

Citrine Informatics Technical Challenge

Data and AI Research Engineer

Version: e6b4a7f0da67f298a09e9550d848bfb86bbc028b

The following technical challenge is meant to be representative of the math and engineering required to solve the problems in scope for our group. There is no time limit, but the problem is designed to take approximately 10 hours so please do not spend much more time on it than that! In addition to your code, please include a brief write-up that explains your solution.

software

One of the core capabilities of Citration is the ability to efficiently sample high dimensional spaces with complex, non-linear constraints. In this challenge, we are asking you to efficiently generate candidates that systematically explore as much of the valid space as possible.

Problem

The “API” of the challenge is file based: you must deliver a script that can be run as:

```
./sampler <input_file> <output_file> <n_results>
```

along with installation instructions. The input file starts with a single line header that gives the dimensionality of the problem, which is defined on the unit hypercube. The next line is a single example feasible point. The remaining lines are a list of constraints as python expressions containing `+`, `-`, `*`, `/`, and `**` operators. They have been transformed such that they all take the form `g(x) >= 0.0`. We’ve included several example input files for reference. We’ve also included a python class that can parse the input file and produce a function that evaluates correctness. The output file should contain a list of vectors (space delimited within the vector; one vector per line). We will evaluate for `n_results = 1000`, and expect the execution to take less than 5 minutes.

Write-up

In addition to the library implementation, we ask that you also include a write-up of your approach. This part of the challenge reflects the fact that, as a team, we often find ourselves spending as much time convincing ourselves and stakeholders that a solution yields quality results as we do actually designing and building the solution itself. Ultimately, our goal when reading the write-up is to be persuaded that the approach is sound and that the performance and limitations are understood.

Evaluation

We will run your script on a held-out constraint list. The evaluation criteria has a few components. Keep these in mind when planning how to focus your effort:

- Write-up: the analysis is compelling - we’d be able to use it to decide whether or not to use this solution in production.
- Correctness: all of the produced vectors satisfy all of the constraints (both on the provided examples and also a set of hold-outs we test your solution against).
- Solution quality: the vectors cover as much of the valid space as possible. You can think of this as them being “spread out” in the Euclidean space in which the valid space is embedded.

You are welcome to use whatever programming language/environment you’d like, as long as we are able to easily install and execute your script on our test constraint list.