* with

$$t^* = \arg \min_{t_j, j=1, \dots, N} \min_{i=1, \dots, n} \rho_i(t_j).$$

Then a Gaussian Process [16] is fitted for each ρ_i using the above training data. This yields an approximate model for each ρ_i . These models are used to figure out a test t_{N+1} which is likely to give smaller robustness values than t^* . The selection of t_{N+1} is done by maximizing expected improvement as is common in Bayesian optimization [1]. More precisely: a candidate test t'_i with highest expected improvement EI_i (with respect to t^*) is selected separately for each ρ_i , and the final candidate t_{N+1} is chosen to be t'_k

with $k = \arg \max_{i=1, \dots, n} \text{EI}_i$. The test t_{N+1} is executed on the SUT and the results are added to the training data. The best test t^* is updated if needed, and the above is repeated until the execution budget is exhausted.

In order to evaluate the minBO algorithm described above, two experiments are proposed in [13]. The first experiment is artificial and concerns properties of predefined but complicated and nonlinear functions. The second experiment concerns two industry benchmark models: the automatic transmission (AT) model of [5] and the ground collision avoidance system (GCAS) autopilot model for the F-16 fighter jet [9]. The findings of [13] are that the minBO algorithm performs always at least as well as a similar Bayesian optimization approach which has only access to the minimum $\min_i \rho_i(t)$. The minBO algorithm performs statistically significantly better in cases where the scale problem is present (both use cases in the first experiment and GCAS in the second experiment). The minBO algorithm is proposed as a performant solution to the scale problem.

III. THE ONLINE GAN ALGORITHM FOR ROBUSTNESS-BASED FALSIFICATION

In this section, we present an algorithm which solves (1) using online GANs. We falsify a single requirement meaning we assume that we can only observe the minimum robustness of multiple requirements.

A. Online GAN Training

At the heart of the algorithm is the idea of using a GAN for optimization. The idea is the same as in [15] where an online GAN is used for maximization. The idea is the following. Let φ be an STL formula, and suppose that the test robustness function $t \mapsto \rho_\varphi(\mathcal{M}(t))$ takes values in $[0, 1]$. We deem φ to be falsified if there exists a test t such that $\rho_\varphi(\mathcal{M}(t)) = 0$.

In addition to the SUT, we have two components: a generator G and a discriminator D . Both G and D are machine learning models, and in this paper they are neural networks. The generator G is a mapping from a latent space to the space of tests, and the aim is to train G in such a way that when the latent space is sampled uniformly, we obtain, via the map G , tests t for which the robustness $\rho_\varphi(\mathcal{M}(t))$ is low. The discriminator D simply learns the map $t \mapsto \rho_\varphi(\mathcal{M}(t))$.

Initially a random search is performed to obtain training data for D . Then we find new tests and train G as follows. We generate candidate tests with G and estimate their robustness with D (this avoids executing tests on the SUT). Whenever a test with high estimated robustness is found, we execute it on the SUT, add the test and its robustness to the training data, and train D using this updated training data. We then sample random points x_1, \dots, x_N from the latent space and form the artificial training data

$$(x_i, 0), \quad i = 1, \dots, N.$$

Using this training data, we train the composite model $D \circ G$ with the model parameters of D frozen. Since 0 is the smallest

value the robustness function can attain, this encourages the parameters of G to change in such a way that it generates tests which D estimates to have low robustness. As more and more data is collected, D should become more accurate and G should thus generate tests with low robustness. Ideally G generates a test with robustness 0 provided that φ is falsifiable.

Notice that our approach does not follow the traditional GAN training [6] where the discriminator is trained to distinguish between fake and real samples. The traditional approach is not possible here as we need to find the tests online.

Our online GAN approach can be seen as an instance of the idea of studying a SUT via a surrogate model which is refined over time as more information is available. The minBO algorithm of [13] fits into the same framework: their surrogate model is a Gaussian process instead of an online GAN.

