Optimization to measure performance in the Tailorshop test scenario — structured MINLPs and beyond

Sebastian Sager **Carola M. Barth** **Holger Diedam**
Michael Engelhart **Joachim Funke**

Interdisciplinary Center for Scientific Computing
Department of Psychology
University of Heidelberg
INF 368, 69120 Heidelberg, Germany

sebastian.sager@iwr.uni-heidelberg.de

Abstract

Obtaining objective means to measure performance is of crucial importance in the research field Complex Problem Solving. While for traditional tests like the Tower of Hanoi the correct solutions were known, this is more difficult for modern, complex, simulation-based test scenarios, as the *Tailorshop*. We derive a problem class of non-convex mixed-integer nonlinear programs (MINLPs) which stem from such economic test scenarios. In a round based scenario participants need to make decisions. A posteriori a performance indicator is calculated and correlated to their ability of emotion regulation. We solve altogether 2088 optimization problems with different size and initial conditions. They are based on real world experimental data from 12 rounds of 174 participants. The goals are twofold: first, from the solutions we gain additional insight into a complex system, which facilitates the analysis of a participant's performance in the test. Second, we propose a methodology to automatize this process by providing a new criterion based on the solution of a series of optimization problems. We disprove the assumption that the "fruit fly of complex problem solving", the *Tailorshop* scenario that has been used for dozens of published studies, is not mathematically accessible. By providing a detailed mathematical description and the computational tool *Tobago* [**12**] for an optimization-based analysis we hope to foster further interdisciplinary research between psychologists and applied mathematicians and provide a source for benchmarking of MINLP solvers.

Keywords: mixed integer programming, nonlinear programming, cognitive psychology.

1. Introduction

Psychologists define complex problem solving as a *high-order cognitive process*. The complexity may result from one or several different characteristics, such as a coupling of subsystems, nonlinearities, dynamic changes, intransparency, or others [**6**]. The main intention of the research field *complex problem solving* of human beings is the desire to understand how certain *variables* influence a solution process. In general, *personal and situational variables* are differentiated. In our study we analyze the personal variable *emotion regulation*. Other interesting personal variables are *working memory*, *amount of knowledge*, and *intelligence*.

Psychologists have been working in the research fields of problem solving for approximately 80 years. Since the 1970s and 1980s also computer-based test scenarios are in use. The overall idea, compared to early works in problem solving, is still the same: one evaluates the performance of a participant by calculating an *indicator function* and either correlates it to personal attributes or analyzes the influence of different experimental conditions for groups of participants. The main difference is that for the early test scenarios the correct solution is known at every stage. For more complex scenarios the performance evaluation is not so straightforward. The availability of an objective performance indicator is an obstacle for analysis and it has often been argued that inconsistent findings are due to the fact that

> *"... it is impossible to derive valid indicators of problem solving performance for tasks that are not formally tractable and thus do not possess a mathematically optimal solution. Indeed, when different dependent measures are used in studies using the same scenario (i.e., Tailorshop* [**7, 13, 11**]*), then the conclusions frequently differ."*

as stated by Wenke and Frensch [**15**, p.95]. The *Tailorshop* is sometimes referred to as the "Drosophila" for problem solving researchers [**9**] and thus a prominent example for a computer-based test scenario. In Section 2 we will derive a mathematical model for the *Tailorshop*. In Section 3 we will discuss mathematically optimal solutions, and finally formulate a valid indicator function in Section 4.

To our knowledge, numerical optimization methods have only scarcely been used for the analysis of participants' decisions. The general approach to compare performance to optimal solutions has been discussed by [**10**]. However, the authors do not provide a mathematical model for their test scenario *EPEX*. Hence, they need to use the software as a black box for brute-force simulation or derivative free strategies, such as Nelder-Mead. Such strategies result in significantly higher computational runtimes, give less insight, and have poor theoretical convergence properties.

2. Tailorshop MINLP Model

The *Tailorshop* has been developed and implemented as a test scenario in the 1980s by Dörner [**6**]. It has been used in a large number of studies. Also comprehensive reviews on studies and results in connection with the *Tailorshop* have been published, e.g., [**8**].

A participant has to take economic decisions to maximize the profit of a small company, specialized in the production and sales of shirts. The scenario comprises twelve rounds (months), in which the participant can modify infrastructure (employees, machines, distribution vans), financial settings (wages, maintenance, prices), and logistical decisions (shop location, buying raw material). As feedback he gets some key indicators in the next round, such as the current number of sold shirts, machines, employees, and the like. Arrows next to the indicators show if the value increased or decreased with respect to the previous round.

