Structural and Multidisciplinary Optimization manuscript
(authors' post-print)

Automated Optimal Design of a Two-Stage Helical Gear Reducer

Tudose, L. · Buiga, O. · Ştefanache, C. · Sóbester, A.

Received: date / Accepted: date

Abstract The design space of multi-stage transmissions is usually very large and heavily constrained. This places significant demands on the algorithm employed to search it, but successful optimization has the potential to yield considerably better designs than conventional heuristics, at the same time enabling a better understanding of the trade-offs between various objectives (such as service life and overall weight). Here we tackle a two-stage helical gear transmission design problem (complete with the sizing and selection of shafts, bearings, housing, etc.) using a two-phase evolutionary algorithm in a formulation that can be extended to include additional stages or different layouts.

Keywords Evolutionary optimization · Gear train design · Spur gear sets · punctuated equilibria · Multi-objective optimization

1 Introduction

The complexity of the design of multi-stage reducers lies in the strong and often intractable connections between the design variables defining its sub-systems. In other words, an optimal reducer is generally not an assembly of components optimized in isolation, a fact overlooked by many conventional design heuristics. For instance, the impact of a certain choice of gear width and center distance may yield a minimum mass gearing, but the selection of this gearing may cascade through subsequent steps of the design process (sizing of shafts, further stages, bearings, housing, etc.) to ultimately lead to a heavier reducer than if a slight compromise had been made on the choice of that first gearing.

A typical example might be that selecting a smaller than optimal gear diameter (and a correspondingly greater contact width) could yield a somewhat heavier gearing, but a more compact layout and therefore a much lighter housing; it is worth mentioning though that in reality the impact on the overall objective tends to be much less direct and therefore much more obscure than in this example.

Of course, in all but a few trivial cases, it is impossible to tell what that first compromise should have been, let alone what any subsequent choices should have been made with the overall goal in mind, instead of concentrating on the subsystem in hand. The chief reasons impeding a truly 'holistic' reasoning at every step of the design heuristic are the sheer number and the highly non-linear nature of the constraints and the objectives, the large number of design variables and the complexity of the interactions between them. Additionally, analytical models may not be available for these interactions and constraints, precluding higher level analytical calculations that could predict the global effect of local design decisions.

The last two decades have seen an increasing awareness amongst the power transmission design community of the shortfalls of simple trial and error type methods conventionally used to tackle this highly constrained class of design problems and potential replacements have begun to emerge in the shape of expert systems

Tudose, L., Buiga, O. and Ştefanache, C.
Technical University of Cluj-Napoca
Faculty of Machine Building, Dept. of Machine Elements and Tribology, Bd. Muncii 103-105, 400641, Cluj-Napoca, Romania
Tel.: +40 745 640 604
E-mail: Lucian.Tudose@omt.utcluj.ro
Ovidiu.Buiga@omt.utcluj.ro

Sóbester, A.
University of Southampton, School of Engineering Sciences
Southampton, SO17 1BJ, United Kingdom
Tel.: +44 23 8059 2350
E-mail: a.sobester@soton.ac.uk

(Ferguson et al. [6], Abersek et al. [3]), synthesis tools based on spatial grammars (see the Simulated Annealing-driven, grammar based topological
gearbox design tool described by Lin et al. [10]), particle swarm searches (Ray and Saini, 2001 [13]), algorithms based on the modeling of civilizations and societies (Ray and Liew, 2003 [12]), constrained quasi-Newton local searches (see the study by Thompson et al.[15] into the fatigue life versus gearing volume trade-off) and evolutionary algorithms (the work of Li et al. [9] on the application of a fuzzy-controlled genetic search to the optimization of a simple reducer model and the study by Gologlu and Zeyveli [7] for recent examples). In fact, the latter category – headlined by genetic algorithms (GAs) – appears to be the direction of choice at present and there are two key reasons for this.

