Multiobjective Monotonicity Analysis: Pareto Set Dependency and Tradeoffs Causality in Configuration Design

Sigurdarson, Nökkvi S.; Eifler, Tobias; Ebro, Martin; Papalambros, Panos Y.

Published in:
Journal of Mechanical Design

Link to article, DOI:
10.1115/1.4052444

Publication date:
2022

Document Version
Peer reviewed version

Journal of Mechanical Design. Received December 17, 2020;
Accepted manuscript posted August 31, 2021. doi:10.1115/1.4052444

Multiobjective Monotonicity Analysis: Pareto Set Dependency and Tradeoffs Causality in Configuration Design

Nökkvi S. Sigurdarson*
Mechanical Engineering,
Technical University of Denmark,
Kgs. Lyngby, Denmark,
noksig@mek.dtu.dk

Tobias Eifler
Mechanical Engineering,
Technical University of Denmark,
Kgs. Lyngby, Denmark,
tobeif@mek.dtu.dk

Martin Ebro
Device R&D,
Novo Nordisk A/S,
Hillerød, Denmark,
mixe@novonordisk.com

Panos Y. Papalambros
Mechanical Engineering,
University of Michigan,
Ann Arbor, MI 48109,
pyp@umich.edu

Multiobjective design optimization studies typically derive Pareto sets or use a scalar substitute function to capture design trade-offs, leaving it up to the designer's intuition to use this information for design refinements and decision making. Understanding the causality of trade-offs more deeply, beyond simple post-optimality parametric studies, would be particularly valuable in configuration design problems to guide configuration redesign. This paper presents the method of Multiobjective Monotonicity Analysis to identify root causes for the existence of trade-offs and the particular shape of Pareto sets. This analysis process involves reducing optimization models through constraint activity identification to a point where dependencies specific to the Pareto set and the constraints that cause them are revealed. The insights gained can then be used to target configuration design changes. We demonstrate the proposed approach in the preliminary design of a medical device for oral drug delivery.

Nomenclature

$\mathcal{A}$ attainable set
$\mathcal{C}$ Pareto Set
$\mathbf{c}$ vector of bound objectives in the upper bound problem
D_s indices of the constraint functions that depend on a shared variable x_i
f primary objective function in the upper bound problem
$f(x^+)$ a function increasing monotonically w.r.t. x
$f(x^-)$ a function decreasing monotonically w.r.t. x
$\mathbf{F^*}$ $[k,j]$-matrix of Pareto optima
$\mathbf{E}$ $[k\text{-}1,j]$ dimensional matrix of sampled values of ε
$\mathbf{g(x)}$ vector of inequality constraints for the design problem
$\mathbf{G^*}$ matrix of$\mathbf{g}(\mathbf{x}^*)$ values stored for every run
$\mathbf{h(x)}$ vector of equality constraints for the design problem
$\mathbf{H^*}$ matrix of $\mathbf{h}(\mathbf{x}^*)$ values stored for every run
j number of computational iterations ε is sampled over
k number of objectives
n number of design variables
$\mathbf{x}$ vector of design variables
$\underline{x}$ argument of the infimum of the design problem
$\overline{x}$ argument of the supremum of the design problem
$\overline{x_i}$ monotonic trade-off variable
$\mathcal{X}$ feasible domain
$\mathcal{X}\varepsilon$ feasible domain for a given upper bound value, ε
ε k-1 dimensional vector of upper-bound parameters
ε_i upper-bound parameter for the *ith* bound objective
$\varepsilon_\mathbf{L}$ lower limit of objective bounds
$\varepsilon_\mathbf{U}$ upper limit of objective bounds
$\tilde{\varepsilon}_i$ reduced-objective variable
$\tilde{\varepsilon}^*_{i,j}$ optimal value of $\tilde{\varepsilon}_i$ implied by the activity case where the Pareto-constraint $g_j(\mathbf{x},\tilde{\varepsilon})$ bounds $\tilde{\varepsilon}_i$
λ Lagrange multiplier vector of inequality constraints

*Corresponding Author

1 Introduction

Designers naturally aim to embody solutions that trade off a range of functionality and production objectives. Over time, competitive pressures require designers to improve performance while integrating more features with each new product generation [1]. Additional trade-offs arise as a result. While optimization methods are commonly used at the embodiment stage, systematic, quantitative analysis of trade-offs is less common ahead of important decisions such as concept selection, iterative redesign, or requirement setting [2]. Whether due to time constraints, task complexity, or early-stage design uncertainties, knowledge about trade-offs is largely experience-driven in design practice [3].

