Convergence of distributed symbolic regression.

Ben Cardoen
ben.cardoen@student.uantwerpen.be

August 3, 2017

1 Abstract

Symbolic regression (SR) fits symbolic expression to a dataset of expected values. Amongst its advantages over other techniques are the ability for a practitioner to interpret the resulting expression, determine important features by their usage in the expression, and an insight into the behavior of the resulting model (e.g. continuity, derivation, extrema). SR combines a discrete combinatoric problem (combining base functions) with a continuous optimization problem (selecting and mutating constants). One of the main algorithms used in SR is genetic programming. The convergence characteristics of SR using GP are still an open issue. In this work we will use a distributed GP-SR implementation to evaluate the effect of topologies on the convergence of the algorithm. We introduce and evaluate a topology with the aim of finding a new balance between diffusion and concentration. We introduce a variation of k-fold cross validation to estimate how accurate a generated solution is in predicting unknown datapoints. This validation technique is implemented in parallel in the algorithm combining both the advantages of the cross validation with the increase in covered search space for each instance. We validate our work on several test problems and a real world use case.

2 Introduction and design

Our tool is implemented in Python. The language offers portability, access to rich libraries and fast development cycles. The disadvantages are speed and memory usage compared with compiled languages (e.g. C++) or newer scripting languages (e.g Julia). Furthermore, Python's usage of a global interpreter lock makes shared memory parallelism not feasible. Distributed programming is possible using MPI.

2.1 Algorithm

The algorithm accepts a matrix X = n x k of input data, and a vector Y = 1 x k of expected data. It will evolve expressions that result, when evaluated on X, in an 1 x k vector Y' that approximates Y. N is the number of features, or parameters, the expression can use. We do not know in advance if all features are needed, which makes the problem statement even harder. The goal of the algorithm is to find f' such that

$$Y = f(X), Y' = f'(X), d(Y, Y') = \epsilon$$

results in ϵ minimal. F is the process we wish to approximate with f'.
Not all distance functions are equally well suited for this purpose. A simple root mean squared error (RMSE) function has the issue of scale, the range of this function is [0, +∞), which makes comparison problematic, especially if we want to combine it with other objective functions. A simple linear weighted sum requires that all terms use the same scale. Normalization of RMSE is an option, however there is no single recommended approach to obtain this NRMSE.
In this work we use a distance function based on the Pearson Correlation Coefficient. Specifically, we define

$$d(Y, Y') = 1 - |\frac{\sum_{i=0}^{n}(y_i - E[Y]) * (y'_i - E[Y'])}{\sqrt{\sum_{j=0}^{n}(y_j - E[Y])^2 * \sum_{k=0}^{n}(y'_k - E[Y'])^2}}|$$

R has a range of [-1, 1] where 1, -1 indicate linear and negative linear correlation respectively, and 0 indicates no correlation. This function has a range [0,1] which facilitates comparison across domains and allows combining it with other objective functions. The function reflects the aim of the algorithm. We not only want to assign a good (i.e. minimal) fitness value to a model that has a minimal distance, we also want to consider linearity between Y an Y'. The use of the Pearson correlation coefficient as a fitness measure is not new, a variant of this approach is used in [14].

2.1.1 Genetic Programming Implementation

We use Genetic Programming (GP) [9] to find solution to the problem statement. The algorithm controls a population of expressions, represented as trees, that are initialized, evolved using operators, and selected to simulate evolution. The algorithm is subdivided in a set of phases, each phase initializes the population with a seed provided by an archive populated by previous phases or by the user. A phase is subdivided in runs, where each run selects a subset of the population, applies operators and if the application of the operator leads to fitness improvement replaces expressions. At the end of a phase the best expressions are stored in an archive to seed consecutive phases. At the end of a phases the best expressions are communicated to other processes executing the same algorithm with a differently seeded population. The next phase will then seed its population using the best of all aggregated expressions. We use a vanilla GP implementation, with 'full' initialization method. Expressions trees are generated with a specified minimal depth. The depth of the expressions during evolution is limited by a second maximal parameter. GP differs from most optimization algorithms in this variable length representation. We use 2 operators in sequence, mutation and crossover. Mutation replaces a randomly selected subtree with a randomly generated tree. Mutation introduces new information, and leads to exploration of the search space. Crossover selects 2 trees based on fitness and swaps randomly selected subtrees. Crossover tends to lead to exploitation of the search space. Selection for crossover is random biased by fitness value. A stochastic process decides if crossover is applied pairwise (between fitness ordered expressions in the population) or at random. The initialization of expression trees can lead to invalid expressions for the given domain. The probability of an invalid expression increases exponentially with the depth of the tree. A typical example of an invalid tree is division by zero. While some implementations opt for a guarded implementation, where division is altered in semantics to return a 'safe' value if the argument is zero, we feel that this alters the semantics of the results, and is somewhat opaque to the user. Our implementation will discard invalid expressions and replace them with a valid expression. Another approach is assigning a maximal fitness value to such an expression, but this can lead in corner cases to premature convergence when a few valid expressions dominate the population early on. We implement an efficient bottom up approach to construct valid trees where valid subtrees are merged. In contrast to a top down approach this detects invalid expressions early and avoids unnecessary evaluations of generated subtrees. Nevertheless, the initialization problem leads to a significant computational cost in the initialization stage of each phase and in the mutation operator.