Firstly, GAs can handle the highly discretised design spaces of transmission systems. Standardisation and the favouring of off-the-shelf (as opposed to purpose-designed) subcomponents are the main reasons for most design variables only being permitted a pre-determined set of discrete values (as we shall see, this is the case with our own application too). Secondly, the full description of a class of reducers (say, that of the two-stage, helical gear family) generally requires a large number of design variables – typically, well over ten – and GAs have a fine track record in the global search of very large design spaces, especially when the computation of the goal function and the constraints is comparatively inexpensive.

The motivation behind the work described here is that evolutionary computing technology has now reached the level where, we believe, it is computationally feasible to consider the automated optimal design of complete reducers. The experiments referenced above have been instrumental in highlighting the importance of using modern global optimization techniques in transmission design (as opposed to conventional, trial and error type design algorithms) even when considering certain subproblems – here we propose to extend the technology to the broader design space of a two stage reducer, whose every element (bearings, seals, shafts, etc.) is subject to change throughout the optimization process.

The industrial relevance of the exercise is ensured by the consideration of all design constraints typically encountered in practice – we bound the design space by a total of 77 constraints categorized into 24 groups. In sections 3 and 4 we discuss this formulation in detail, with section 5 containing an account of the results of its deployment. First, however, we need to introduce a key element of our evolutionary computing methodology, necessitated by the need to handle such a large number of constraints in an efficient manner.

2 An Evolutionary Paradigm

The inevitable paucity of the fossil record makes it rather difficult to estimate the rate of evolutionary change along any given lineage. Nevertheless, it is almost certain that evolution proceeds with varying speed. The extrema of these speed variations are, however, subject to some debate in the evolutionary biology community. It is clear from neo-Darwinian synthesis that large changes (*macromutations*) are almost always deleterious and this is an evolutionary *upper* speed limiting factor. According to the school of thought usually associated with the seminal paper of Eldridge and Gould [5], the *lower* limit is practically zero. That is, they suggest that populations evolve in bursts, which punctuate long periods of equilibrium (*stasis*), when no variations occur.

From the perspective of evolutionary algorithm design it is almost entirely irrelevant whether the theory of punctuated equilibria is correct or not. After all, GAs incorporating spells of *Lamarckian learning* are not made less successful by the fact that the Lamarckian theory of evolution is now known to be incorrect! Similarly, macromutations are often beneficial in evolutionary optimization, reminding practitioners that evolutionary computation is not an exact simulation of nature (nor was it meant to be). Consequently, it is not surprising that the idea of punctuated equilibria has seen steady exposure over the years in the evolutionary algorithms community, in spite of the debate surrounding it in biology. This exposure is associated with two main lines of intellectual inquiry.

Firstly, many practitioners of evolutionary optimization have noted periods of stasis on multi-modal fitness landscapes. The familiar pattern sees the population converging on a local optimum, where a number of generations of stagnation (*metastability*) precede a beneficial mutation, which propels the population into the next, better basin of attraction (see, for example, [11] for a detailed account of the dynamics of this phenomenon on a one-dimensional bistable fitness landscape).

Secondly, and this is the angle we are interested in here, some success has been reported over the years in attempts to actually engineer metastable states in GAs, which would then be followed by bursts of rapid convergence. There is no unique, well-defined template for designing such heuristics; the literature contains a fairly broad range of models that are loosely based on the concept of punctuated equilibria.

Much of the work along these lines is based on multi-population architectures. An example is the integrated circuit design application of Cohoon *et al.* [4], where each subpopulation (*environment*) is allowed to evolve independently for a number of generations (an *epoch*), after which a 'natural catastrophe' causes genetic material to move between subpopulations, followed by another epoch of isolated evolution, etc. Multiple epochs separated by catastrophes can be seen in single population implementations too – see, as an example, the GA of Hamada *et al.* [8], where the catastrophe is represented by a steep increase in mutation rates.