B. Online-GAN Algorithm for a Single Conjunctive Requirement

Here we present an online GAN algorithm which attempts to falsify a single STL formula φ with robustness function ρ defined by $\rho(t) = \rho_\varphi(\mathcal{M}(t))$. In our context of falsifying the safety requirements of a SUT, the formula can be thought to be $\varphi_1 \wedge \dots \wedge \varphi_n$, that is, we can only observe the minimum robustness. The algorithm is presented in detail in Algorithm 1. We remark that finding the robustness function ρ is straightforward and implementations can be found in [14], [7].

Algorithm 1: Online GAN test generation algorithm for a single requirement.

```

1 T := Latin hypercube sampling(initial sample size);
2 GAN := new model with generator GN and
  discriminator DN;
3 repeat
4   train(GAN, T);
5   target := 0;
6   repeat
7     target := target + Δ;
8     t := generate(GN);
9     until predict(DN, t) ≤ target;
10  outcome := robustness(t);
11  T := T ∪ {(t, outcome)};
12 until outcome ≤ 0 ∨ |T| = budget;
13 result test suite T

```

The algorithm simply does what is described on a higher level in Subsection III-A. The first task is to perform a random search using a small part of our testing budget. In this algorithm we propose to use Latin hypercube sampling [10] for this purpose. After that, the algorithm proceeds with the search driven by the GAN. The generator is used to find a test with high estimated fitness in the inner loop. Initially we set the target estimated fitness to be 0 (falsified) but, as it is not necessarily possible to generate such a test (especially just after the algorithm has started), we raise the target on each

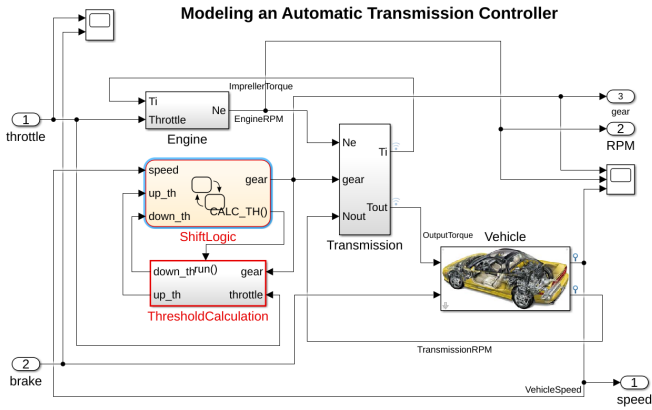


Fig. 1. Overview of the MATLAB model for the Automatic Transmission Controller (AT) problem. Model source: [5].

execution of the loop until a candidate test is found. As more training data is available, we should be able to achieve lower and lower targets.

IV. EXPERIMENT 1: AUTOMATIC TRANSMISSION (AT) MODEL

In this section, we evaluate Algorithm 1 on the automatic transmission (AT) model which is standard benchmark model proposed in [11]. The model performs automatic gear selection for a car when two input signals throttle and brake are provided to it. The model outputs two signals: engine speed (in RPM) and vehicle speed (in mph).

We are interested falsifying the requirement $\varphi_1 \wedge \varphi_2 \wedge \varphi_3$ where

$$\begin{aligned}\varphi_1 &= (\square_{[0,30]} \text{RPM} < 3000) \rightarrow (\square_{[0,4]} \text{SPEED} < 35), \\ \varphi_2 &= (\square_{[0,30]} \text{RPM} < 3000) \rightarrow (\square_{[0,8]} \text{SPEED} < 50), \\ \varphi_3 &= (\square_{[0,30]} \text{RPM} < 3000) \rightarrow (\square_{[0,20]} \text{SPEED} < 65).\end{aligned}$$

These requirements are given, e.g., in the ARCH workshop 2021 competition [4]. They require that during the first 30 time units (which is the complete duration of the signals) the initial vehicle speeds should take values 35, 50, and 65 provided that the engine speed is below 3000 RPM during the whole execution. The same falsification problem is considered in [13].