2.2 Distributed algorithm

GP allows for both fine grained and coarse grained parallelism. Parallel execution of the fitness function can lead to a speedup in runtime, but will not alter the search process. Python's global interpreter lock and the cost of copying expressions for evaluation makes this approach unfeasible. A more interesting approach is coarse grained parallelism where we execute k instances of the algorithm in parallel and let them exchange their best expressions given a preset topology. The topology will determine both the convergence of the algorithm and the runtime.Exchanges of messages can introduce serialization and deadlock if the topology contains cycles. Our tool supports any user provided topology so must be able to deal with both issues effectively. After each phase a process looks up its targets given the topology. It then sends its best k expressions to the set of targets, either by copying all expressions to all targets or by spreading them over the target set. When the sending stage is complete, the process looks up its set of sources and waits until all source processes have sent their best expressions. To avoid deadlock a process sends its expressions asynchronously, not waiting until the receiving process has acknowledged receipt. The sent expressions are stored in a buffer for this purpose, together with an associated callable waiting object. After the sending stage the process synchronously collects messages from its sources, and executes the next phase of the algorithm. Before the next sending stage, the process will then check each callable to verify that all messages from the previous sending phase have been collected. Once this blocking call is complete, it can safely reuse the buffer and start the next sending phase. This also introduces a delay factor between processes. The phase runtime between processes will never be exactly identical, especially not given that the expressions have a variable length representation and differing evaluation cost. Without a delay factor processes would synchronize on each other, nullifying any runtime gains. With this delay factor a process is able to advance k phases ahead of a target process k steps distant in its topology. For hierarchical, non-cyclic topologies this can lead to a perfect scaling, where synchronization decreases as the number of processes increases.

2.3 Approximated k-fold cross validation

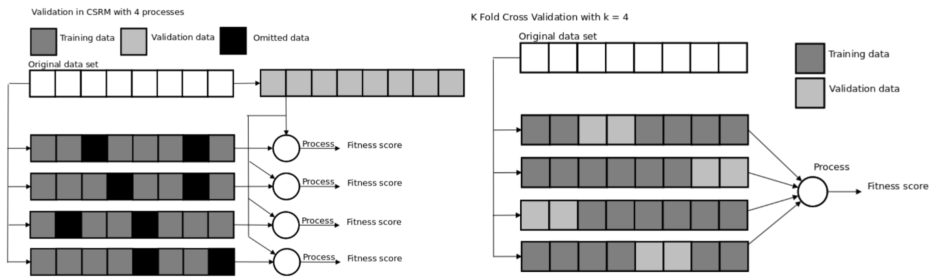
We divide the data over k processes with each process use a random sample of 4/5 of the total data. Each process operates on a further 4/5 split between training and validation data. The aggregate distributed process then approximates k-fold cross validation. Independent of the topology each randomly chosen pair of communicating processes will have the same probability of overlapping data. When this probability is too low, overfitting will be introduced and highly fit expressions from one process will have a high probability to be invalid for another process' training data. When the overlap is too great both processes will be searching the same subspace of the search space.

2.4 Topologies

A topology in this context is the set of communication links between the processes. The topology influences the convergence characteristics of the entire group. In a disconnected topology, there is no diffusion of information. If a process discovers a highly fit expression, that knowledge has to be rediscovered by the other processes in order to be used in the process. An edge case where this is an advantage is if we see the group of processes as a multiple restart version of the sequential algorithm. If the risk for premature convergence due to local optima is high, we can try to avoid those optima by executing the algorithm in k instances, without communication. Such an approach is sometimes referred to as partitioning, as we divide the search space in k distinct partitions.

Diffusion and concentration Our aim is for the processes to share information in order to accelerate the search process. With a topology we introduce two concepts : concentration and diffusion. Concentration refers to processes that share little or no information and focus on their own subset of the search space. Like exploitation concentration can lead to overfitting and premature convergence. It is warranted when the process is nearing the global optimum. Diffusion, in this work, is the spread of information over the topology. Diffusion accelerates the optimization process of the entire group. It is not without risk, however. If diffusion is instant, a single suboptimal solution can dominate other processes, leading to premature convergence. The distance between processes and connectivity will determine the effect diffusion has.

Grid The grid topology is 2 dimensional square of k processes with k a square of some natural number. Each process is connected with 4 other processes in north/south, east/west direction. The grid allows for delayed diffusion, to reach all processes an optimal expressions needs to traverse $\sqrt{k}$ links. This prevents premature convergence, with all processes still interconnected.



(a) Approximation of k fold cross validation with parallel processes, k = 4, r = $\frac{3}{4}$.

(b) Visualization of k fold cross validation with k = 4.

Figure 1: Topologies in CSRM.

Tree A binary tree with with root as a source and leafs as sinks with unidirectional communication, the tree topology is an efficient example of a hierarchical topology. For k processes there are k-1 communication links, reducing the messaging and synchronization overhead significantly compared to the grid topology (4k). Diffusion can be hampered, each link is unidirectional so an optimal expression will not travel upwards to the root. On the other hand, with a spreading distribution policy where the optimal expressions are spread over the outgoing link, a partitioning effect will occur which can prevent premature convergence. There are no cycles, synchronization overhead is low.

Random In a random topology it is hard to predict how the convergence of the aggregated process will behave. It is possible that the topology contains cycles, but at the same time it can contain cliques. Diffusion is not guaranteed, and runtime performance will differ based on the actual instance. With grid and tree convergence and synchronization behavior is deterministic. The advantage of the random topology is that it can avoid certain patterns that occur in the other deterministic topologies. If the grid tends to lead to premature convergence for a given problem instance, it is possible that a random topology will avoid this.