Indeed, a single-population, multi-epoch framework defines the structure of our proposed search strategy too. We depart, however, from the previously mentioned heuristics in the way we control the evolution of the population within each of two distinct types of epoch (or phase). In the first, the population is only subjected to the selective (evolutionary) pressure of the constraints. The epoch concludes when a sufficient number of feasible individuals has been generated – this is one of the run-time parameters of the heuristic. Typically this threshold value is set to around 40% of the population – clearly, this value controls the balance between feasibility and diversity. The next epoch then sees most of the control of the selective pressure relinquished to the objective function itself; at this stage, the algorithm works like a standard GA, which penalises constraint violations in the conventional manner. Upon registering a drop in the feasibility percentage of the population below a certain threshold value, the algorithm reverts once more to the constraint-led mode of operation for the next epoch and so it proceeds until some termination condition is met (typically, the objective value of the fittest individual reaches a pre-set threshold).

As noted before, this novel constraint-handling mechanism is the result of our efforts to solve a problem subject to a very high number of constraints. We shall review the constraints of our reducer design problem shortly, but first let us consider the 'genotype', or the set of design variables, that defines the two-stage reducer.

3 The 'Genotype' of the Two-Stage Reducer

The class of reducers we are considering here (see Figure 1 for an example) is highly standardised, both in terms of the design of their gearings and bearings and in terms of their layout. The set of 18 design variables that define the reducer unequivocally (see Table 1) reflects this, with standardisation imposing discrete value sets on most of them.

Nevertheless, the resulting design space is still vast. Assuming a discretisation of the few continuous variables into 25 steps, we could obtain a number of possible reducer designs of the order 4×10^{26} (it is notoriously hard to gain an intuitive 'feel' for such numbers, but it is worth considering that this is comparable to the number of atoms in about eight kilograms of C-12!).

Two conclusions can be drawn from here. Firstly, it is clear that, although the computational cost of evaluating the objective function and the constraints for a given design is quite low (less than a second per design), an exhaustive (full factorial) search of the design space is not feasible. Secondly, of all the search heuristics one could consider using instead, population-based evolutionary algorithms appear to be the most suitable, reinforcing the observation we made earlier regarding the popularity of such approaches in the transmission design literature.

4 A Highly Constrained Design Space

This is probably a good time to turn our attention to the constraints of our design problem, which, as hinted earlier, constitute the key challenge of this class of problems in general and of optimizing the structural elements of the reducer shown in Figure 1 in particular. As it will become apparent, they are all of the inequality type, mostly involving geometrical or structural considerations. There are a total of 77 constraints, which we have organized into 24 groups (e.g., the same type of constraint applied to all four gears constitutes one group, though it is actually implemented as four separate constraints).

In the interest of conciseness we shall not dwell on the details of their calculation (the interested reader may find all the details of the gearing calculations in the relevant industrial standard document[2] and the catalogs and calculation methods used for bearings and seals in the SKF online catalog[1]); we do note though that they are mostly based on a combination of lookup tables and simple analytical models, their computational cost therefore being minimal.

Additionally, it is worth noting that to enable the calculation of some of the constraints (such as some of the geometrical inequalities, as well as those related to lubrication and operating temperature), the housing was also designed automatically once the virtual design of each two-stage gearing was generated.

The following list of constraints should be viewed with reference to the sketch in Figure 1.

Table 1 The 18 design variables defining the reducer.