As in [4], we assume that $0 \leq \text{THROTTLE} \leq 100$ and $0 \leq \text{BRAKE} \leq 325$ during the whole execution (both signals can be positive simultaneously). Our input signals are piecewise constant functions with 6 pieces meaning that each input is constant for 5 time units at a time. We discretize the signals by sampling every 0.2 time units. We represent our tests as vectors in \mathbb{R}^{12} whose components satisfy the preceding requirements.

A general way to find a robustness function for STL formulas described in [2]. However such a direct approach is problematic as RPM and SPEED are measured on very

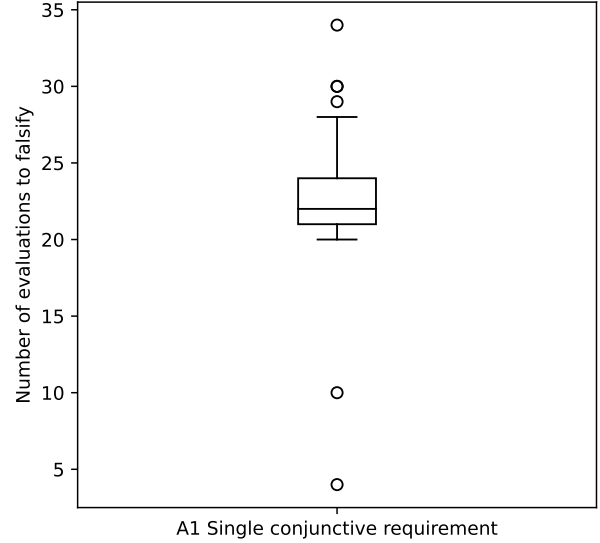


Fig. 2. Experiment 1: Box plot for number of executions needed for falsification.

different scales. We use the following ad hoc robustness function ρ_1 for φ_1 :

$$\rho_1(\text{RPM}, \text{SPEED}) = \begin{cases} \frac{1}{2}(35 - M_{\text{SPEED}})/35, & \text{if } M_{\text{RPM}} < 3000 \\ M_{\text{RPM}}/3000 - \frac{1}{2}, & \text{otherwise} \end{cases},$$

where

$$\begin{aligned}M_{\text{RPM}} &= \sup_{t \in [0,30]} \text{RPM}(t) \text{ and} \\ M_{\text{SPEED}} &= \sup_{t \in [0,4]} \text{SPEED}(t).\end{aligned}$$

We define robustness functions ρ_2 and ρ_3 analogously for φ_2 and φ_3 . It is readily checked that φ_i is falsified if and only if there exists a signal (RPM, SPEED) such that $\rho_i(\text{RPM}, \text{SPEED}) < 0$.

We attempt to falsify the requirement $\varphi_1 \wedge \varphi_2 \wedge \varphi_3$ using Algorithm 1. Due to the stochastic nature of the algorithm, we repeat the falsification task 50 times. We allow 80 executions on the SUT and use 25% of the execution budget for random search (Latin hypercube sampling). We have implemented Algorithm 1 in Python using TensorFlow for GAN implementation. The SUT is implemented in MATLAB as a Simulink model available at the ARCH verification competition repository.¹ The model is called via the MATLAB Python engine.

The box plot for the number of executions on the SUT required for falsification is found in Figure 2. Algorithm 1 succeeded in falsification in each replication with mean 22.8 (SD 4.4) executions required for a successful falsification. Since 20 executions were allowed for random search, we

¹<https://gitlab.com/goranf/ARCH-COMP>

conclude that typically the random search was unable to falsify the requirement but the GAN could with just a few extra executions. This shows that Algorithm 1 is capable for falsification as intended.

Interestingly, the minBO algorithm of [13] needed on average 68.7 executions for falsification. We explain the difference as follows. During research, we observed that the AT requirements are easily falsified for constant input signals. When the GAN is trained for the first time, it is typical that the components of its output are close to the middles of the allowed ranges. The first test proposed by the GAN is thus approximately constant and has high chance succeeding in falsification. We take this as an indication that the tests proposed by the minBO algorithm are not approximately constant.

We should note that we have not reproduced the results presented in [13] ourselves. Instead, we used the figures of the article, and therefore we cannot be certain that all the algorithms are evaluated under the same conditions.