3 Experiments

3.1 Reproducability

All benchmarks were performed on an Intel Xeon E5 2697 processor with 64GB Ram, with Ubuntu 16.04 LTS, kernel 4.4.0 as operating system. CSRM is implemented using Python3, the test system uses Python 3.5. The experiments use a fixed seed in order to guarantee determinism. Where relevant the configuration is given. An open source repository holds the project's source code, benchmark scripts, analysis code and plots. The project dependencies are minimal making the project portable across any system that has a working Python3 implementation and pip as an installation manager. In order to run distributed the project requires an MPI implementation, which is available for most platforms.

3.2 Benchmark problems

Recent work on the convergence of GP-based SR [7, 8] featured a set of benchmark problems that pose convergence problems for SR implementations. These 15 problems use at most five features, and are non linear arithmetic expressions using the standard set of base functions : (sin, cos, tan(h), log,a^x, /, *, modulo, abs, min, max, +, -). CSRM does not know which features are used, making the problem harder. In other words it assumes each problem is a function of 5 features which may or may not influence the expected outcome. This is an extra test in robustness for the algorithm, while also testing the algorithm's capability as a classifier.

3.3 Experiment setup

Each process has a population of 20, initial depth of 4, max depth of 8, executes 20 phases of 20 generations each, with an archivesize of 20. Per phase the 4 best expressions are archived, the spreading policy distributes the best expressions over the communication links. The benchmark functions have at most 5 features, each with 20 sample points in the range [1,5]. The Grid and Tree topology will have single expression per link, the Random topology will have 2 per link. 25 processes are used, resulting in respectively 400, 15 ,50 being sent per phase. The random topology in this case contains 2 cliques of cycles.

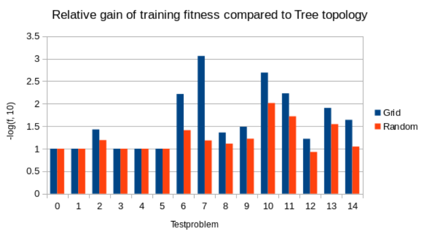
3.3.1 Measures

When the experiment ends the best 20 expressions from all processes are collected and scored. We measure best fitness on the training data, and best fitness on the validated data. Finally we record the mean fitness values of the best 5 expressions, both on the training and validation data. The mean is restricted to the upper quarter of the population specifically to measure how the best expressions are distributed. This measure records the convergence more accurately as the fittest expressions drive the convergence rate.

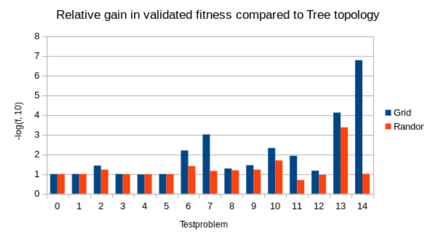
Calculation The fitness values fluctuate strongly between the test problems and even between topologies. We apply a negative logarithmic scale : $f_t = -\log_{10}(f)$ where f is either the best fitness value or the mean. Then we scale the results relative to the values obtained for the tree topology in order to measure relative gain or loss in orders of magnitude.

3.3.2 Results

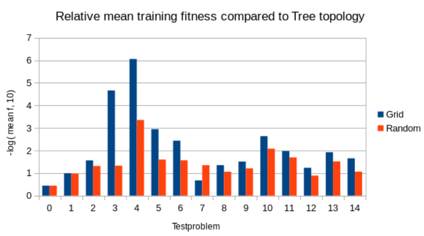
Convergence In Figure 2 we see how the topology affects convergence. The first observation we make is that the first 5 testproblems, with exception of the second, all have identical values for best fitness on training and validation data. The processes



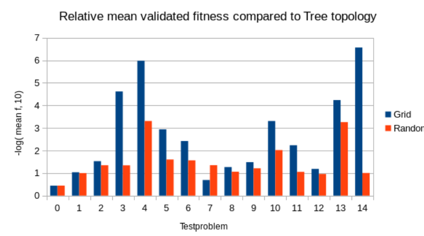
(a) Relative gain in best fitness of training data



(b) Relative gain in best fitness on full data.



(c) Relative gain in mean fitness on training data.

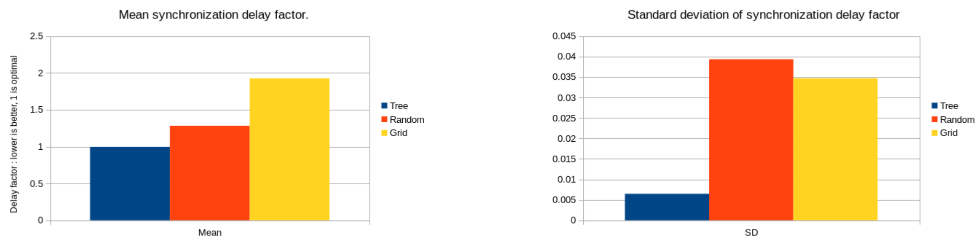


(d) Relative gain in mean fitness on full data.

