Using the Kolmogorov-Smirnov statistic for evaluating test generation quality

Benoît Barbot
LACL, Université Paris-Est Créteil, France

Nicolas Basset, Thao Dang, and Ouri Maler
VERIMAG/CNRS, Université Grenoble Alpes, France

I. Introduction

The heterogeneity of cyber-physical system (CPS) components requires different modelling and design paradigms the combination of which is not easy to analyse rigorously. For this reason, industrial CPS are mostly validated by simulation or testing, which entail resolving uncertainty and non-determinism to produce one single behaviour at a time. A naturally problem that arises is to judge the goodness of a finite set of behaviours for assessing the system correctness. Inspired by the Monte Carlo and quasi-Monte Carlo approaches, we aim at defining a figure of merit which quantifies the uniformity or equidistribution of a set of behaviours. To this end, we propose to use the Kolmogorov-Smirnov (KS) statistic. This figure of merit can be used, on one hand, to derive guarantees about how reliable the result yielded by a given tested set of behaviours is (such as in terms of error bounds on the estimated worst-case or average property satisfaction robustness), or to suggest efficient test generation methods. In the following we first describe the testing framework and then explain how the KS statistic can be used as a test quality measure. We then describe how the KS statistic can be estimated.

II. Testing framework

We consider temporal specifications described by timed automata [1]. Roughly speaking, a timed automaton is an automaton equipped with a finite set of real-valued clock variables X. Each transition connecting two automaton locations is associated with an event, a guard (defined by a finite conjunction of clock constraints of the form $x \sim c$ or $x-y \sim c$ where $\sim \in \{\leq, <, =, >, \geq\}$ with $x, y \in X,\ c \in \mathbb{N}$) and a reset function specifying a subset of clock variables to be reset to 0. A run is a sequence $(q_0, \mathbf{x}_0) \xrightarrow{t_1,\delta_1} (q_1, \mathbf{x}_1) \dots \xrightarrow{t_n,\delta_n} (q_n, \mathbf{x}_n)$ where $\mathbf{x}_i$ are clock vectors and δ_i are transitions. The *delay* t_i represents the time before firing the transition δ_i. A run is called accepting if it starts in the specified initial state and ends in a specified final state. Given a discrete path $w = \delta_1 \cdots \delta_n$, we denote by P_w the set of timed vector $\vec{t}$ such that $(q_0, \mathbf{x}_0) \xrightarrow{t_1,\delta_1} \dots \xrightarrow{t_n,\delta_n} (q_n, \mathbf{x}_n)$ is an accepting run. This set is called the *timed polytope* associated to the path w, and all sequences $(t_1, \delta_1) \dots (t_n, \delta_n)$ are called timed words. Timed automata are a powerful formalism to specify time constraints on sequences of events. To use them for CPS behaviours expressed as real-valued signals, each event is associated with a set of constraints on the signal values, for example to specify threshold crossing, entering a region and more complex temporal patterns. Since a major goal of testing is to detect bugs, if a behaviour does not correspond to an accepting run of the timed automaton, a bug is found and reported. If none of the tested behaviour falsifies the specification, the associated set of accepting runs is collected for test quality assessment, to yield the confidence on the correctness of the system.

III. Kolmogorov-Smirnov statistic as test quality measure

Let us first recall the KS statistic. We consider a random point $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_d)$ in R^d, with joint distribution function $F(\mathbf{x})$. The Kolmogorov-Smirnov statistic is a test of goodness of fit that compares the empirical distribution function $\hat{F}_n$ of a sample S of n points ($S = \{\mathbf{x}^1, \dots, \mathbf{x}^n\}$) against some (expected) distribution function F. This statistic can be used to test whether a sample is indeed drawn from F. The empirical distribution function $\hat{F}_n$ is defined as $\hat{F}_n(\mathbf{x}) = \frac{1}{n}\sum_{1\leq j\leq n} \mathrm{I}\{\mathbf{x}^j \preceq \mathbf{x}\}$ where I is the indicator function and $\mathbf{x}^j \preceq \mathbf{x}$ is true iff $\mathbf{x}^j_i \leq \mathbf{x}_i$ for every coordinate axis $i \in \{1, \dots, d\}$. The KS statistic measures the difference between $\hat{F}_n$ and F as follows: $\mathbf{KS}_n = \sup_{\mathbf{x}} |\hat{F}_n(\mathbf{x}) - F(\mathbf{x})|$. In general, this statistic is not distribution free; however, the distribution of $\mathbf{KS}_n$ is the same for all absolutely continuous distributions F. Indeed, if we consider the following transformation T from $\mathbf{x}$ to $\mathbf{y}$: $\mathbf{y}_1 = F_1(\mathbf{x}_1)$ and for $i = 2, \dots, d$ $\mathbf{y}_i = F_i(\mathbf{x}_i|\mathbf{x}_1, \dots, \mathbf{x}_{i-1})$, then $\mathbf{y}_1, \dots, \mathbf{y}_d$ are a sample from the uniform-$[0, 1]$ distribution [7]. Then, the KS statistic in question can be expressed as $sup_{\mathbf{y}\in[0,1]^d}|\hat{G}_n(\mathbf{y}) - G(\mathbf{y})|$ where $\hat{G}_n$ is the empirical distribution of the transformed sample $\mathbf{y}$ and the uniform distribution function $G(\mathbf{y}) = \mathbf{y}_1\mathbf{y}_2 \dots \mathbf{y}_d$. Here we can see that the KS statistic is indeed distribution free (for all continuous underlying distributions F).