Symbol	Range	Unit	Description
m_{G12}	$\in \{1.12, \dots 40\}$	–	Gear ratio of the first stage. Standardised, discrete real values.
a_{w1}, a_{w2}	$\in \{71, \dots, 400\}$	mm	Center distances of first and second stage respectively. Standardised, discrete, real values.
x_{n1}, x_{n3}	$\in \{-0.5, \dots, 1\}$	–	Normal tooth addendum coefficients of first and second stage pinions respectively. Standardised, discrete, real values.
Ψ_{a1}, Ψ_{a2}	$\in [0.2, \dots, 0.8]$	–	Gear width to center distance ratio of first and second stage respectively. Real values.
β_1, β_2	$\in [7.25, 15]$	degrees	Helix angle measured at the pitch diameters of the first and second stage respectively. Discrete real values.
z_1, z_3	$\in \{17, \dots, 15\}$	–	Number of teeth of first and second stage pinions respectively. Integer values.
i_1	$\in \{0, \dots, 63\}$	–	Catalogue index of standardised end for the input shaft. Integer values.
i_2	$\in \{0, \dots, 127\}$	–	Catalogue index of rotary seal for the input shaft. Integer values.
i_3	$\in \{0, \dots, 63\}$	–	Catalogue index of tapered roller bearing for the input shaft. Integer values.
i_4	$\in \{0, \dots, 63\}$	–	Catalogue index of tapered roller bearing for the intermediary shaft. Integer values.
i_5	$\in \{0, \dots, 63\}$	–	Catalogue index of tapered roller bearing for the output shaft. Integer values.
i_6	$\in \{0, \dots, 127\}$	–	Catalogue index of rotary seal for the output shaft. Integer values.
i_7	$\in \{0, \dots, 63\}$	–	Catalogue index of standardised end for the output shaft. Integer values.