Figure 2: Convergence differences between topologies.

converged to zero fitness values for these problems, hence the identical results. The training fitness results in Figure 2a indicate that the grid and random topology have superior convergence characteristics compared to the tree topology, with grid outperforming random on several testproblems. When we look at the fitness values on the validation data in Figure 2b we see a more nuanced result. The grid topology is overall still the best choice, but the random topology has worse results for problems 11 and 12. Next we look at the mean fitness values on training and validation data. Interestingly enough for problem 0 both random and grid score far worse than the tree topology. Overall the grid topology is still better for most problems, with the exception of problem 7. Note the similarity in pattern in the results here between training and validation data indicating that the predictive capability of the results is still good. If overfitting would have taken place we would see a reverse pattern in the results for the validation data.

Measuring Overhead Cycles in the topology will lead to excessive synchronization and even serialization. We measure the mean execution time for testproblem 6. The convergence characteristics using the three topologies differed significantly making this a good testcase. The processes will communicate 25 times. If the runtime of a single phase is too long, the overhead of communication will become hard to measure. If it is too short the overhead dominates the entire runtime. This second case is one we should try to avoid, it will unfairly penalize topologies with cycles forcing them to serialize. The runtime of a phase is dependent on the generations, population and depth ranges of the expressions. Ideally we would like for a practitioner to choose



(a) Mean synchronization delay factor.

(b) Standard deviation of synchronization delay factor.

Figure 3: Synchronization overhead introduced by topologies.

these parameters based on the problem at hand and not be constrained by synchronization considerations. We compare the three topologies and use the disconnected or 'none' topology as a reference point, which has zero synchronization overhead. This last topology has an ideal speedup of n, where n is the processcount, compared to a sequential process. From the synchronization overhead we can then derive the speedup each of the topologies is able to offer. In practice even the 'none' topology will have some synchronization overhead, as the root process has to collect all results from the other processes. In Figure 3 we see that the tree topology has a near zero delay introduced by the synchronization. This is due to the delay tolerance we have built in in our implementation. The random topology has a mean delay factor of 1.3, the grid topology scores worst with a mean delay factor of nearly 2. This is easily translated in terms of speedup. A tree topology will have near linear speedup, a grid will have a speedup roughly half of that value and a random topology will have a speedup bracketed between those two. The standard deviation for the tree topology is significantly smaller indicating that a tree topology will have a far more predictable speedup.

4 Use Case

4.1 Problem statement

We want to apply symbolic regression on the output of a simulator. The simulator we use for our experiment is a high performance tool [13] that models the spread of infectious diseases. Epidemiological simulation is a vital tool for policy makers. Simply observing the real world process is a measure of last resort at unacceptable cost in human suffering, and the resulting data is not strictly predictive for new outbreaks. A single outbreak is a sample of a very large configuration space. The focus in policy making lies on prevention and insight, and for these aims simulation and surrogate modelling are essential. A theoretical model cannot approximate within a reasonable error margin the complexity of a real world process, while simulation, with a configurable set of parameters mimicking the real world process, can. The simulator is configured to model a measles outbreak in a population of 5e5 in the city of Antwerp, Belgium. With immunization for this disease a worldwide concern we would like to obtain a surrogate model that can offer policy makers insights leading to preventative measures. Of vital importance here is the immunization aspect. Our research question for this case is : How does the immunization fraction influence the outbreak of measles? We investigate this use case within the context of this work, that is, we focus on the convergence characteristics of the process evolving the model rather than the domain specific implications of the surrogate model itself. We are interested in the value of the surrogate model at an intermediary stage in the process. How closely does this model exhibit the same trends as the underlying process? This relation is vital in order to justify our usage of partial results in the feedback loop between practitioner, simulator and regression tool. A single simulation instance is computationally expensive. We would like to construct a surrogate model that approximates the simulator. A surrogate model can offer insights into the underlying process that the resulting data cannot. Symbolic regression offers a white box model in addition to this advantage. We can use symbolic regression to obtain such a model, but in order to do so we need to obtain input and output data. Generating all simulation output in sequence leads to significant downtime for the practitioner. Using our incremental support detailed we offer the practitioner partial results during this downtime. These results can be used by a domain expert to modify the design of experiment instead of having to wait until the entire process has completed. Our tool is able to reuse the partial results as seeds for new runs. We will investigate if this approach can lead to an improved model. The value for the practitioner in this approach is twofold : incremental informative results are offered during an otherwise inactive time allowing for a feedback loop with the simulator, and a possible improvement in the final model can be obtained by seeding our regression tool.

Design of experiment We would like to have a space filling design that maximizes the sampling of the parameter space while minimizing the number of evaluation points. In this experiment we apply a Latin Hypercube Design (LHD). In this work we will use the Audze-Eglais [4, 3, 2] (AE) LHD, which uses the Euclidean distance measure but in addition obtains a uniform distribution of the individual points. The AE LHD is based on the concept of potential energy between design points, a measure based on the inverse of the euclidean distance.

$$E^{AE} = \sum_{i=0}^{p-1} \sum_{j=i+1}^{p-1} \frac{1}{d_{ij}}$$

The potential energy measure is minized in AE.

4.1.1 Experiment

Simulator configuration We construct a DOE with 3 dimensions, 30 points in total using the tool introduced in the work of [6]. The following are the parameters used:

- Basic reproduction number (R0) : the number of persons an infected person will infect, [12-20]

- Starting set of infected persons (S) : Number of persons in the population that is an infected person at the start of the simulation, [1-500]
- Immunity fraction (I) : Fraction of the population that is immune to the disease. [0.75, 0.95]

The output parameter represents the attack rate, measured as the rate of new cases in the population at risk versus the size of the population at risk. For each parameter we obtain 30 points uniformly chosen in their range. These are then combined in the DOE. The simulator is run once for each configuration.

