Report draft number 3
Communication in the Mathematical Sciences Spring 202

Using Gaussian Processes to estimate excursion sets of black box functions

Irina Espejo

1 Introduction

Simulators are ubiquitous in applied sciences since the onset of the computational era. They are a powerful technique to study phenomena with accessible local rules yet complex global effects. Although, simulators are not new to researchers, they have recently become the rock stars of newspapers due to the covid-19 outbreak crisis. The predictions of how to "flatten the curve" with different social distancing restrictions are calculated using simulators. Because simulators work in a wide range of length scales, they are popular in such distinct disciplines like cosmology, molecular biology, material science, epidemiology, particle physics, climate science, protein folding, economics, neuroscience or population genetics.

Nevertheless, simulators have critical limitations. For instance, each call of a simulator might be computationally expensive and can last from minutes to days. Another limitation is that when simulators are internally complex, we lack important in-between steps information. For example, suppose we want to test different policies for social distancing during an epidemic to find a global minimum while allowing some degree of business mobility. We would have to run the simulation multiple times, save the outcome and then compare to find a minimum. A naive evaluation strategy, like a grid strategy, will make this task computationally unfeasible. Trying to find the global minimum of a simulation is only one example of multiple properties we might be interested in. Those functions where no extra information is given are called black box functions and simulators, in general, are considered a type of black box function. In this report we present the problem of finding an efficient evaluation strategy that allows us to estimate properties of black box functions.

Here we review a method called Active Learning for Simulation-Based Inference. Given a desired property we want to find, it provides a strategy that is computationally cheaper than a traditional grid search. In particular, we are interested in excursion sets, synonym for level set, of black box functions. The method is supported on concepts from Gaussian Processes, Bayesian Optimization and Reinforcement Learning.

This paper is organized as follows: each section deals with a different ingredient of the method. The authors will try to connect the ideas between sections whenever possible. In the last section we will tie all the concepts together into the description of an algorithm. To sum up, our main goal is to estimate excursion sets of black box functions where the output can be interpreted as a p-value of one hypothesis vs the null.

2 Excursion sets and thresholds

In this section we will motivate the importance of obtaining excursion sets of black box functions. To illustrate this, let's suppose that a black box function $f(x)$ with $x = (x_1, x_2)$ is a weather forecast simulation. The variable x_1 is temperature and x_2 humidity and the output $f(x_1, x_2)$ is the probability that, given the weather today, tomorrow the temperature will be x_1 and humidity x_2. Suppose a scientist has the hypothesis H_1 that tomorrow it will be specially hot with $x_1 = 113$ F and $x_2 = 99\%$. Then, he or she should test the hypothesis of a hot day against the null hypothesis H_0: that tomorrow the weather will be approximately the same as today. Not surprisingly, this is related to output of the simulation $f(x_1 = 113, x_2 = 99\%)$ in the following way: $f(x_1 = 113, x_2 = 99\%)$ is the p-value for the hypothesis H_1 vs H_0. Realistically, the choice of $x_1 = 113$F and $x_2 = 99\%$ is rather arbitrary and ideally the scientist would like to know the p-value of as many pairs (x_1, x_2) as possible. However, there is no free lunch and the scientist must still introduce subjectiveness by defining a threshold for the p-value function. This threshold is also known as confidence, for instance a threshold of $t = 0.95$ means that if $f(x_1 = 113F, x_2 = 99\%) \leq 1 - t$, then it is quite unrealistic to think that tomorrow we will have the temperature-humidity pair $(x_1 = 113F, x_2 = 99\%)$ because the probability is lower than 0.05. In that case we would reject the hypothesis that tomorrow there will be a temperature of $113F$ and humidity of 99%.

In the last example, it was unavoidable to introduce a threshold, more precisely we are interested in excursion sets of black box functions. Given a threshold $t \in \mathbb{R}$ the excursion set for $f(x)$ at t is

$$E_f(t) = \{x \in \mathbb{R}^n \,|\, f(x) = t\}$$

One could also be interested in calculating multiple excursion sets for different thresholds $t_1, .., t_k$ of the same black box function f. For now, for the sake of simplicity, we will restrict ourselves only to one level set. Finally, the connection between excursion sets and simulators is that often the output of a simulator $f(x)$ is a test statistic, like a p-value, and we are interested in the excursion set for a particular confidence.

3 Gaussian Processes as a way to interpolate between black box samples

If we hope to estimate excursion sets of a black box function, first we should have a model for the black box function itself. Recall that black box functions are expensive to evaluate so the number of samples available is typically not large enough to fully characterize excursion sets. The idea is to use a surrogate, cheap, analytical model to infer the black box function where we have no samples. Our approach uses the Gaussian Process framework to construct the surrogate function. The Gaussian Process framework consists on two steps: first we impose a Gaussian prior (or belief) over the space of functions $p(f(x))$. More precisely, $p(f(x))$ follows a

$$\mathcal{N}(\mu_{prior}(x), k_{prior}(x, x'))$$

where $\mu_{prior}(x)$ is the mean function and $k_{prior}(x, x')$ is the covariance function between two points. This prior means that we believe out black box function $f(x)$ is sampled from a Gaussian Process with an ansatz mean and covariance close to $\mu_{prior}(x)$ and $k_{prior}(x, x')$.