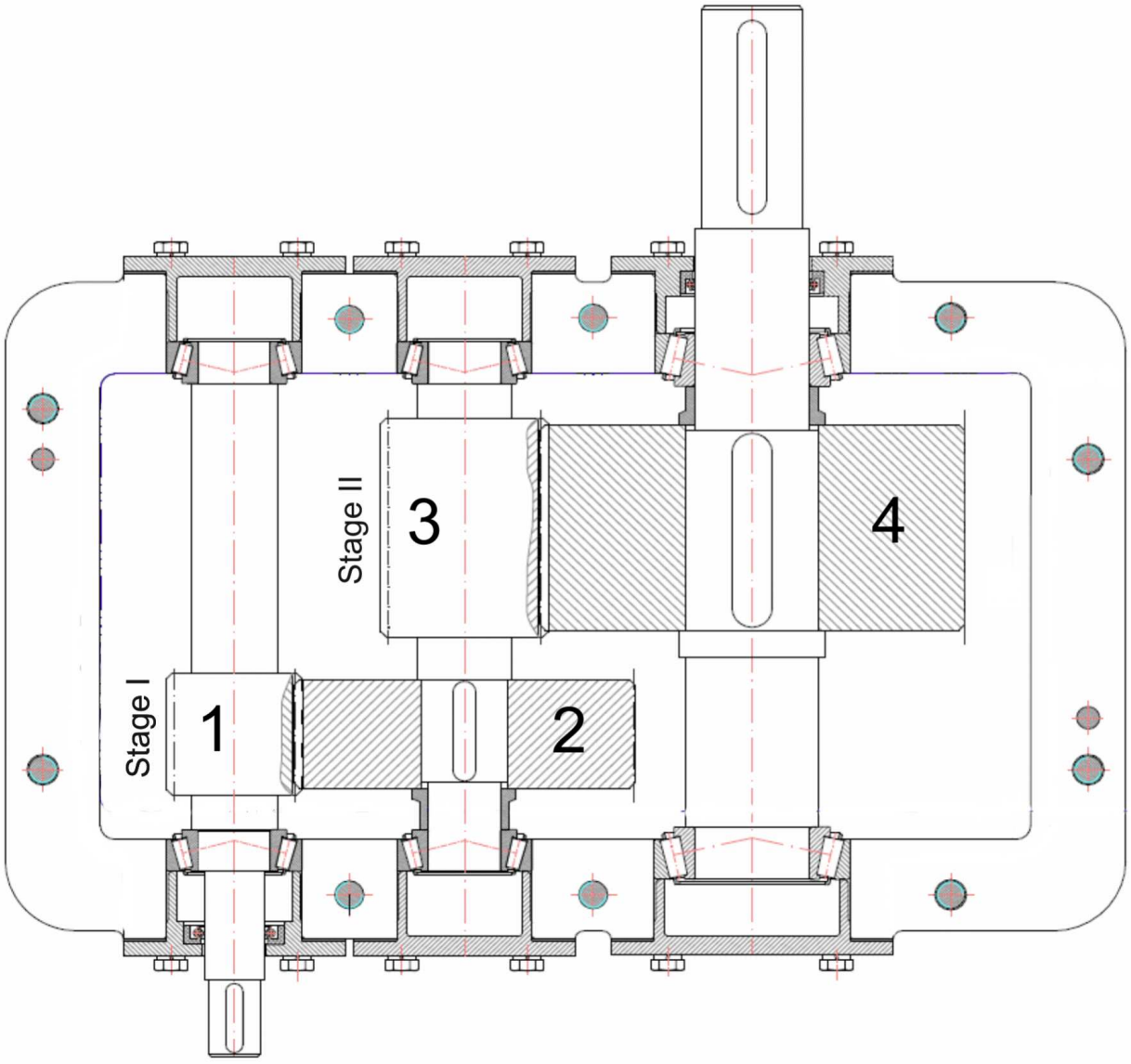

Fig. 1 Sketch of the two-stage reducer.

Constraint group 1 *The relative difference between the required and the actual gearing ratio must be within the range* $[-2.5\%, 2.5\%]$ *on both stages.*

Constraint group 2 *The Hertzian contact pressure on the teeth of both stages must not exceed a specified value.*

Constraint group 3 *The bending stress on the teeth of gears 1 through 4 must not exceed a specified value.*

Constraint group 4 *The teeth on gears 1 through 4 must not be undercut.*

Constraint group 5 *The top land of the teeth on gears 1 through 4 must not vanish.*

Constraint group 6 *The contact ratio on stages I and II must be greater than a specified value.*

Constraint group 7 *The normal addendum coefficients on both stages should be in the range* $[-0.5, 1]$.

Constraint group 8 *Measurability constraint for all four gears.*

Constraint group 9 *The numbers of teeth on the gears of both stages must be relative primes.*

Constraint group 10 *Gear 2 must not interfere with the output shaft.*

Constraint group 11 *Lubrication constraint – the margin between the minimum and maximum allowable lubricant levels should be no less than 10mm.*

Constraint group 12 *The input and output shaft ends must have a sufficient diameter step to allow the mounting of a belt wheel.*

Constraint group 13 *The inside diameter of the tapered roller bearings on the input and output shafts must be less than the mounting diameter of the the seal.*

Constraint group 14 *Geometrical constraint relating to the space required by the outside ring of the tapered roller bearings on the input and output shafts.*

Constraint group 15 *Set of manufacturability constraints on all four gears.*

Constraint group 16 *Input, output and intermediary shaft stress constraints (radial and axial loads originating from the gearings).*

Constraint group 17 *The fatigue life safety factors on the three shafts must not fall below a specified value.*

Constraint group 18 *The bending strains on the three shafts in key locations must be below certain threshold values to enable the correct functioning of the gearings and the bearings.*

Constraint group 19 *The torsional strains in the three shafts must be below a threshold value.*

Constraint group 20 *The service life of the tapered roller bearings must exceed a specified value.*

Constraint group 21 *The shearing and crushing stresses must not exceed a specified value on the keys and keyways of the input and output shafts.*

Constraint group 22 *The minimum distance between adjacent roller bearings must be greater than 15mm.*

Constraint group 23 *The shearing and crushing stresses must not exceed a specified value on the keys and keyways used the attach the four gears to the shafts.*

Constraint group 24 *The operating temperature of the reducer must not exceed a specified value.*

5 A Mass Minimization Problem

Let us now consider the following design problem. A 2.9kW two-stage reducer is to be designed for minimum weight and a service life of 8,000 hours, given an input speed of 925 RPM and a transmission ratio of 7.6. The gears should be based on an ISO 53 basic rack profile, with the pinions and wheels made of quenched and tempered 42CrMo4 and 41Cr4 respectively.