Symbolic regression configuration We run the experiment with a population size of 20, 30 phases with 60 generations per phase, (10,20,30) datapoints for 3 features, an initialdepth of 3 and maximum depth of 6. The 4 best expressions of each phase are archived. We compare 3 approaches. First we run the CSRM tool on the entire dataset. This is the classical approach, the tool is not seeded and so starts a blind search. In a real world setting this would mean waiting until the simulator has run all 30 configurations. In our second approach we split the data into incremental sections. After 10 configurations have completed we start the tool on this dataset. The best 4 results are saved to disk, then we run the tool with the result of 20 configurations and use the results from the previous run as a seed. The overlap between the two datasets will influence the effects of the initialization problem. Finally we use the results of the 20-point dataset as a seed for the 30 point run. The cost of running the 10 and 20 point runs to use as seed for the 30 point run is similar to the cost of the 30 point run. To ground the comparison our last approach runs the tool on the data from 30 configurations with double the amount of phases. This means that it has approximately the same number of fitness evaluations as the 10-20-30 combination. We compare all three to see which gains are made and at what cost. We run the experiment distributed to observe the change in convergence characteristics using the seeds of the 10-20-30 combination to combine the distributed and incremental features of our tool.

4.2 Results

4.2.1 Fitness improvement

We compare both seeded runs and the extended run with the normal 30-point run. In Figure 4 we see that the fitness is improved by using the best results of the previous run on a partial data set. We have deliberately split our data set in such a was as to expose a risk here. If we run the tool with 20 datapoints seeded by a run of 10 datapoints, we see that the validated fitness actually decreases compared to a non seeded run. The ratio between new and known data is too large, leading to overfitting. If we seed the best results from the 20-point run into a 30 point configuration we see that both the training and validated fitness values significantly improve. The 30 point run with 60 phases has the same computational cost as the 10-20-30 runs combined, but gains little to nothing in convergence. We see that convergence is slowing, with training fitness improving by a factor of 1.1, but validation fitness worsens. This is a typical example of overfitting. The combined 10-20-30 run increases validation fitness with a factor of 1.13.

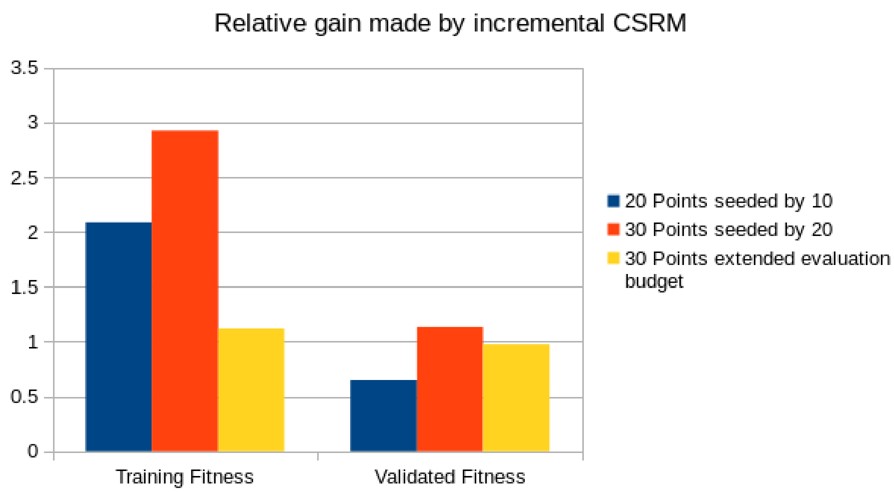


Figure 4: Incremental fitness gain in CSRM.

4.2.2 Convergence behavior

Fitness distribution In Figures **??**,**??**,**??** we see how the convergence process evolves over time. Note the clustering patterns in the incremental runs. When seeded with solutions of previous runs we introduce information that the algorithm otherwise would have to discover. These seeds have fitness values somewhere between the randomly generated and optimal samples. Applying operators on them leads to a niching effect visualized in the fitness distribution. Seeding will not guarantee this effect, it is a possibility depending on how the seeds fit in the trajectory in the search space that the algorithm generates during its execution. If we compare the fitness plots we clearly see that the seeded process has a different distribution compared to the unseeded runs. Around generation 1000 the convergence rate slows. Doubling the phases has little effect on the fitness distribution.

4.2.3 Distributed

We seed a distributed run with the results of the 10/20/30 run and compare the topologies in terms of fitness improvement and speedup. We run the same configuration as before with 25 processes, we use as seeds the 4 best expressions from the 10/20/30 run, and use Tree, Grid, Random and Disconnected topologies.

Results In Figure 5 we compare the gain in fitness on the validation data for the tree, grid and randomstatic topologies compared to the disconnected topology. We can clearly see that the diffusion in the grid topology leads to the highest gain, followed by the tree topology. Interestingly enough, the random topology scored worse than the disconnected topology. This can occur when a local optimum is communicated early to

