1.1 Testing Heuristics, we have it all wrong

Imagine, one day you want to buy a horse, and not just any horse, but the best horse in the world. You want that single horse, that will make you win the gold medal at the London 2012 Olympic games. Hence, you search all newspapers and reports of the last years, and you quickly find out that this special horse is called Secretariat. Secretariat is one of only two horses that ever broke the two minute frontier in the Kentucky Derby, and is widely acknowledged to be the fastest horse in the history of horse racing. And, more importantly, he is for sale.

A few weeks pass, and together with the best horse, you have invested in the best trainers, the best jockey, the best stables and the best food, such that all parameters that might influence the performance of your horse are optimal. Nothing can get between you and the gold medal, you are sure of that.

Finally the day is there, the Olympic anthem is played, and the flag is hoisted into the air. Your horse is next, its name resounds through the stadium. And then ..., the music starts. “Dance of Devotion”; the same song Anky van Grunsven had, when she won her gold medal in 2008, and now its your turn.

Of course, you became last; you bought a race horse and let it compete in a dressage. An attempt that is not only utterly useless, but also shows a lack of knowledge about horses. Still, in the field of evolutionary algorithms, this happens all the time. A horse race is the common practice; algorithms compete on different kinds of benchmarks and
the one that reaches the finish line first is the winner. And, just as in the example, this winner is then used in a wide range of applications that have nothing to do with the original benchmark. In other words, the race horse is put in the dressage stadium. Even if this special race horse is capable of performing a gold-medal dressage, it is often the case that it was fully prepared for running, while a dressage would require a different trainer, a different jockey and possibly a different diet and horseshoes.

Therefore, there is no such thing as the ‘best horse’. Which one is the best fully depends on what you intend to do with it, and how it is prepared. The same thing holds for evolutionary algorithms; letting them compete in a horse race and awarding the champion as being ‘the best horse’ is just doing it wrong.

1.2 Parameter Tuning and Scientific Testing in Evolutionary Algorithms

In this thesis we address this issue of evaluating and comparing evolutionary algorithms, and more specific, the role of parameter tuning in this context. The subject of this thesis is heavily inspired by the 1995 publication by J.N Hooker titled ”Testing Heuristics: We Have It All Wrong”[70], in which Hooker criticizes the current paradigm of testing heuristics (such as evolutionary algorithms). He concludes that we have confused research with development, and focused on showing that our new algorithm performs better than the current state-of-the-art, rather than why. In this thesis we take the same stance, but focus on how parameter tuning can be used for a more ”scientific approach” to testing heuristics.

To this end, we first give a general introduction of designing evolutionary algorithms in Chapter 1. There we address which questions arise and what decisions need to be made in such a process. Furthermore, we introduce the different terms related to algorithms, parameters and parameter tuning that are used throughout this thesis. We conclude this chapter by identifying two different approaches for comparing algorithms, namely competitive and scientific testing, and relate these to the issues that arise when designing algorithms. Most parts of this chapter originate from [43] and [41].

In Chapter 2, we give an extensive overview of the current approaches to automated parameter tuning, and we introduce three taxonomies by which these can be classified. We classify each of the approaches by the taxonomies and compare them based on theoretical arguments. The performance of the state-of-the-art approaches are then thoroughly tested by means of a big experiment, in which each of them needs to tune an advanced algorithm on a well-known test suite. We end this chapter with some
conclusions and recommendations about the use of the parameter tuners in different situations. Most parts of this chapter originate from [43], [104], [130], and [132].

Chapter 3 contains the results of three case studies that are conducted, each of which had a different goal. They are used to illustrate how parameter tuning can be used to improve the current practice of competitive testing for benchmarking, competitive testing for comparison, and the use of parameter tuning in scientific testing. Most parts of this chapter originate from [126], [128], [127], and [58].

In the fourth chapter, we discuss possible future use of parameter tuning methods in different areas. We show how it can be used for numerical optimization of noisy fitness landscapes, and the role that parameter tuning methods can have in parameter control. Most parts of this chapter originate from [43] and [83].

The thesis ends with a general discussion and conclusions about parameter tuning and scientific testing. It summarizes the thesis, and addresses the question that is fundamental to it: if we are all doing it wrong now, what is the correct way?