V. ONLINE-GAN ALGORITHMS FOR MULTIPLE REQUIREMENTS

In this section, we present two Algorithms 2 and 3 which better take into account the information on multiple requirements.

A. Multiple Requirements

In Algorithm 2, we present an online GAN algorithm which attempts to falsify multiple requirements $\varphi_1, \dots, \varphi_n$ with corresponding robustness functions ρ_1, \dots, ρ_n .

Algorithm 2: Online GAN test generation algorithm for multiple requirements.

```

1 T := Latin hypercube sampling(initial sample size);
2 GAN := nproperties new models with generators GN
  and discriminators DN;
3 repeat
4   for i ∈ [1..nproperties] do
5     | train(GAN, i, T);
6   end
7   target := 0;
8   repeat
9     | target := target + Δ;
10    | for i ∈ [1..nproperties] do
11      | t[i] := generate(GN[i]);
12      | p[i] := predict(DN[i], t[i]);
13    | end
14    | best := argmin(p);
15  until p[best] ≤ target;
16  outcome := robustness(t[best]);
17  best = min(outcome);
18  T := T ∪ {(t, outcome)};
19 until best ≤ 0 ∨ |T| = budget;
20 result test suite T

```

The difference to Algorithm 1 is that we train n online GANs, one for each requirement φ_i . When we search for a candidate test, we consult the generators of each GAN and select the test which the corresponding discriminator estimates to have the lowest robustness. Again, we raise the target robustness until a candidate test is found.

B. Multiple Requirements with Property Selection

Next we introduce a new idea which addresses an obvious problem with Algorithm 2: the requirements $\varphi_1, \dots, \varphi_n$ are not equal, so they should not be given the same amount of consideration. Indeed, if there is a falsifiable requirement, then there is a requirement that is easiest to falsify and we should focus only on it. This saves both generation time and executions on the SUT. The problem is, of course, that we do not know which requirement (if any) is the easiest. Moreover, we are studying the requirements indirectly via the online GANs $\mathcal{G}_1, \dots, \mathcal{G}_n$ which act as surrogate models for the requirements and change over time. Our problem is thus similar to the nonstationary multi-armed bandit (MAB) problem [18]: how to explore all the surrogate models \mathcal{G}_i and how to exploit the ones we deem to be the best?

We propose the following simple approach for solving the problem. Whenever a candidate test is executed on the SUT, we record which surrogate model \mathcal{G}_i achieved the lowest robustness. Using these records we keep track of success frequencies p_1, \dots, p_n for each surrogate model. When a new candidate test needs to be generated, we proceed as follows: we pick a random surrogate model \mathcal{G} according to the frequencies p_1, \dots, p_n and use \mathcal{G} to generate a new candidate test. In an initial warm-up period, we consult all surrogate models in order to obtain good initial estimates for the success frequencies.

The described strategy clearly satisfies the requirements of exploration and exploitation: the historically most successful surrogate model is most likely being selected again, but there is a chance that another is selected, which perhaps leads to favoring an easily falsifiable requirement which initially looks unpromising. The success of this approach obviously depends on the nature of the requirements $\varphi_1, \dots, \varphi_n$ and the quality of the initial random search. We leave it as an open problem to develop a better strategy which would take into account the available information better. Recall that the minBO algorithm of [13] uses expected improvement to detect tests which likely have low robustness.

The variant of Algorithm 2 with the above MAB-inspired approach is described in detail in Algorithm 3. On line 3, we pick a single GAN to be used for candidate generation based on the success frequencies. On line 13, we record which GAN achieved the lowest robustness when the selected test was executed on the SUT. Notice that this GAN might be different from the GAN used for candidate test generation.