Pahl & Beitz [4] note that optimizing the "carrier of several combined functions" can be difficult. Yet, they also argue that decisions on what parts and subsystems contribute to different aspects of product functionality are made early on, typically during embodiment design. This is a term used somewhat interchangeably with preliminary design, configuration or topology design [5], layout [6], system design [7], or system architecture [8]. Here, we use the term configuration design for consistency with the design optimization literature. Andreasen and Howard [9] similarly argue that identifying and managing trade-offs is a key challenge in embodiment design. Different configurations will ultimately be affected by different trade-offs, a notion well accepted also in design optimization.

Multiobjective Design Optimization (MODO) techniques study what is achievable in a design subject to trade-offs, typically identifying, comparing, simplifying, and visualizing Pareto sets. Procedures for selecting a point on a Pareto set are essentially post-optimality analyses. Occasionally, they lead to converting the multiobjective problem to a scalar one with a new objective at a higher level, such as going from engineering design to design for market systems, e.g., Shiau and Michalek [10].

There is paucity in discussing *why* the result is a set rather than a dominant optimum. Yet, existing methods seem to focus only on selecting points within the set or on measures to describe how the objectives compete. Selecting a point in a Pareto set includes work on modeling preferences [11, 12, 13], identification of compromise solutions by measuring the distance to a utopia point [14], scaling methods to account for objective weighting [15], and strategies for making trade-offs aggressively or conservatively [16]. Substantial work exists for sensitivity, robustness [17], uncertainty [18], visualisation [19], dimensional reduction [20], and identification of competing objectives in a n-dimensional objective space [11]. Furthermore, structural topology optimization (TO) [21] is a notable contribution in the context of multiobjective configuration design problems. Somewhat uniquely, TO optimizes a functional representation of the design without the actual embodiment of the functions, and the results inform configuration design.

Some post-optimality analyses aim at understanding how objectives compete. Multiple measures for Pareto frontier shape exist e.g., [11, 22, 23]. Frischknecht and Papalambros [24] developed metrics to measure the alignment of objective pairs and later suggested a Pareto set analysis using local measures of objective coupling [25] to compare system topologies. Metrics describing the *quality* of a Pareto set such as hypervolume, Pareto-spread, and generational distance [26] have also been suggested to compare Pareto sets for alternative configuration designs. To tackle the comparison of multi-dimensional objective spaces, Athan and Papalambros [27] introduced the notion of meta-Pareto sets, which consist of the union of Pareto sets of multiple alternative configurations. Mattson and Messac [18] similarly put forward an approach to concept selection using s-Pareto frontier to compare alternatives.

There are three challenges with the current MODO approaches to design. First, the main focus is on optimizing a fixed design rather than questioning why the objectives compete. Second, the analysis done at earlier time points in the product's evolution may become obsolete at a later design stage. Finally, if the Pareto set contains no points acceptable to the designer, e.g., due to non-modelled considerations, there is little guidance for what to do next. A rigorous approach to gain insights into the root cause of the trade-offs inherent to the design would substantially increase the value of optimization at an early stage of product development.

Originally developed by Papalambros and Wilde [28], Monotonicity Analysis (MA) is a rigorous, yet opportunistic, method used to identify active constraints. When applicable, it allows model reduction and assessment of model boundedness, and in some cases, it reveals global optima with little or no computation. Michelena and Agogino [29] expanded the method to multiobjective problems by applying MA to a weighted sum formulation. Gobbi et al. [30] and Mastinu et al. [31] later applied MA in a procedure to derive Pareto sets analytically. Unlike the other approaches discussed above, MA can be performed prior to numerical computation and even before a full optimization model has been built. To date, MA has largely been used as a model "debugging" tool.