Running the algorithm described earlier yielded a reducer with a 2.8×2.7 division of the transmission ratio and axial distances of 80mm and 100mm on stages one and two respectively, weighing 44.3 kg. This solution was found on the boundary of the second transmission stage Hertzian contact pressure constraint, very near to four additional constraint boundaries. These are highlighted in Figure 2.

In contrast to the successful determination of this global optimum found in an 'awkward' corner defined by several constraint boundaries, consider the outcomes of a set of benchmark experiments. The natural basis for comparison is, of course, the standard GA the multi-epoch heuristic is built upon, run with the epoch switching feature disabled. Multiple repeated runs of this GA failed to reach the objective value of 44.3kg within a budget of 300,000 evaluations. In fact, this standard GA failed to reach even a slightly relaxed threshold objective value of 45.8kg by an arbitrarily selected cut-off point of 300,000 evaluations. By comparison, the multi-epoch algorithm has attained this

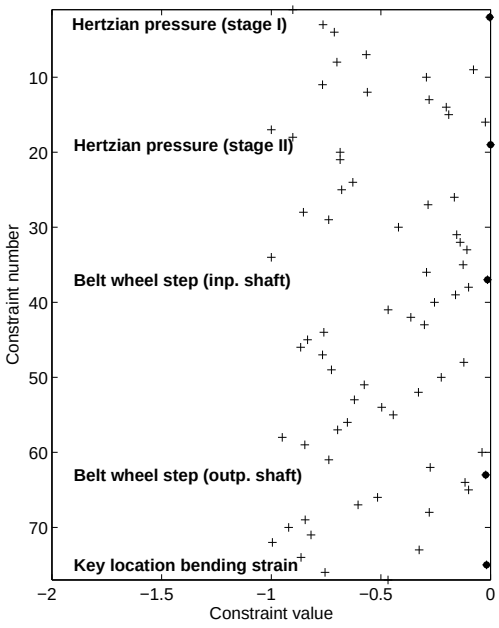

Fig. 2 The '+' symbols indicate the values of the constraints of the problem at the optimum design, with black dots highlighting the five constraints whose boundaries are closest to that optimum (note that the value g_i of constraint number i is defined as $g_i = a_i/b_i - 1$, where the constraint is of the form $a_i < b_i$) .

value on every one of 50 independent runs (started from different random initial populations) of up to 300,000 evaluations – Figure 3 is a histogram of the number of evaluations required by it to do so (average of just under 75,000 evaluations, standard deviation of just over 54,000). Perhaps even more tellingly, as the first bar of the histogram shows, 17 of the 50 runs passed the threshold weight value after less than 50,000 evaluations.

As an additional benchmark, we have also tested a Simulated Annealing optimizer (an implementation of a recent version of the heuristic by Talbi [14]) as a means of tackling the reducer problem. Experiments with five different types of cooling schedules (linear, exponential, parabolic, hyperbolic and power) run to 1,000,000 evaluations of the objective function, failed, just as in the case of the "plain" Genetic Algorithm, to reach the threshold weight, once again underscoring the extreme difficulty of this mixed variable problem.