VI. EXPERIMENT 2: MO3D

In this section, we evaluate the three algorithms on a synthetic problem proposed in [13]. It concerns the nonlinear

Algorithm 3: Online GAN test generation algorithm for multiple requirements with property selection.

```

1 Apply Algorithm 2 for N% of the total budget.
2 repeat
3   chosen := pick one from [1..nproperties] with
   weights winner[1..nproperties] ;
4   train(GAN, chosen, T);
5   target := 0;
6   repeat
7     target := target + Δ;
8     t := generate(GN);
9     p := predict(DN, t);
10  until p ≤ target;
11  outcome := robustness(execute(t));
12  best, best_index = min(outcome), argmin(outcome);
13  winner[best_index] := winner[best_index] + 1;
14  T := T ∪ {(t, outcome)};
15 until best ≤ 0 ∨ |T| = budget;
16 result test suite T

```

TABLE I
MAIN STATISTICS OF EXPERIMENT 2.

	A1	A2	A3
Falsifications after 50 repetitions	16	39	46
% of falsifications	32%	78%	92%
Median observed minimum	2.84	-1.36	-2.01
Mean observed minimum	3.39	-1.00	-1.67
SD observed minimum	4.59	1.72	1.25

function $\text{mo3d}: \mathbb{R}^3 \rightarrow \mathbb{R}^3$, $\text{mo3d}(x) = (h_1(x), h_2(x), h_3(x))$ where

$$h_1(x_1, x_2, x_3) = 305 - 100 \sum_{i=1}^3 \sin\left(\frac{x_i}{3}\right),$$

$$h_2(x_1, x_2, x_3) = 230 - 75 \sum_{i=1}^3 \cos\left(\frac{x_i}{2.5} + 15\right), \text{ and}$$

$$h_3(x_1, x_2, x_3) = \sum_{i=1}^3 (x_i - 7)^2 - \sum_{i=1}^3 \cos\left(\frac{x_i - 7}{2.75}\right).$$

This function achieves its componentwise minimum value at $x^* = (7, 7, 7)$ with $f(x^*) = h_3(x^*) = -3$. The other two components achieve their minimum value of 5 at the points $3\pi/2(1, 1, 1)$ and $-37.5(1, 1, 1)$ respectively.

We are interested in requiring that all elements of $f(x)$ should be greater than 0 in the input domain $[-15, 15]^3$. This can be represented in STL as

$$(\Box h_1(x) > 0) \wedge (\Box h_2(x) > 0) \wedge (\Box h_3(x) > 0).$$

We simply use the functions h_1 , h_2 , and h_3 themselves as robustness functions. We repeat the falsification task 50 times for each algorithm. We allow 80 executions on the SUT and use 25% of the execution budget for random search.

The results written in Table I show that Algorithm 1 succeeds in falsifying the given requirement only in 16 of the

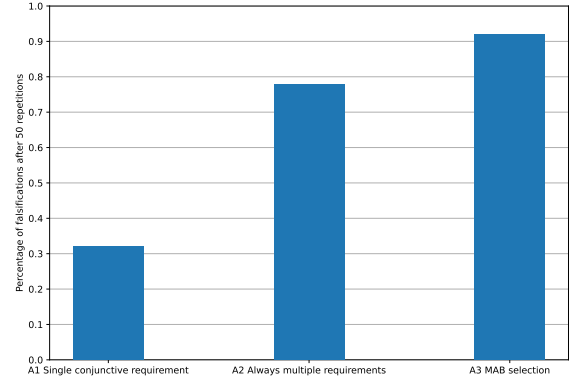


Fig. 3. Experiment 2: Percentage of successful falsifications.

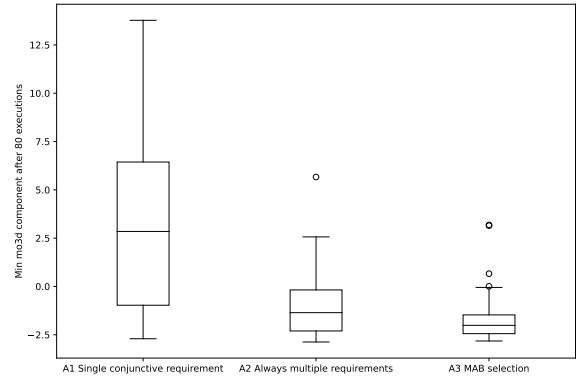


Fig. 4. Experiment 2: Box plots for minimum mo3d component over 50 experiments.