We derive a mathematical formulation as an optimization problem. The basic idea is that for different initial values (the current state in round n_s of a participant's test run) the optimal solution for the remaining $N - n_\mathrm{s}$ rounds can be calculated. The optimal solution can then either be used for a manual comparison and analysis of the participant's decisions, Section 3, or for an automated indicator function, as discussed in Section 4.

The *Tailorshop* has been developed as a test scenario in `GW-Basic` code. On the basis of this code we derived a mathematical optimization problem for a participant and month $0 \le n_\mathrm{s} < N$ as

$$
\begin{array}{lllll}
\max\limits_{x,u,s} & F(x_N) & & & \\
\text{s.t.} & x_{k+1} & = & G(x_k, u_k, s_k, p), & k = n_\mathrm{s} \dots N-1, \\
& 0 & \le & H(x_k, x_{k+1}, u_k, s_k, p), & k = n_\mathrm{s} \dots N-1, \\
& u_k & \in & \Omega, & k = n_\mathrm{s} \dots N-1, \\
& x_{n_\mathrm{s}} & = & x^p_{n_\mathrm{s}}. &
\end{array}
\tag{44.1}
$$

The model is dynamic with a discrete time $k = 0 \dots N$, where $N = 12$ is the number of rounds. The control vector $u_k = u(k)$ has 15 (or 13 when van purchase is fixed) entries for each $k = 0 \dots N-1$ corresponding to the decisions the participant can make in the test. The vector of dependent state variables $x_k = x(k)$ comprises 16 entries. We define

$$(x^\mathrm{p}, u^\mathrm{p}) = (x^\mathrm{p}_0, \dots, x^\mathrm{p}_N, u^\mathrm{p}_0, \dots, u^\mathrm{p}_{N-1})$$

to be the vector of decisions and state variables for all months of a participant. Certain entries $x^p_{n_\mathrm{s}}$ enter (44.1) as fixed initial values. The goal is to find decisions u_k that maximize the overall balance at the end of the time horizon. The objective function is given by $F(x_N) = x^{OB}_N$. The resulting problem is a nonconvex mixed-integer nonlinear program with n_s–dependent dimension.

3. Optimization and numerical results

We want to solve a series of optimization problems of the form (44.1) for different participant data that has been obtained experimentally.

3.1. Implementation

To be able to analyze and visualize the data in a convenient way, to have a simulation environment for own studies, and to be able to automatize the optimization of all $2088 = 174 \cdot 12$ problems, we implemented the software framework *Tobago* [**12**]. It is publically available under an open source license, includes a GUI, and may as well be used for experimental setups. This data generation and analysis tool can be hooked to a variety of optimization solvers. Currently the software supports *AMPL* interfaces. This allows for the usage of solvers from the *COIN-OR*

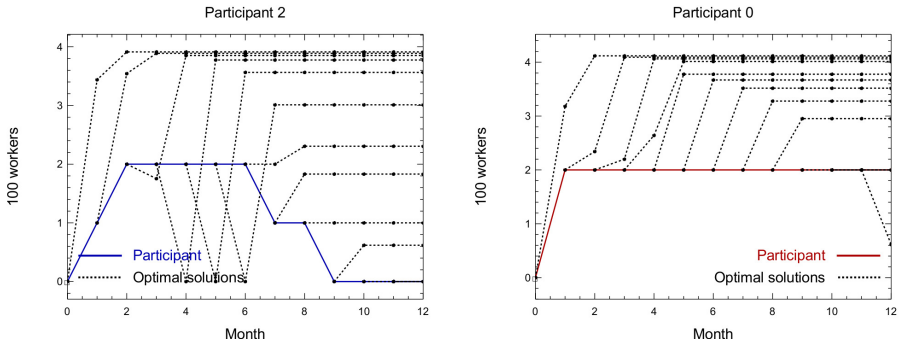

Figure 1. Top row: state variable $x_k^{W_{100}}$ that indicates how many workers for the 100 machines are employed. The left and right column show the results for two different participants. For both the optimal strategy is to have a fixed number of 0 to 4 workers which is decreasing as n_s increases. Note that the values are solutions of the relaxed problem where also non-integer values are possible.

initiative, which are also available under a public license. In this study we use the global solver *Couenne* [**3**] and the local solvers *Bonmin* [**4**] and *Ipopt* [**14**]. We used the currently latest stable version 0.2.2 of *Couenne*, and for better comparability the versions 1.1.1 of *Bonmin* and 3.6.1 of *Ipopt* it is interfaced with. For all solvers we used the default settings exclusively and the MA27 sparse solver for numerical linear algebra.