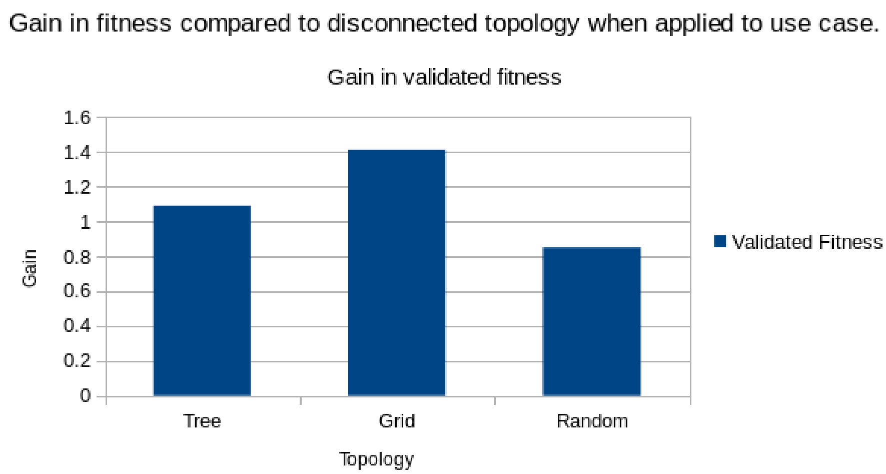


Figure 5: Incremental distributed CSRM applied to use case.

the other processes which then dominates the remainder of the process. The effect on the runtime is measured in Figure 6. We see that the tree topology has minimal overhead and runs nearly as fast as the disconnected topology where no synchronization or communication overhead is present. The grid topology suffers a 2x performance penalty and the random topology finds the middle ground between the two. During the experiment we observed that the processes in the tree and disconnected topologies varied as much as 4 phases. This is what we expected, in this tree topology the distance between two processes is at most 4 (depth of a 25-node binary tree). This is an important observation, if we increase the number of processes the tree topology will actually scale better. The delay tolerance allows the tree topology this scaling effect.

Statistics So far we have focussed on the end results only of the symbolic regression process. Our tool measures several statistics on the convergence characteristics of the process which we will discuss now. We are interested in the convergence characteristics of the different processes and how the evolution of the fitness values calculated on the training data correlate with those of the validation data. We use the pearson r correlation coefficient (for populations) as we do for the fitness calculation itself:

$$d(f_{training}, f_{validation}) = 1 - |r(f_{training}, f_{validation})|$$

The measure is 0 for perfect correlation, 1 for no correlation. A value of 1 in this setting indicates overfitting, a value of 0 indicates that the predictive value of the generated models is perfect. In other words, the validation data are fit perfectly by the generated expressions even though the expressions have been evolved without those data points. In Figure 7 we can clearly see how the topology affects the correlation between fitness values on training and validation data. Our validation configuration

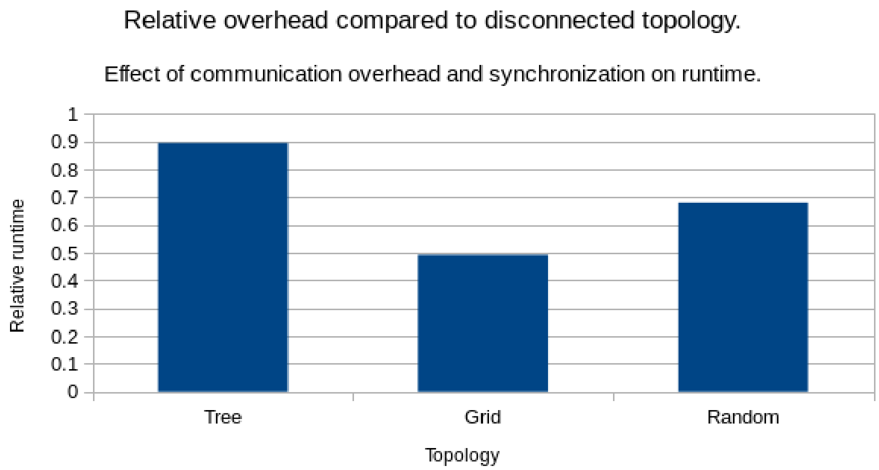
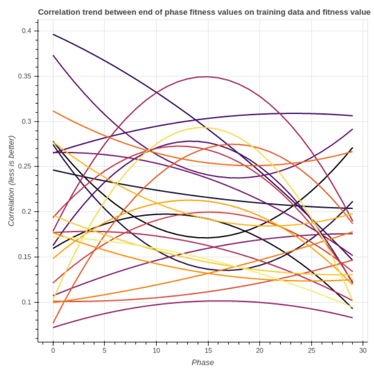


Figure 6: Runtime impact of synchronization and communication overhead.

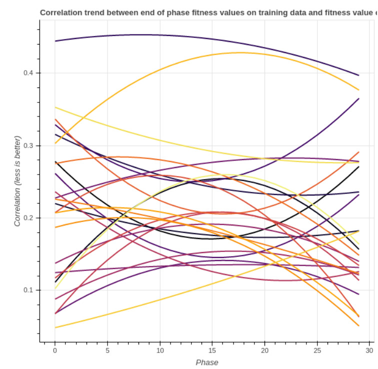
introduced in section **??** promotes this behavior, the group of processes approximates k-fold cross validation. In a fully connected topology with minimal distance such as the grid, the correlation behavior will tend to the same value. We see a similar effect for the random topology. The disconnected topology has a near uniformly distributed correlation, there is no communication between the processes so they cannot train on the validation data. This is a parallel execution of 25 sequential processes with differing seeds and data. The tree topology finds the middle ground, as we descend toward the leaves the amount of shared information increases and with it the coverage of the validation data. We can also see from the plots how the process evolves over time. The grid and random topology quickly stabilize without significant gain. The tree topology has a subset of processes that starts to converge to a lower correlation value.