Secondly, we take as surrogate model the posterior w.r.t a dataset. In the Bayesian setting, the posterior w.r.t a dataset is the product of the prior and the likelihood of that dataset. One can think the posterior as an update of the prior (our beliefs) after we observe the dataset. More precisely, let $f(x)$ be a black box function from $\mathbb{R}^n$ to $\mathbb{R}$ and let

$$D_n = \{(x_i, f(x_i)), i \in 1..n\}$$

be a dataset consisting of pairs of black box function evaluations. The x component of D_n will be relevant and denoted by the vector ${X_n}^T = (x_1, ..., x_n) \mid x_i \in D_n$. From the literature we know that the posterior also follows a normal distribution over functions.

$$p(f(x)|D_n) = \mathcal{N}(\mu_{post}(x), k_{post}(x,x')) \tag{1}$$

The quantities $\mu_{post}(x)$ and $k_{post}(x,x')$ have the following analytical form

$$\mu_{post}(x) = \mu_{prior}(x) + k_{prior}(x,X_n)k_{prior}(X_n,X_n)^{-1}f(X_n) \tag{2}$$

$$k_{post}(x,x') = k_{prior}(x,x') - k_{prior}(x,X_n)k_{prior}(X_n,X_n)^{-1}k_{prior}(x',X_n)^T \tag{3}$$

We take the function $\mu_{post}(x)$ with uncertainty $k_{post}(x,x')$ as the surrogate model for the black box function $f(x)$. Intuitively, definition 1 means that at each point x we have a Gaussian distribution over the possible values of $f(x)$ with analytical parameters that only depend on the dataset and the point x, as shown in equation 3. Figure 5 gives a visual example of a Gaussian Process in one dimension.

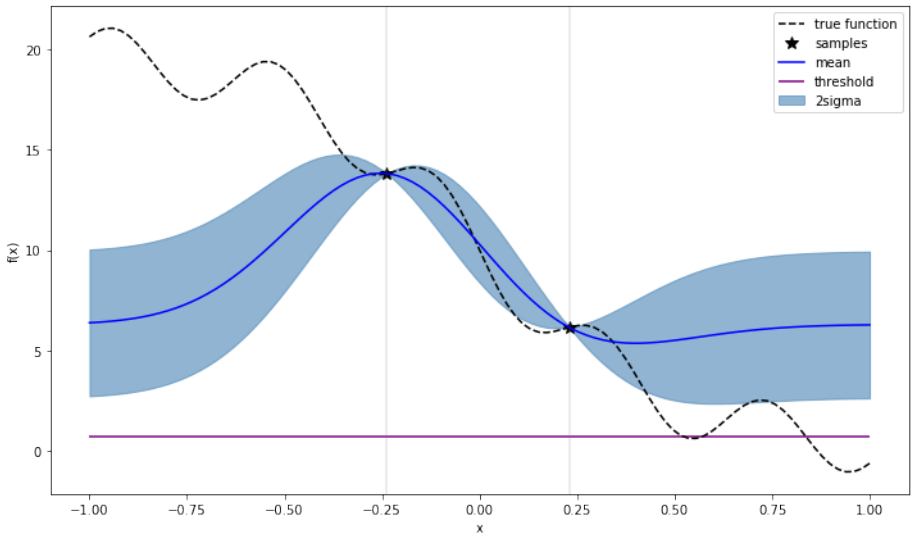

Figure 1: The dotted line is the true, unknown, black box function $f(x)$. We have only two samples $(x, f(x))$ from it marked as black stars. The threshold we are interested in is marked in purple $t = 1$. We are interested in estimating the excursion set $\{x \in [-1,1] | f(x) = t\}$. The blue line is the mean $\mu(x)$ of the Gaussian process fitted to the two available true samples. The blue shadow is the covariance at each point.

4 The acquisition function or how to choose a smart sampling strategy

Suppose we are in the situation depicted in Figure 5 and that we have only available two samples (two black stars). How can we accurately estimate the excursion set $E_f(t=1)$? We need to invoke the black box function at more points. A naive strategy would be to set up a coarse grid, for instance starting at -1, stopping at +1 with step 0.01. This grid would have 200 points, if each evaluation of the black box function $f(x)$ lasts 10 minutes, then evaluating the whole grid would take approximately half an hour. The example depicted in Figure is a simple one: one dimension, on the segment [-1,1] and also 10 minutes for a simulation is a short estimation. Therefore, if we intend to study real world cases, we need a more efficient sampling strategy in terms of computational cost. The less evaluations of the black box function, the less computation time. Our main goal in this section is to develop a criteria to estimate excursion sets of black box functions with as few evaluations as possible.