Monotonicity analysis is of interest here for its implications in a design context. Jain and Agogino [32] demonstrated how MA could be used to support the conceptualization of a multi-speed gearbox and explore configuration changes that lead to superior designs compared to proportional changes alone. Ishii and Barkan [33] applied MA in an interactive expert system, intending to help designers identify bottlenecks in the design caused by active constraints. Cagan and Agogino [34] also used MA to reveal previously hidden relationships through back-substitution of active constraints into objective functions. They identified ways to expand the design space to widen the search for design improvements. Deb and Srinivasan [35] meanwhile, discussed the similarities between MA and their 'innovation through optimization,' or *innovization* procedure aimed at deriving design principles using commonalities among Pareto-optimal designs. They argued that both MA and their NSGA-II-based innovization approach help identify important relationships at the optimum.

Most of this prior work [32, 34, 35] has focused on understanding the common characteristics of Pareto-optimal designs to allow reuse in future designs but within a single

configuration. Yet, if a configuration has limitations or is just not very good, one would simply find the best compromise for a poor design. If MA can identify relationships for the design variables at optimality, then arguably, it might also be able to identify relationships that *limit* optimality. In a multiobjective formulation, such analysis could lead to the discovery of the root cause for trade-offs between objectives.

In the remainder, Section 2 articulates the aims of this work, Section 3 provides some theoretical foundation for the Pareto set Dependency Analysis method developed in Section 4. Section 5 presents the methodology applied to the SOMA (Self-Orienting Millimeter-scale Applicator) drug delivery device currently in development. We offer a discussion in Section 6 and conclude in Section 7.

2 Aim of this work

Multiobjective optimization quantifies trade-offs among competing objectives. While trade-offs can be studied computationally, understanding the underlying causes is typically left to designers' ability to interpret results and to identify redesigns aimed at improving performance. Optimization has been claimed to have an intrinsic value in the design process beyond just providing numerical optimal solutions [2, 34, 36]. How to extract such design knowledge systematically is left to the designer, particularly in the early design stages. This then begs the question:

How can conceptual or configuration design limitations reflected in the Pareto set be identified rigorously? In particular, what specific design dependencies and constraints cause trade-offs?

This work seeks to demonstrate how the limitations of a design configuration may be identified through rigorous analysis rather than through tacit knowledge and heuristics alone, using novel extensions to monotonicity analysis.

To this end, we apply MA to multiobjective problems posed in the upper-bound formulation, also known as the bound objective [14] or ε-constraint method [37]. While MA is often used to check the *validity* of a model, here we demonstrate extensions to MA that allow it to be used to check the design itself when it exhibits global or regional monotonic behaviour. We use constraint activity identification and systematic reduction of the model's degrees of freedom to reveal often hidden dependencies among variables and objectives at the optimum, which cause the trade-offs. Designers will still need experience and intuition to convert this knowledge into actionable redesign decisions, but these decisions are informed by deeper understanding from rigorous analysis.

3 Theoretical Foundation

The multiobjective design optimization problem is stated in negative-null form as:

$$\text{min.} \quad \mathbf{F}(\mathbf{x}) \tag{1}$$
$$\text{subject to} \quad \mathbf{g}(\mathbf{x}) \leq 0 \tag{2}$$
$$\mathbf{h}(\mathbf{x}) = 0 \tag{3}$$
$$\mathbf{x} \in \mathbb{P} \tag{4}$$

where **F(x)** is a vector of design objectives f_i, $i = [1,2,...,k]^T$, **x** is a vector of design variables, and **h(x)** and **g(x)** are the equality and inequality constraints respectively. If $\mathcal{X}$ denotes the feasible domain, then the attainable set $\mathcal{A}$ contains all values of $\mathbf{F}(\mathbf{x})|x \in \mathcal{X}$. A point $\mathbf{F_0}(\mathbf{x}^*)$ in the attainable set $\mathcal{A}$ is said to be Pareto-optimal if and only if there exists no point in the attainable set that fulfills:

$$\mathbf{F}(\mathbf{x}) \leq \mathbf{F_0}(\mathbf{x}^*) \qquad \wedge \qquad f_i(\mathbf{x}) < f_i(\mathbf{x}^*) \tag{5}$$

The set of all Pareto-optimal points is the Pareto set $\mathcal{C}$ sitting on the boundary of the attainable set $\mathcal{A}$ [38]. There are many ways to construct the Pareto set; in the trade-off analysis that follows, we use the *upper-bound formulation* also known as the ε-*constraint method* (see [14] for a methods overview).