4.3 Resulting Model

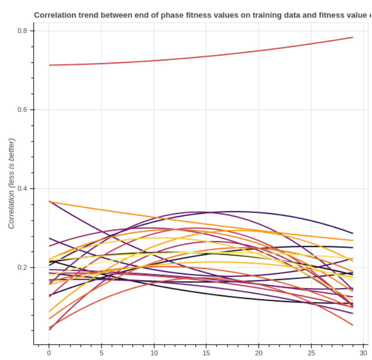
We select the best expression returned by the distributed application of CSRM with a tree topology, given its benefits in runtime and scaling. The distributed run is seeded by the 10/20/30 run. The resulting expression has a fitness value of 0.039 on the full data set. While this value is low, it is still 10 orders of magnitude removed from the optimal. This expression therefore represents an intermediate result and gives us an indication of the value partial results can offer. We use response plots for each parameter in order to isolate the effect each parameter has. We vary each of the parameters while keeping the other two constant. For the constant value we select the midpoint of the range. We then observe the effect on the attack rate. It is important to note that the range of the attack rate is [0,1]. We observe 2 important effects. First, our model produces an attack rate outside of the valid range of [0,1]. There is a scaling factor of 10 between the output of the model and the actual output data from the simulator. This



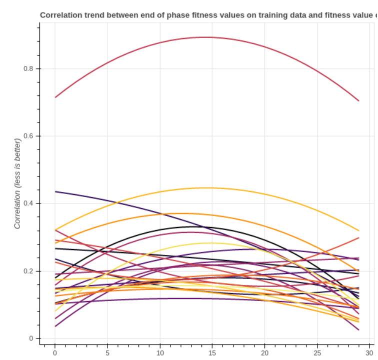
(a) Correlation behavior without communication.



(b) Correlation behavior with tree topology.

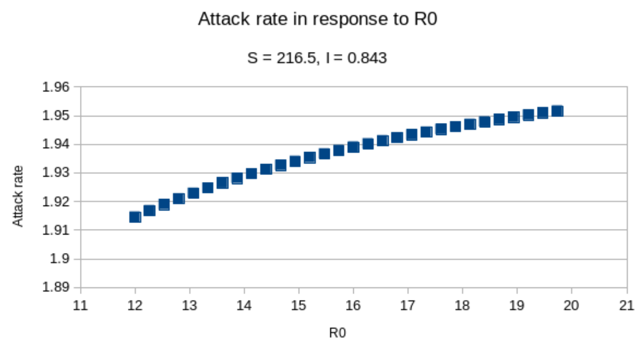


(c) Correlation behavior with grid topology.

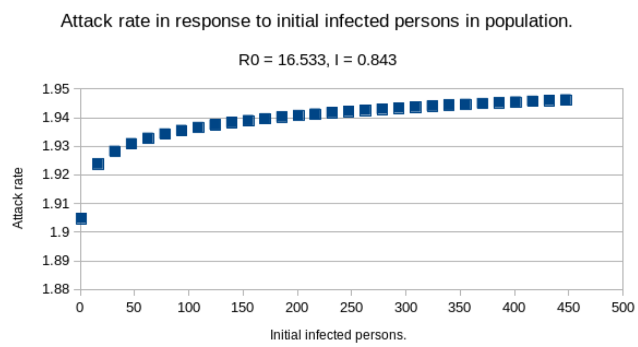


(d) Correlation behavior with random topology.

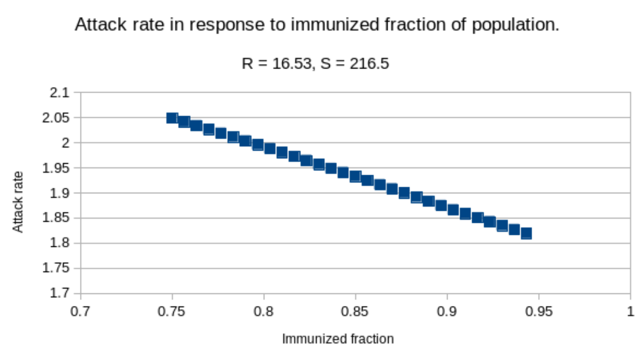
Figure 7: Use case : Effect of topology on fitness correlation of distributed processes.



(a) Response of attack rate to R.



(b) Response of attack rate to S.



(c) Response of attack rate to I.

Figure 8: Response plots of intermediate surrogate model.

is simply due to the fact that convergence is still in an intermediary phase. An important observation here is that our tool evolves the model based on 30 data points and not the full factorial design. This means that the response plots will use the model to evaluate points that are not necessarily available to our tool to train on. Second, the trend in the response plots is in line with what we expect to see in such a surrogate model. When R0 increases the attack rate increases, which is in line with theoretical and empirical results. A similar trend is visible with the initial number of infected persons, where R0 shows a logarithmic response. Finally, as the immunization fraction increases we see a negative linear response in the attack rate. We have chosen this suboptimal surrogate model to demonstrate that while the exact values of the attack rate are not yet correctly modelled, the expected trends are. This conclusion is vital to justify our incremental approach. We can see that surrogate models will focus on matching the trend first, rather than matching individual points. This is in part due to our usage of the Pearson R correlation coefficent as a basis for the fitness function.