50 repetitions while Algorithms 2 and 3 succeed 39 and 46 times respectively. These correspond to 32%, 78%, and 92% of the repetitions. The success ratios are represented visually in Figure 3. The two proportion Z-test reports that the differences in falsification success rates between Algorithms 2 and 3 are statistically significant albeit with an observed p -value of 0.05.

Figure 4 shows box plots for the minima found by each algorithm after 80 executions on the SUT. It clearly shows again that Algorithms 2 and 3 perform better than Algorithm 1. While Algorithm 1 produces a median greater than 0, Algorithms 2 and 3 exhibit a median smaller than 0. When comparing Algorithms 2 and 3, the first one yields a median of -1.36 while the later yields a median of -2.01 . The Wilcoxon signed-rank test reports a p -value of 0.02 under the null hypothesis that the median of differences is 0.

Finally, Figure 5 shows the evolution of the mean observed minimum over the sequence of function evaluations. We observe how the mean for the minimum reported by Algorithm 3 becomes smaller than 0 with less evaluations than the other

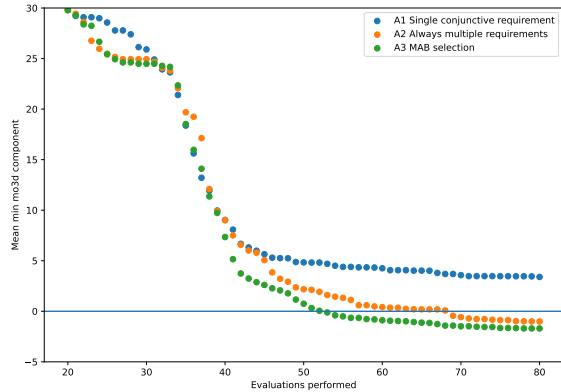


Fig. 5. Experiment 2: Evolution of minimum mo3d component over number of function evaluations, mean of 50 experiments.

two algorithms.

We have thus empirically confirmed that observing the robustness of each requirement separately can lead to significant improvement in falsification rate and the number of executions needed for falsification. Moreover the multi-armed bandit approach of focusing only on the most promising requirement can bring further improvements in falsification in addition to saving significantly on computational resources.

VII. CONCLUSIONS

In this paper we present a series of algorithms for robustness-based falsification of cyber-physical systems based on the online GAN framework.

The online GAN framework for requirement falsification is a novel concept that apparently has not been proposed before in the research literature. Both our approach and the Bayesian optimization approach of [3] combine surrogate model creation and test generation in an iterative loop. One difference is that Bayesian optimization methods use a Gaussian process as a surrogate model while we use a neural network. However we consider that the main difference is that Bayesian optimization methods must query the Gaussian process to determine the next input to evaluate while our approach uses a generative neural network that serves as a model of the space of falsifying inputs. The surrogate model and the generator model are trained together as in a GAN. Our early results indicate that Algorithm 1, based on the online GAN framework, exhibits a competitive performance when compared to the results published in [13].

We also present an extension of our basic algorithm to process conjunctive requirements more efficiently. The approach adopted in Algorithm 2 is based on the solution provided in [13]. Algorithm 2 confirms that the idea of unfolding a single conjunctive requirement in multiples requirements can also be used with the online GAN framework. Algorithm 2 is then extended with the addition of a multi-armed bandit to produce Algorithm 3. In our experiment, Algorithm 3 is more

efficient than Algorithm 2 both in achieving the falsification with less function evaluations and using less computational resources.

We should note that the presented evaluation of the proposed algorithms is based on a small number of numerical experiments. Also the performance comparisons between the algorithms presented here and the algorithm presented in [13] are only indicative.

As a future work, we acknowledge that a more comprehensive evaluation with a larger set of problems is necessary in order to establish the benefits of the proposed algorithms. Finally, the theoretical underpinnings of the online GAN framework should be studied in more detail in order to understand better its benefits and limitations as an optimization tool.