It turns out, however, that the size and complexity of the problems presented in this paper leads to extremely long runtimes of the global solver and can only be used on a small subset of the problems. We present a problem-specific cut to avoid bad local minima and guarantee monotonicity of the analysis function that builds on the locally optimal objective function values.

3.2. Optimal Solutions

In total, 2088 optimization problems have been solved. Depending on the value of n_s in (44.1), each consists of $13(N - n_s)$ control, $16(N - n_s)$ state, and $5(N - n_s)$ slack values. The total number of optimized variables for all 174 participants sums up to

$$n_{\text{var}} \quad = \quad 174 \sum_{n_s=0}^{N-1} 34(N - n_s) = 174 \cdot 2652 = 461448.$$

This many variables are obviously difficult to discuss and visualize comprehensively. As an illustration, in Figure 1 the state variable $x_k^{W_{100}}$ is depicted. It indicates how many workers for the 100 machines are employed at time k.

3.3. Local minima and integer solutions

The optimization problems (44.1) are nonconvex. Depending on initial values for the optimization variables different local minima can be found. Hence one has to use a global optimization solver, such as *Couenne* or one of the solvers listed on [**5**]. As mentioned above, we used three different solvers to obtain solutions. Table 1

S	0	1	2	3	4	5	6	7	8	9	10	11
1	0.15	0.13	0.17	0.11	0.1	0.06	0.05	0.03	0.02	0.02	0.0	0.0
2	1183	264	1552	1464	356	36	5	4	16	3	0.2	0.2

Table 1. CPU times in seconds for the solution of (44.1) for one participant. The columns show the start month n_s. Solver S 1: *Ipopt* for the relaxation of (44.1). Solver S 2: *Bonmin*. The global solver *Couenne* could only solve the problem for $n_s = 11$ in 3 seconds, for $n_s = 10$ the B&B tree grew too fast.

shows an overview of computational times that have been obtained with *Ipopt* and *Bonmin*. Note that the runtime is not monotonically increasing as n_s is reduced. The reason is that the solution process strongly depends upon the local minima of the relaxations that need to be solved.

The global solver *Couenne* was able to solve (44.1) for $n_s = 11$ in 3 seconds. For the next larger problem, $n_s = 10$, however, no results could be obtained. The solver terminated after processing 600.000 nodes in 7 hours, because the computer ran out of memory. The stack comprised about 2.000.000 open nodes at that time. To reduce the search space, we introduced and tightened the bounds on all variables to extremal values found with the local approaches. However, even with this restriction and a relaxation of all integer variables the same happened, now after 8.800.000 processed nodes with 2.9 million NLPs still on the tree. The best solution at that time was 500497 with the upper bound of 506610 still leaving a certain gap. For comparison: the objective function values found by *Bonmin* and *Ipopt* are 490385 and 500779, respectively. When heuristic non-convexity options `num_resolve_at_root` and `num_resolve_at_node` are used with a value of 1 (or 2) for *Bonmin*, an integer solution with value 500188 (500438) is found after 142 (317) seconds, which is considerably higher than the 0.2 seconds with the standard settings.

Obviously already for one participant data set the computational times are prohibitive for global approaches. For the analysis of all 174 participants we therefore solved 2088 NLP relaxations with the local optimizer *Ipopt*.

A crucial feature of our method is that the *How much is still possible*–function, see Section 4.1, decreases monotonically with n_s increasing. To take this into account, we exploit this knowledge in our a posteriori analysis. We define

$$(x^*, u^*, s^*) = (x^*_{n_s}, \ldots, x^*_N, u^*_{n_s}, \ldots, u^*_{N-1}, s^*_{n_s}, \ldots, s^*_{N-1})$$

as a locally optimal solution obtained by solving problem (44.1) for month n_s.

We initialize the variables for problem (44.1) with a feasible solution. To avoid local minima with a worse performance, we add the additional cut

$$x_N^{OB} \quad \geq \quad x_N^{*,OB} \tag{44.2}$$

to (44.1).

Computational experience shows that the primal-dual interior point solver we are using cannot exploit the initialization to its full extent and in many cases *Ipopt* converged to locally infeasible points although it started from a primally feasible one. Future studies should therefore include active set based solvers. For this study we iterated in an inner loop with random initializations until the objective function cut (44.2) was fulfilled for all problems.