Application to sets of CPS behaviours. The KS statistic is defined for samples of points, we need thus to map a set of CPS behaviours to a set of points, via a timed abstraction. Note that each accepting run corresponds to a timed vector in a timed polytope, hence we can evaluate the uniformity of the set of runs by the uniformity of the set of timed vectors over the timed polytopes. The latter can be judged by the KS statistic. To do so, as explained in the above, we need a transformation T that maps the timed words to a set of points in the unit box. In [3], such a transformation is defined by the conditional cumulative distribution function (CDF) $F = (F_1, \dots, F_n)$,

which for each path in the timed automaton can be written as $F_i(t_i \mid t_1, \ldots, t_{i-1}) = p_i(t_1, \ldots, t_{i-1})/q_i(t_1, \ldots, t_i)$ where p_i and q_i are polynomials of degree at most i. Each CDF F_i can be computed in time polynomial in the path length. The computation of these conditional CDF was implemented in the tool Cosmos [2].

Kolmogorov-Smirnov statistic estimation. We remark that the KS statistic definition using the distributions $\hat{G}_n$ and G of the transformed sample in the unit box $[0,1]^d$ is exactly the star-discrepancy of the transformed sample [6]. In [4] the star-discrepancy is used to define a coverage measure for continuous and hybrid systems, and an upper and a lower bound of the star discrepancy for a set of points can be estimated using a grid-based method [8]; however accurate estimations often require fine grids and computation time is thus exponential in dimension. In this work, we alternatively use an approximate version, proposed in [5], which considers only the maximal difference between $\hat{G}_n$ and G only over the points in the transformed sample denoted by $T(S)$, that is $\tilde{\mathbf{KS}}_n = max_{\mathbf{y} \in T(S)} |\hat{G}_n(\mathbf{y}) - G(\mathbf{y})|$. The larger the size of the sample is, the more accurate this approximation is. We also remark that the computation of the KS statistic in the multivariate case is much more complex than in the univariate case due to the partial order $\preceq$; indeed the empirical distribution changes not only at the points in the sample but also at the points formed by combining the coordinates of the sampled points. Nevertheless, compared to the implementation of the grid-based estimation in [4] this approximation is less sensitive to dimension, which allows us to handle higher-dimensional point sets. Furthermore, in view of using the KS statistic to dynamically guide the test generation method we derive an incremental method to update the estimation with new points. The incremental method keeps a list of the values of the empirical distribution at the points as long as they are processed. When a new point is added, these values are updated according to the position of the new point, and the value for the new point is added in the list. To illustrate this estimation, we consider a timed automaton modelling bounded uncertainty in the signal period. We generate two sets of 98000 timed vectors of dimension 5 (that is, the timed words are of length 5), one is generated by the uniform random generation [3] and the other by a low-discrepancy method (which also uses the same CDF but the unit box is sampled using a low-discrepancy generation instead of the uniform random method). With such a large number of timed words in 5 dimensions, it was not possible to use the grid-based estimation to obtain bounding intervals which are accurate enough for comparison purposes. Using this new estimation method, we are able to compare the quality of the two sets generated by the two methods. For this particular example, the low-discrepancy method is better in terms of the KS statistic. Concerning computation time, the static version (which handles the full set of points) took about 6100 seconds, and the incremental version took about 9600 seconds. The figure 1 shows the KS statistic estimated incrementally for the set generated by the low-discrepancy method. We can observe a convergence of the estimate towards the exact result when the number of the considered points increases.

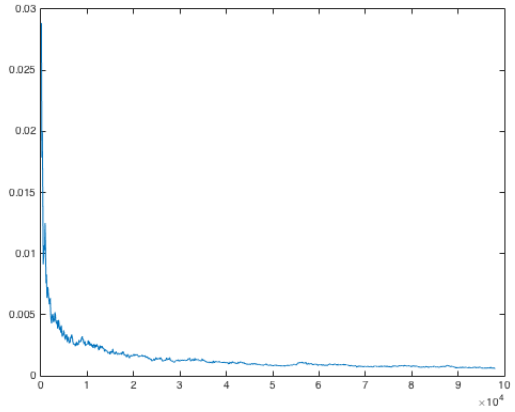

Figure 1. The KS statistic incrementally estimated (for the set generated by the low-discrepancy method)

Concluding remarks. We present in this abstract an important element of our CPS testing framework: defining and evaluating the test quality. We are currently working on a more efficient implementation of the KS statistic estimation by using the data structures that can handle better the geometry of point sets.

REFERENCES

[1] R. Alur and D. L. Dill. A theory of timed automata. *Theoretical Computer Science*, 126:183–235, 1994.

[2] P. Ballarini, B. Barbot, M. Duflot, S. Haddad, and N. Pekergin. HASL: A new approach for performance evaluation and model checking from concepts to experimentation. *Performance Evaluation*, 90(0):53 – 77, 2015.

[3] B. Barbot, N. Basset, M. Beunardeau, and M. Kwiatkowska. Uniform sampling for timed automata with application to language inclusion measurement. In *Quantitative Evaluation of Systems - 13th International Conference, QEST 2016, Quebec City, QC, Canada, August 23-25, 2016, Proceedings*, pages 175–190, 2016.

[4] T. Dang. *Model-Based Testing for Embedded Systems*, chapter Model-based Testing of Hybrid Systems. CRC Press, 2011.

[5] A. Justel, D. Peña, and R. Zamar. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters*, 35(3):251 – 259, 1997.

[6] H. Niederreiter. *Random Number Generation and quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[7] M. Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.

[8] E. Thiémard. An algorithm to compute bounds for the star discrepancy. *Journal of complexity*, 17(4):850–880, 2001.