Monotonicity Analysis [28] leverages any existing monotonic behavior of objective and constraint functions to check for boundedness and identify constraint activities thus reducing the problem's degrees of freedom. A function is said to be monotonically increasing with respect to a variable x if $f(x_2) > f(x_1)$ for any $x_2 > x_1$. This monotonic relationship between f and x is denoted $f(x^+)$. Correspondingly, a function is monotonically decreasing with respect to x if $f(x_2) < f(x_1)$ for any $x_2 > x_1$, and denoted $f(x^-)$. In the presence of monotonicity, the following principles [36] can be exploited in single-objective problems to identify activity of certain constraints, without computing the optimum first:

First monotonicity principle (MP1)
In a well-constrained minimization problem, every increasing variable is bounded below by at least one non-increasing active constraint.

Second monotonicity principle (MP2)
In a well-constrained minimization problem, every nonobjective variable is bounded both below by at least one non-increasing semi-active constraint and above by at least one non-decreasing semi-active constraint.

Constraint activity means that the location of the optimum is altered if the constraint is deleted. Active inequality constraints will be satisfied as strict equalities at the optimum, thus reducing the degrees of freedom accordingly. By identifying active constraints, one can solve the constraint functions with respect to (w.r.t.) one of their dependent variables and substitute the solution for that variable into the remaining constraint functions and objectives, thereby eliminating the active constraints and the substituted variables.

3.1 Modelling and Computation

As mentioned, multiobjective MA was originally demonstrated using a weighted-sum formulation [29]. For the trade-off analysis method development that follows in section 4, we use the upper-bound formulation [14]. This formulation involves converting the problem in Eq. 1-4 into

a single objective one:

$$\begin{aligned} \text{min.} \quad & f(\mathbf{x}) && (6)\\ \text{s.j.t} \quad & \mathbf{c}(\mathbf{x};\boldsymbol{\varepsilon}) \leq 0 && (7)\\ & \mathbf{g}(\mathbf{x}) \leq 0 && (8)\\ & \mathbf{h}(\mathbf{x}) = 0 && (9)\\ & \mathbf{x},\boldsymbol{\varepsilon} \in \mathbb{P} && (10) \end{aligned}$$

In this formulation, originally put forward by Carmichael [37], $\mathbf{c}(\mathbf{x};\boldsymbol{\varepsilon})$ is a $k-1$ dimensional vector of *bound objectives* expressed in the form $c_i(\mathbf{x},\varepsilon_i) = f_{i+1}(\mathbf{x}) - \varepsilon_i \leq 0$ or $c_i(\mathbf{x},\varepsilon_i) = \varepsilon_i - f_{i+1}(\mathbf{x}) \leq 0$, $i = [1,2,..,(k-1)]$. The vector $\boldsymbol{\varepsilon}$ of parameters ε_i represents the upper bounds of the bound objectives. When $f(\mathbf{x})$ is minimised for given values of ε_i, then the solution $\mathbf{x}^*$ is Pareto optimal if all of the bound objectives are active with non-zero Lagrange multipliers. Pareto points are thus identified by varying $\boldsymbol{\varepsilon}$ systematically between lower $\boldsymbol{\varepsilon}_\mathbf{L}$ and upper limits $\boldsymbol{\varepsilon}_\mathbf{U}$. See [14, 37, 39] for an overview of works on the upper bound formulation, the underlying mathematics, and approaches to defining suitable limits for $\boldsymbol{\varepsilon}$. The Pareto set is constructed by sampling a set of $\boldsymbol{\varepsilon}$ parameter values

$$\mathbf{E} = (\boldsymbol{\varepsilon}_\mathbf{U} - \boldsymbol{\varepsilon}_\mathbf{L})\mathbf{R} + \boldsymbol{\varepsilon}_\mathbf{L} \tag{11}$$

where $\mathbf{R}$ is a matrix of uniformly distributed quasi-random numbers between 0 and 1 of the dimension [k-1; j], where j is the number of computational iterations, and k is the number of objectives. A low discrepancy quasi-random set (e.g., a Halton set) can be used to reduce bias in $\mathbf{R}$ to reduce the computational cost of achieving a Pareto set with low sparsity. After sampling, the optimization problem is solved iteratively, as in the following pseudo-code:

for i = 1..j **do**
  Set upper bound on constrained objectives, $\boldsymbol{\varepsilon} = \mathbf{E}(:,i)$
  Solve optimization problem w.r.t $\boldsymbol{\varepsilon}$
  Store optimum, $\mathbf{F}^*(:,i) = [f^*, \boldsymbol{\varepsilon}^T]^T$
  Store arguments, $\mathbf{X}^*(:,i) = \mathbf{x}^*$
  Store Lagrange multipliers, $\Lambda(:,i) = \lambda$
  Store constraint values ,$\mathbf{G}^*(:,i) = \mathbf{g}(\mathbf{x}^*)$ and $\mathbf{H}^*(:,i) = \mathbf{h}(\mathbf{x}^*)$
end for

The sparsity of the approximated Pareto set decreases as j increases, while the span increases with j and the difference between ε_U and ε_L. With an increased j, one identifies more Pareto points resulting in a more dense Pareto set. A high j can be necessary to approximate the shape of the Pareto set, should it have interactions between the objectives that exist locally in the attainable set, for instance creating knee like shapes [23]. Beyond a certain limit, the Pareto set will have been exhaustively constructed, meaning no additional feasible solutions can be found by further increasing the difference between ε_U and ε_L. Thus, one can also solve the MODO problem multiple times with a relatively low j, increasing the difference between ε_U and ε_L, until the boundaries of the Pareto-set seem to have been identified, and then subsequently increasing j to the desired level of density.

As discussed in [14], the $\boldsymbol{\varepsilon}$-constraint formulation does have certain limitations. It results in the identification of non-optimal solutions when the bound objectives are inactive, computational iterations are "wasted" on values of $\boldsymbol{\varepsilon}$ that lie between the Pareto set and the Utopia point, and it might only identify local optima in non-convex attainable sets. In situations where the problem $\mathcal{A}$ is non-convex or computationally expensive, one could use another formulation for numerical solution and only use the $\boldsymbol{\varepsilon}$-constraint formulation in pre- and post-optimality analysis. One could also rely on one of the many implementations of the $\boldsymbol{\varepsilon}$-constraint method that have addressed these limitations (e.g. AUGMECON by Mavrotas [39]). However, the aim of this contribution is to identity the properties that affect the optimum rather than to identify the optimum itself efficiently. As such, we will forego further treatment of specific implementations and computational efficiency and use the general problem form in eqs. 6-10 due to its benefits in MA:

1. *Maintaining monotonic properties*: Converting a set of objectives into a composite function, e.g., a weighted-sum, can result in loss of monotonic properties when the objectives share variables. Using an upper-bound formulation avoids this issue.
2. *Objective elimination*: Introducing objectives as constraints in an optimization model allows one to parametrically study the *activity* of the bound objective across the attainable set using monotonicity analysis. If a bound objective can be determined to be active through monotonicity analysis, the objective itself can be 'optimized out' of the model through back-substitution [36], revealing how the objectives affect each other at the Pareto frontier.
3. *Sensitivity data*: Solving a constrained optimization problem yields non-zero Lagrange multipliers for active constraints, revealing the local sensitivity of the optimum w.r.t. changes in each active constraint. In the upper-bound formulation, the Lagrange multipliers of the bound objectives describe whether and to which degree the bound objectives compete with the primary objective, which some term the *trade-off ratio* [40].

It is often suggested that the most important objective should be modelled as the function being minimised [14], while the remaining objectives should be bound. To simplify monotonicity analysis, however, the most suitable approach would be to select the objective with the largest number of design variables. Doing so allows the broadest application of *MP1* in problem reduction.

4 Pareto Set Dependency Analysis

This section develops novel theory for the systematic reduction of multiobjective problems (Subsection 4.1) and the analysis of the relationships that *bound* the Pareto set (Subsection 4.2). We then use these developments to define an overall analysis procedure that allows the identification of

the dependencies between the objectives and constraints that create the Pareto set (Subsection 4.3). These developments, collectively referred to as Pareto Set Dependency Analysis, are demonstrated on algebraic models. The methods could, in principle, also be applied to meta-models and numerical models through either computational experiments or implicit model reduction [36].