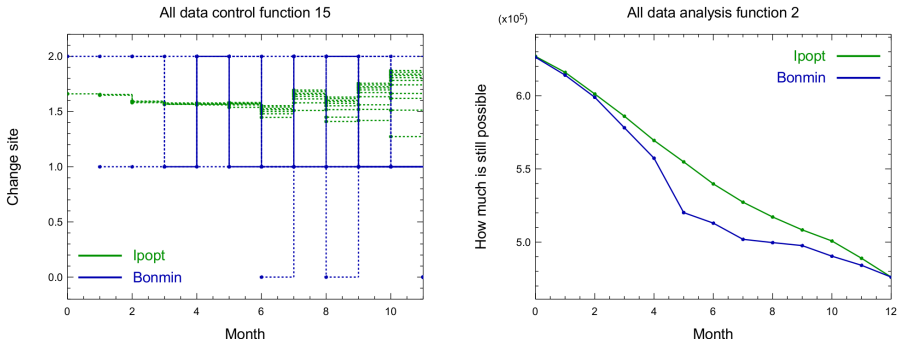

Figure 2. Left: optimal choices of site for one participant and all start months n_s, calculated with *Ipopt* (green, relaxed values between 1.1 and 1.9) and *Bonmin* (blue, integer values of 0, 1, and 2). Right: *How much is still possible*–function for one participant, calculated with *Ipopt* (green, upper curve) and *Bonmin* (blue, lower curve). The integer gap seems to be largest for intermediate values of n_s.

Within our analysis approach, local minima can lead to a violation of the goal to have an objective measurement for participant performance. Whenever possible, global solvers with a guaranteed, deterministic global minimum should be used. If the size of the problem is still too large for current algorithms and computational platforms, we propose to use relaxations and include the cut (44.2) as a compromise. The difference between participant's performance and global optimum seems to be so far apart compared to the distance between global and local minimum, especially when the cuts (44.2) are used, that the analysis based on a local *How much is still possible*–function should still be valid.

Several of the control variables are restricted to integer values. A comparison of (locally) optimal relaxed and integer solutions shows that some of the variables show typical behavior for most $x^p_{n_s}$, such as the maintenance u_k^{MA} or the purchase of raw material $u_k^{\Delta MS}$. Others, in particular the numbers of machines and workers, the shirt price u_k^{SP}, and the choice of the site u_k^{CS} are more sensitive to local optima and/or the fixation of some of the variables to integer values. Figure 2 shows an example.

3.4. Analyzing Lagrange Multipliers

Using optimization as an analysis tool yields insight on several levels. Structural properties of the problem, e.g., the unboundedness, can be understood. Also the performance of a participant can be compared to the optimal solution, and the *How much is still possible*–function to be discussed in Section 4 delivers a temporal resolution of this performance. But even a more detailed analysis is possible. While an analysis of the *How much is still possible*–function indicates at what rounds the participant made particularly good or bad decisions, the question of what of the decisions contributed significantly to the success or failure remains and might be of importance in a given test scenario.

We propose to combine two concepts. First, the comparison of the participant's decisions at month n_s with the optimal solution, $u^p_{n_s} - u^*_{n_s}$, gives a global indication of differences in the controls. However, it is unclear from this comparison how

significant a single deviation is. Therefore we use, second, Lagrange Multipliers for the participant's decisions to measure the effect on the objective function. We augment problem (44.1) with the additional constraint

$$u_{n_s} = u^{\mathrm{p}}_{n_s} \tag{44.3}$$

Note that necessarily it holds $x^*_{n_s+1} = x^{\mathrm{p}}_{n_s+1}$, hence the augmented problem for month n_s has the same solution as problem (44.1) for $n_s + 1$. Hence there is no need for additional optimization problems to be solved. The advantage is that an optimization code will also calculate the dual variables or *Lagrange multipliers* λ_{n_s} for the constraints (44.3). It is well known that the Lagrange multipliers indicate the shadow prices, i.e., how much the objective function will vary if the corresponding constraints were relaxed, assumed that the active set stays constant.

4. A correct indicator function for Tailorshop

We propose to use the solutions of (44.1) for all n_s as an indicator function for the performance of a participant. The approach described in Section 4.1 is generic and should also be used for other test scenarios in complex problem solving in the future. In Section 4.2 we describe the results we obtained by using this indicator function for a psychological study.

4.1. How much is still possible

On an individual basis, the performance of every participant can be better understood by a comparison with optimal solutions as illustrated in Section 3. For an evaluation of large data sets that shall be related to characteristics of participants or experimental setup, an automatization and a reduction to an indicator function is necessary. To measure performance within the *Tailorshop* scenario different indicator functions have been proposed in the literature, e.g., the evolution of profit or overall worth of the tailorshop. An obvious drawback of comparing the results of several rounds with one another is that the main goal of the participant is to maximize the value at the end of the test, not necessarily in between.