5 Related Work

Enrique Alba's book [1] on parallel metaheuristics is the reference work for the field. It provides a broad overview of the challenges and advantages of parallel metaheuristics. SR can be implemented using a parallel metaheuristic such as GP. A Python toolbox for evolutionary algorithms has been developed [5], not specifically directed at symbolic regression but as a repository of evolutionary algorithms in a distributed context. An interesting parallel GP SR implementation [12] introduces a random islands model where processes are allowed to ignore messages, contrary to our approach. The authors argue that this promotes niching, where 'contamination' of locally (per process) fit individuals could otherwise introduce premature convergence. The clear advantage of such a system is speedup, since no process ever waits on other processes. Another difference is the message exchange protocol. Whereas our tool exchanges messages after each phase, their tool uses a probability to decide per process if messages are sent or received interleaved with the generations. Such a setup allows for a heterogeneous set of processes. A different approach is shown in [11] where a master slave topology is used in combination with a load balancing algorithm in order to resolve the imbalance between the different slaves executing uneven workloads. The slaves do not form separate processes, they are assigned a subset of the population and execute only the fitness function. The selection and evolution steps are performed by the master process. This a a fine grained approach, and while it offers a speedup in comparison with a sequential GP SR implementation it does not increase the coverage of the search space. In Distributed Genetic Progamming (DGP) [10] a ring and torus topology are used. The two way torus topology is similar to our grid topology. The study finds that sharing of messages is essential to improve convergence but that the communication pattern is largely defined by the problem domain. It concludes that diffusion is a more powerful technique compared to partitioning. In partitioning no communication between subgroups is possible, which can protect against premature convergence.

6 Conclusion

7 Future work

Using our distributed architecture it is possible to give each process a different optimization algorithm, creating a heterogeneous set of communicating algorithms. This has two advantages. First, it allows for comparison of different optimization algorithms within the same framework. Second, it would make the SR tool more robust. We know from the NFL theorem that no optimization algorithm is optimal for all optimization problem instances. A cooperative set of optimizations algorithms could offer an optimal solution for all problem instances by balancing the disadvantages and advantages of each algorithm.

Topologies The inverted tree topology, where the root is a sink and leaves are sources, is an interesting alternative to the original tree. Future work will evaluate other communication strategies such as random sampling. A random tree topology could offer a balance between convergence and speedup. This approach would aim to combine the advantages of a stochastic approach with the speedup gains of the tree structure itself.

Incremental DOE Dividing the DOE generated input points into sections and using them in the regression tool can lead to issues regarding the structural properties of the design. While our results indicate that the model obtained from an incremental run has a higher quality compared to that of a unseeded run, this does not exclude the possibility that the seed we used corresponds with a biased coverage of the parameter space. An alternative approach would be to increment not datapoints but parameters. The Latin Hypercube Design maintains its structural properties when parameters collapse or are removed.

References

[1] ALBA, E. *Parallel Metaheuristics: A New Class of Algorithms*. Wiley-Interscience, 2005.

[2] AUDZE, P., AND EGLAIS, V. New approach for planning out of experiments. *Problems of dynamics and strengths 35* (1977), 104–107.

[3] BATES, S., SIENZ, J., AND TOROPOV, V. Formulation of the optimal latin hypercube design of experiments using a permutation genetic algorithm. In *45th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference* (2004), p. 2011.

[4] BATES, S. J., SIENZ, J., AND LANGLEY, D. S. Formulation of the audze–eglais uniform latin hypercube design of experiments. *Adv. Eng. Softw. 34*, 8 (June 2003), 493–506.

[5] FORTIN, F.-A., DE RAINVILLE, F.-M., GARDNER, M.-A. G., PARIZEAU, M., AND GAGNÉ, C. Deap: Evolutionary algorithms made easy. *J. Mach. Learn. Res. 13*, 1 (July 2012), 2171–2175.

[6] HUSSLAGE, B. G. M., RENNEN, G., VAN DAM, E. R., AND DEN HERTOG, D. Space-filling latin hypercube designs for computer experiments. *Optimization and Engineering 12*, 4 (2011), 611–630.

[7] KORNS, M. F. *Accuracy in Symbolic Regression*. Springer New York, New York, NY, 2011, pp. 129–151.

[8] KORNS, M. F. *A Baseline Symbolic Regression Algorithm*. Springer New York, New York, NY, 2013, pp. 117–137.

[9] KOZA, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[10] NIWA, T., AND IBA, H. Distributed genetic programming: Empirical study and analysis. In *Proceedings of the 1st Annual Conference on Genetic Programming* (Cambridge, MA, USA, 1996), MIT Press, pp. 339–344.

[11] OUSSAIDÈNE, M., CHOPARD, B., PICTET, O. V., AND TOMASSINI, M. Parallel genetic programming: An application to trading models evolution. In *Proceedings of the 1st Annual Conference on Genetic Programming* (Cambridge, MA, USA, 1996), MIT Press, pp. 357–362.

[12] SALHI, A., GLASER, H., AND DE ROURE, D. Parallel implementation of a genetic-programming based tool for symbolic regression. *Inf. Process. Lett. 66*, 6 (June 1998), 299–307.

[13] WILLEM, L., STIJVEN, S., TIJSKENS, E., BEUTELS, P., HENS, N., AND BROECKHOVE, J. Optimizing agent-based transmission models for infectious diseases. *BMC bioinformatics 16*, 1 (2015), 183.

[14] WILLEM, L., STIJVEN, S., VLADISLAVLEVA, E., BROECKHOVE, J., BEUTELS, P., AND HENS, N. Active learning to understand infectious disease models and improve policy making. *PLOS Computational Biology 10*, 4 (04 2014), 1–10.