4.1 Multiobjective Monotonicity Analysis (MOMA)

The reasoning behind the desire to develop a systematic approach to multiobjective monotonicity analysis is as follows. Consider that the Pareto set $\mathcal{C}$ exists on the boundary of the attainable set $\mathcal{A}$ but is not necessarily defined by the constraints alone, as unconstrained multiobjective problems also yield Pareto sets [38]. It follows that the occurrence of Pareto sets must have two causes:

1. *Trade-off variables* In negative-null form, a variable x that influences two objectives, $f_1(x)$ and $f_2(x)$, causes a trade-off if $\arg\min f_1(x) \neq \arg\min f_2(x)$. In design, this mostly occurs when an objective pair is oppositely monotonic w.r.t. a variable, either globally or regionally.
2. *Active constraints* Active constraints reduce the degrees of freedom (DOF) in optimization problems, affect the feasible domains for the remaining DOF, and change the optimum. Eliminating and back-substituting active constraints into objective functions can introduce new variables to the expression, and can change its monotonicity w.r.t. the original variables, revealing additional trade-off variables *hidden* in constraints.

Hence, multiobjective monotonicity analysis (MOMA) may allow the systematic identification of trade-off variables and reveal relationships between the objectives at the optimum that are *hidden* by constraints. This can help designers understand the root causes of trade-offs in a configuration design. Demonstrating this requires certain extensions of MA to deal with multiobjective problems.

4.1.1 Definitions and Theorems

For upper-bound formulations, the extension of MA into multiple objectives is relatively straightforward as this merely involves handling more constraints. The principles and procedures originally developed by Papalambros and Wilde [36] mostly still apply. The exception is that the bound objectives, $\mathbf{c}(\mathbf{x};\boldsymbol{\varepsilon})$, cannot be treated as traditional inequality constraints. Firstly, as we wish to vary the upper-bound values, ε, these cannot be regarded as fixed parameters when performing monotonicity analysis. Secondly, we seek to partially minimize the bound objectives, which has implications for the use of MP1 and MP2. Hence, it is necessary to introduce some theorems of relevance to how $\mathbf{c}$ is handled:

Definition 1 Trade-off Variables
If an objective pair f and c_i have a variable x_1 in common, but differ in monotonicity w.r.t. x_1, e.g., $f(x_1^+)$ and $c_i(x_1^-)$, then x_1 is said to be a trade-off variable, denoted $\underline{\bar{x}}_1$. Correspondingly, an objective pair of like monotonicity w.r.t. a common variable, indicates that the variable is harmonious and can be used to partially minimise both simultaneously.

Theorem 1 Influence of Monotonic Trade-off Variables
In the presence of monotonic trade-off variables, no dominant minimum exists, resulting in a Pareto set. The proof for this is a corollary to MP1.

Proof. Let f_1 be monotonically increasing w.r.t. $x \in \mathbb{P}$ and f_2 monotonically decreasing, and let x be well bounded from above and below. Then by MP1, $\arg\min \mathrm{f}_1(\mathrm{x}) = \underline{x}$, and $\arg\min \mathrm{f}_2(\mathrm{x}) = \bar{x}$, meaning that the minimizers for the two objectives are defined by the *greatest lower bound* (glb)and the *lowest upper bound* (lub) respectively. Hence any feasible value of x will yield a unique Pareto point. ■

Corollary 1.1 Boundedness of trade-off variables
Following Theorem 1, multiobjective problems can only be said to be well-bounded if all trade-off variables are bounded from above and below.

For instance, if a bound objective, c_i, is critical w.r.t. a monotonic trade-off variable, $\underline{\bar{x}}_1$, then the multiobjective problem is not well bounded, as $\underline{\bar{x}}_1 \to \infty$ or $\underline{\bar{x}}_1 \to 0$ when $\varepsilon_i \to \infty$ and f is minimised. This can either be handled by introducing additional constraints, or by selecting suitable limits for the upper-bound problem $\boldsymbol{\varepsilon}_\mathbf{L}$, $\boldsymbol{\varepsilon}_\mathbf{U}$.