In addition to the solution of the basic design problem, the type of optimization capability demonstrated here opens the possibility of evaluating objective function sensitivities with respect to the elements of the design brief in a timely manner. Consider, for example, the impact of the required service life of the reducer on the mass (our objective function in this study). This is

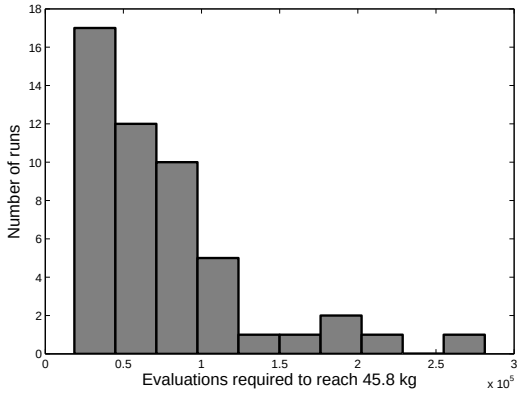

Fig. 3 Histogram of evaluation numbers required by the two epoch algorithm to reach the weight threshold of 45.8kg.

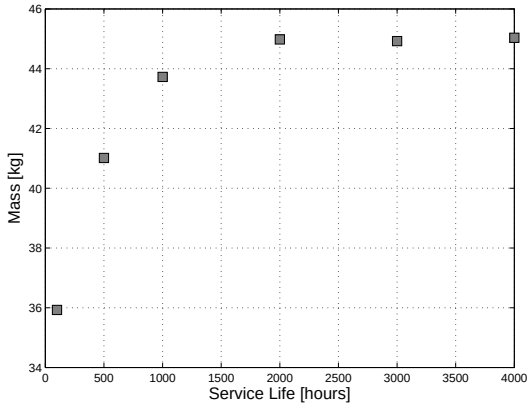

Fig. 4 Minimum weight reducers designed for various service life spans – note that the required service life has to be reduced to less than 2,000 hours before any weight reductions will result.

a high level relationship shrouded by a plethora of low level connections between the design variables and the constraints, whose analytical calculation can be considered, for all practical purposes, intractable. We can, however, obtain discrete handholds on this relationship by simply running the optimizer for different values of the service life – the results of such a study are shown in Figure 4.

Here are some typical conclusions that can be drawn from this type of study, as shown on the plot. If we were aiming for, say, 2,000 hours, we would need to sacrifice 75% of this service life for a 4kg (roughly 10%) weight saving. Halving the required service life would only save us just over one kilogram. At the same time, if weight was our sole concern, there would be no point in making any service life sacrifices if we cannot allow it to drop below 2,000 hours, as no weight saving would result.

6 Conclusions and the Way Forward

Compared to many other industrial products, certain classes of mechanical transmissions (such as the reducers discussed here) offer the allure of a finite design space (as a result of the standardisation of their layout and most of their design variables), which should make global optimization considerably easier.

Alas, as a result of the 'curse of dimensionality', even these design spaces can be vast and can present a considerable challenge to the designer, especially when a multitude of inequalities constrain the design space.

In this paper we have shown how an evolutionary algorithm, augmented by a constraint handling heuristic based on a biological paradigm, can make solving such complex structural design problems a feasible proposition, even when detail design level constraints are taken into account.

We have considered a particular class of transmissions here, but in building the tool for solving the related design problem, we have not encountered any severe scalability issues (though clearly, a greater number of design variables would further increase the size of the design space and therefore longer GA run times may be required) – therefore, broader classes of transmissions could be considered for the same treatment, leading ultimately to a generic transmission system design tool based on the evolutionary optimization concepts described here.

Taking a broader perspective, this type of heuristic might facilitate the solution of other heavily constrained design problems, which, as a result of the complexity of their constraint boundaries and relatively small size of their feasible regions, present unsurmountable obstacles to conventional problems. Complex engineering systems with multiple interactions between their subsystems are often the sources of such optimization problems. Consider, for instance, the conceptual design of airframes: an intensely multi-disciplinary problem typically yielding a variety of constraints related to aerodynamic performance, structural design criteria, environmental impact, cabin design, payload positioning, etc. For similar reasons, the automotive industry encounters many similar problems too, adding crashworthiness to the above list as another typical source of highly restrictive constraints.