The acquisition function $acq(x)$ determines which $x \in [-1,1]$ in the domain should we evaluate $f(x)$ next. The choice of an $acq(x)$ is chosen by the scientist and typically depends on the field. The acquisition function incorporates explicitly the Gaussian Process framework detailed in the last section. Intuitively, the acquisition function is a measure of "how much is it worth it" to evaluate a particular point n the domain, considering how expensive it is to invoke the black box function. We choose the new x by maximizing the acquisition function

$$x_{\text{next}} = \text{argmax}_{x\in[-1,1]}\ acq(x)$$

An example of a popular acquisition function we have the Lower Confidence Bound (LCB) defined as

$$LCB(x) = \mathbb{E}_{p(f(x)|D_n)}\left[\mu_{post}(x) - k_{post}(x,x)\right]$$

We use the LCB when we want to favor x_{new} to be a point x such that the threshold (e.g. $t=1$) is inside the region $[\mu_{post}(x), \mu_{post}(x) - k_{post}(x,x)]$ of possible images $p(f(x_{new}))$. The LCB is only an example of an acquisition function, for the sake of simplicity that will be the only example in this report.

5 The method

Finally, in this section we will tie together the concepts presented along the report into the Active Learning for Simulation-Based Inference method. Let's start defining again the setting. We have a computationally expensive black box function $f(x)$ for which we do not have an analytical formula. Suppose we start our research with a small dataset of $f(x)$ samples, for instance two samples $D_2 = \{(x_1, f(x_1)), (x_2, f(x_2))\}$ as in Figure . Now we are interested in estimating the excursion set for $t=1$ this is $E_f(t=1)$. The only way to obtain more information is by invoking $f(x)$ at a new point x_3 not in an arbitrary way but using an acquisition function. If we use the $LCB(x)$ function then

$$\begin{aligned} x_3 &= \text{argmax}_{x\in[-1,1]}\ LCB(x) \\ x_3 &= \text{argmax}_{x\in[-1,1]}\ \mathbb{E}_{p(f(x)|D_2)}\left[\mu_{post}(x) - k_{post}(x,x)\right] \end{aligned}$$

All the elements in the last equation can be calculated using the Gaussian Process framework as shown in section 3. This sequence of steps is iterated as many times as desired.

The pseudocode for the algorithm is presented below

Algorithm 1: Active Learning for Simulation-Based Inference

Result: A set of points x_{new} to estimate excursion set
//Initialization;
D_2;
X_{grid} over [-1,1];
$\mu_{prior}(x), k_{prior}(x,x')$;
for $i \in range(N_{iter})$ **do**
  //Bayesian inference;
  X_{i-1} is the x component of D_{i-1};
  for $x \in X_{grid}$ **do**
    $\mu_{post}(x) = \mu_{prior}(x) + k_{prior}(x, X_{i-1}) k_{prior}(X_{i-1}, X_{i-1})^{-1} f(X_{i-1})$;
    $k_{post}(x,x') = k_{prior}(x,x') - k_{prior}(x, X_{i-1}) k_{prior}(X_{i-1}, X_{i-1})^{-1} k_{prior}(x', X_{i-1})^T$;
    //Calculate acq(x) using $\mu_{post}(x)$ and $k_{post}(x,x')$;
    //save them in array;
    $values\_acq \leftarrow acq(x)$;
  end
  //Select new x and y;
  $x_{new} = \text{argmax}_{x \in X_{grid}}\ values\_acq$;
  //Invoke black box function;
  $y_{new} \leftarrow f(x_{new})$;
  //Update dataset;
  $D_i \leftarrow D_{i-1} \cup (x_{new}, y_n ew)$;
  //Reset;
  $x_{new} \leftarrow 0$;
  $y_{new} \leftarrow 0$;
  $values\_acq \leftarrow [\,]$;
end

6 Conclusion

To conclude, in this report we have explained all the necessary ingredients to formulate an algorithm that estimates excursion sets of computationally expensive black box functions. Crucially, this algorithm requires less computational time than a naive grid strategy. Despite the complexity of this algorithm and of its implementation, it makes feasible tasks that would be too expensive to calculate otherwise. In this report we have focused on excursion sets because of its importance in hypothesis testing.

The limitations of this method are numerous, the most important one is its scaling with domain dimensionality. If each point in the domain is d dimensional we suffer from the so-called curse of dimensionality. A naive grid search would still need more computational time than the Active Learning method. Further work will be needed to address that regard if we hope to study real world simulators. In particular, an interesting idea would be trying to borrow adaptive mesh refinement techniques from the field of PDEs. Roughly, this approach would consist of creating a coarser grid around the excursion

sets and a sparse grid anywhere else. Since we do not know a priori where the excursion sets are, this we would have to learn the grid refinement online as evaluations of the black box function come.