Hence it might happen that decisions are analyzed to be bad, while they are actually good ones and vice versa. To overcome this problem, we propose to compare the decisions to mathematically optimal solutions. In a certain analogy to the cost-to-go-function in dynamic programming, the optimal objective function values for *all* rounds yield the monotonically decreasing *How much is still possible*–function. We look at the series of optimal objective function values $F^*(x_N; n_s)$ for $n_s = 0, \ldots, N-1$. By comparing $F^*(x_N; n_s = k)$ with $F^*(x_N; n_s = k+1)$ we obtain the exact value of how much less the participant is still able to obtain, assumed he would take the best solutions from now on.

We conclude that the newly proposed methodology based on the *How much is still possible*–function is more reliable and generally applicable to test scenarios in complex problem solving.

4.2. Impact of Emotion Regulation

In the study 174 data sets have been used, every one from a different participant who had but one try. For 42 of them a *positive feedback* was used in the sense that in every round, regardless of the decisions the participant took, a sum of

20.000 money units (MU) was added to the capital. For 42 participants a *negative feedback* in form of a reduction of 8000 MUs was implemented. These modifications are implemented in the model and readjusted in the a posteriori analysis, of course.

In a previous study [**1**] it was shown that participants who receive a negative feedback perform better than those who receive positive feedback. In our new study we additionally considered the ability to regulate emotion. The psychological results of this study are submitted in a separate paper [**2**] in which also details on the experimental setup can be found. As a main result, an interaction between feedback and emotion regulation could be shown: participants with a high ability of emotion regulation perform better when they get negative feedback, while those with a low ability to regulate their emotions perform bad for negative and good for positive feedback.

Acknowledgments

Financial support of the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences is gratefully acknowledged.

References

1. C.M. Barth and J. Funke. Negative affective environments improve complex solving performance. *Cognition and Emotion*, 2009. (in press).
2. C.M. Barth, J. Funke, and S. Sager. Effects of emotion regulation and affect on problem solving. *Journal of Individual Differences*. (submitted).
3. P. Belotti. Couenne: a user's manual. Technical report, Lehigh University, 2009.
4. P. Bonami, L.T. Biegler, A.R. Conn, G. Cornuéjols, I.E. Grossmann, C.D. Laird, J. Lee, A. Lodi, F. Margot, N. Sawaya, and A. Wächter. An algorithmic framework for convex mixed integer nonlinear programs. *Discrete Optimization*, 5(2):186–204, 2009.
5. M.R. Bussieck. Gams performance world. http://www.gamsworld.org/performance.
6. D. Dörner. On the difficulties people have in dealing with complexity. *Simulation and Games*, 11:87–106, 1980.
7. J. Funke. Einige Bemerkungen zu Problemen der Problemlöseforschung oder: Ist Testintelligenz doch ein Prädiktor? *Diagnostica*, 29:283–302, 1983.
8. J. Funke. *Problemlösendes Denken*. Kohlhammer, 2003.
9. J. Funke. Complex problem solving: A case for complex cognition? *Cognitive Processing*, 2010. (in press).
10. S. Kolb, F. Petzing, and S. Stumpf. Komplexes Problemlösen: Bestimmung der Problemlösegüte von probanden mittels verfahren des operations research – ein interdisziplinärer Ansatz. *Sprache & Kognition*, 11:115–128, 1992.
11. W. Putz-Osterloh. Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg. *Zeitschrift für Psychologie*, 189:79–100, 1981.
12. S. Sager, H. Diedam, and M. Engelhart. Tailorshop: Optimization Based Analysis and data Generation tOol. TOBAGO web site https://sourceforge.net/projects/tobago.
13. H.-M. Süß, K. Oberauer, and M. Kersting. Intellektuelle Fähigkeiten und die Steuerung komplexer Systeme. *Sprache & Kognition*, 12:83–97, 1993.

14. A. Wächter and L.T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.
15. D. Wenke and P. A. Frensch. *Is success or failure at solving complex problems related to intellectual ability?*, pages 87–126. The psychology of problem solving. Cambridge University Press, 2003.

Proceedings of the European Workshop on Mixed Integer Nonlinear Programming

12-16 April 2010 - CIRM - Marseille - France

Pierre Bonami [1]
Leo Liberti [2]
Andrew J. Miller [3]
Annick Sartenaer [4]

[1]LIF, Université de la Méditerranée, Marseille, France. pierre.bonami@lif.univ-mrs.fr
[2]LIX, École Polytechnique, Palaiseau, France. liberti@lix.polytechnique.fr
[3]IMB, Université de Bordeaux 1, France. andrew.miller@math.u-bordeaux1.fr
[4]Dept. of Maths, Université de Namur, Belgium. annick.sartenaer@fundp.ac.be