There is room for further development in terms of the fidelity of the analysis, which drives the optimization process described here – of course, given the relatively large number of objective function evaluations demanded by the sheer size of the design space, this would have to be achieved through a careful control of the computational cost of the analysis. In the same vein, further objective functions could also be considered – manufacturing cost is a potential example.

In parallel with such developments, there is also considerable scope for the better understanding of a series of algorithm design questions related to the heuristic introduced here. For example, when is the best time to conclude an equilibrium epoch? A cursory study led to us using 40% as the ratio of feasible individuals being reached as an epoch termination criterion, but it is hard to tell at present whether this is a problem-dependent number. Indeed, it is uncertain whether the feasibility percentage is a good indicator of the optimum switching moment. Similarly, further (broader) studies may reveal more effective ways of deciding on the best moment to switch back for a another constraint satisfaction phase.

7 Acknowledgements

The work of A. Sóbester has been supported by the Royal Academy of Engineering and the Engineering and Physical Sciences Research Council. Romanian Government grant PN II ID PCE 2007-2010, CNCSIS Code ID 1077 supported the work of the other authors.

References

1. http://tinyurl.com/p7u4yx. Accessed on 17/08/09.
2. DIN 3990 Teil 3 Tragfähigkeitsberechnung von Stirnrädern. Deutsches Institut Für Normung E.V. (1987)
3. Abersek, B., Flasker, J., Balic, J.: Expert system for designing and manufacturing of a gear box. Expert Systems with Applications **11**(3), 397–405 (1996)
4. Cohoon, J.P., Hegde, S.U., Martin, W.N., Richards, D.S.: Distributed genetic algorithms for the floorplan design problem. IEEE Transactions on Computer-Aided Design **10**(4), 483–492 (1991)
5. Eldredge, N., Gould, S.J.: Punctuated Equlibria: An alternative to phyletic gradualism, in Schopf, T. (ed.), Models in Paleontolgy. Freeman Cooper, San Francisco (1972)
6. Ferguson, G.L., Robinson, M., Moynihan, G.P.: Expert system for selecting speed reduction components for a power transmission. Journal of Manufacturing Systems **18**(1), 66–74 (1999)
7. Gologlu, C., Zeyveli, M.: A genetic approach to automate preliminary design of gear drives. Computers & Industrial Engineering **In Press**(x), x–xx (2009)
8. Hamada, M., Martz, H.F., Reese, C.S., Wilson, A.G.: Finding near-optimal Bayesian experimental designs via genetic algorithms. The American Statistician **55**(3), 175–181 (2001)
9. Li, R., Chang, T., Wang, J., Wei, X.: Multi-objective optimization design of gear reducer based on adaptive genetic algorithm. 12th International Conference on Computer Supported Cooperative Work in Design (2008)
10. Lin, Y., Shea, K., Johnson, A., Coultate, J., Pears, J.: A method and software tool for automated gearbox synthesis. Proceedings of the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, August 30 - September 2, San Diego, California, USA pp. 1–11 (2009)

11. Oh, S., Yoon, H.: An analysis of punctuated equilibria in simple genetic algorithms. Lecture Notes in Computer Science **1363**, 195–206 (1998)
12. Ray, T., Liew, K.M.: Society and civilization: An optimization algorithm based on the simulation of social behavior. IEEE Transactions on Evolutionary Computation **7**(4), 386–396 (2003)
13. Ray, T., Saini, P.: Engineering design optimization using a swarm with an intelligent information sharing among individuals. Engineering Optimization **33**(6), 735–748 (2001)
14. Talbi, E.G.: Metaheuristics: From Design to Implementation. Wiley (2009)
15. Thompson, D.F., Gupta, S., Shukla, A.: Tradeoff analysis in minimum volume design of multi-stage spur gear reduction units. Mechanism and Machine Theory **35**, 609–627 (2000)