ACKNOWLEDGMENTS

This research work has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 101007350. The JU receives support from the European Union’s Horizon 2021 research and innovation programme and Sweden, Austria, Czech Republic, Finland, France, Italy, Spain.

REFERENCES

- [1] F. Archetti and A. Candelieri. *Bayesian Optimization and Data Science*. SpringerBriefs in Optimization. Springer, 2019.
- [2] A. Donzé and O. Maler. Robust satisfaction of temporal logic over real-valued signals. In K. Chatterjee and T. A. Henzinger, editors, *Formal Modeling and Analysis of Timed Systems*, pages 92–106. Springer Berlin Heidelberg, 2010.
- [3] A. Corso et al. A survey of algorithms for black-box safety validation. *CoRR*, abs/2005.02979, 2020.
- [4] G. Ernst et al. Arch-comp 2021 category report: Falsification with validation of results. In G. Frehse and M. Althoff, editors, *8th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH21)*, volume 80 of *EPiC Series in Computing*, pages 133–152. EasyChair, 2021.
- [5] Gidon Ernst et al. Arch-comp 2020 category report: Falsification. In G. Frehse and M. Althoff, editors, *7th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH20)*, volume 74 of *EPiC Series in Computing*, pages 140–152. EasyChair, 2020.
- [6] I. J. Goodfellow et al. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014.
- [7] J. Cralley et al. TLTK: A toolbox for parallel robustness computation of temporal logic specifications. In J. Deshmukh and D. Ničković, editors, *Runtime Verification*, pages 404–416. Springer International Publishing, 2020.
- [8] L. Mathesen et al. Falsification of cyber-physical systems with robustness uncertainty quantification through stochastic optimization with adaptive restart. In *15th IEEE International Conference on Automation Science and Engineering, CASE 2019*, pages 991–997. IEEE, 2019.
- [9] P. Heidlauf et al. Verification challenges in F-16 ground collision avoidance and other automated maneuvers. In G. Frehse, editor, *5th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH18)*, volume 54 of *EPiC Series in Computing*, pages 208–217. EasyChair, 2018.
- [10] J. C. Helton and F. J. Davis. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliab. Eng. Syst. Saf.*, 81(1):23–69, 2003.
- [11] B. Hoxha, H. Abbas, and G. Fainekos. Benchmarks for temporal logic requirements for automotive systems. In G. Frehse and M. Althoff, editors, *1st and 2nd International Workshop on Applied Verification for Continuous and Hybrid Systems (ARCH14-15)*, volume 34 of *EPiC Series in Computing*, pages 25–30. EasyChair, 2015.

- [12] O. Maler and D. Nickovic. Monitoring temporal properties of continuous signals. In Yassine Lakhnech and Sergio Yovine, editors, *Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems*, pages 152–166. Springer Berlin Heidelberg, 2004.
- [13] L. Mathesen, G. Pedrielli, and G. Fainekos. Efficient optimization-based falsification of cyber-physical systems with multiple conjunctive requirements. In *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, pages 732–737, 2021.
- [14] D. Ničković and T. Yamaguchi. RTAMT: Online robustness monitors from STL. In D. Van Hung and O. Sokolsky, editors, *Automated Technology for Verification and Analysis*, pages 564–571. Springer International Publishing, 2020.
- [15] I. Porres, H. Rexha, and S. Lafond. Online GANs for automatic performance testing. In *IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW 2021)*, pages 95–100, 2021.
- [16] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, 2006.
- [17] S. Sankaranarayanan and G. Fainekos. Falsification of temporal properties of hybrid systems using the cross-entropy method. In *Proceedings of the 15th ACM International Conference on Hybrid Systems: Computation and Control*, page 125–134. Association for Computing Machinery, 2012.
- [18] A. Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1–2):1–286, 2019.
- [19] Z. Zhang, I. Hasuo, and P. Arcaini. Multi-armed bandits for boolean connectives in hybrid system falsification. In I. Dillig and S. Tasiran, editors, *Computer Aided Verification*, pages 401–420. Springer International Publishing, 2019.