In upper-bound formulations, we treat objectives as additional constraints and iteratively identify Pareto points, exploring $\underline{\bar{\mathbf{x}}} \in \mathcal{X}$, for different values of ε, as illustrated in Figure 1. If a bound objective is active, the model is essentially exploring a smaller region of the feasible domain $\mathcal{X}_\varepsilon \in \mathcal{X}$. From this, an additional theorem arises:

Theorem 2 Activity of Bound objectives
A bound objective $c_i(\underline{\bar{x}};\varepsilon_i)$ can either be active, semi-active, dominated, or inconsistent with other constraints, depending on the value of ε_i. The change in activity of $c_i(\underline{\bar{x}};\varepsilon_i)$ across $\mathcal{A}$ affects the shape of the Pareto set.

Consider an objective pair, $f_1(x_i^+)$ and $c_1(x_i^-,\varepsilon_1)$, with the design variable x being bounded from below by $g_1(x_i^-)$ and from above by $g_2(x_i^+)$, where ε is the upper bound parameter. Here, the value of ε determines constraint activity:

1. For the values of ε_1 where $g_1(x_i) < c_1(x_i)$, c_1 is active, and the result will be Pareto-optimal.
2. For the values of ε_1 where $c_1(x_i) < g_1(x_i)$, c_1 is inactive, and the result will not be Pareto-optimal
3. For the values of ε_1 where $g_2(x_i) < c_1(x_i)$, $\mathcal{X}_\varepsilon \in \emptyset$, and thus these constraints are inconsistent. In this case, g_2 shapes a boundary of the Pareto set.
4. For the value of ε_1 where $c_1(x_i) = g_1(x_i)$, the bound objective is semi-active, yielding the single-objective optimum for f_1. Correspondingly, $c_1(x_i) = g_2(x_i)$ yields the single-objective optimum for f_2.

Thus, exploring these changes in the activity of c_1 yields the Pareto set for the objective pair. We can hence utilise MOMA to identify the conditions under which a bound objective is active, dominated, or inconsistent. This can reveal

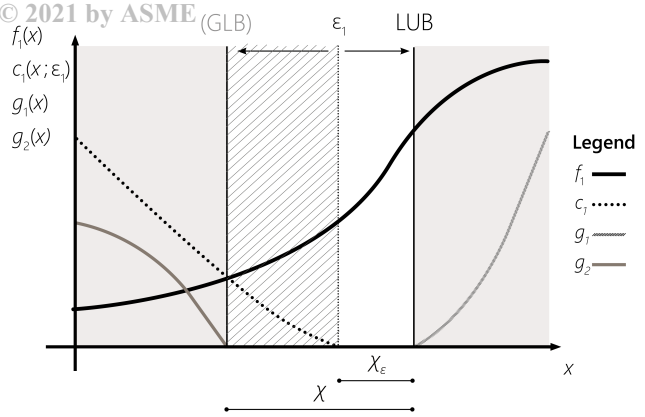

Fig. 1: MOMA allows the partial identification of the Pareto set, by identifying the values of ε where the bound objectives are active, semi-active, violated, and inconsistent

important relationships between the objectives and the constraints g_i that affect the Pareto set. Here, it is important to consider the the influence of ε on the activity of $\mathbf{g}(\mathbf{x})$:

Definition 2 Global Activity
In the monotonicity analysis of an upper-bound problem, a constraint $g_i(\mathbf{x})$ is said to be globally active if and only if $f(\mathcal{X}_i) < f(\mathcal{X}_*)$ for any $\{\varepsilon \in \mathbb{P} \,|\, \varepsilon_L \leq \varepsilon \leq \varepsilon_U\}$.

Trade-off variables can only be optimized out if an active bound objective is used to eliminate it or if the bound objective can be determined to be dominated w.r.t. said trade-off variable by another globally active constraint. This notion of global activity is central to multiobjective monotonicity analysis. A reduced model would potentially only identify parts of the Pareto set if we were to optimize variables out with constraints that are not globally active.

The final extension to MA that is necessary in order to deal with multiobjective problems is the question of how to partially minimise several objectives concurrently:

Definition 3 Partial minimisation of bound objectives
In a well-constrained multiobjective, upper-bound minimization problem, any increasing objective variable not in the primary objective, is bounded below by at least one non-increasing active constraint.

Modelling objectives as constraints is merely a route to identifying Pareto points. It is still desirable to identify partial minima for bound objectives. By simply extending MP1 into multiobjective problems, we can reduce multiple objectives, i.e., identify partial minima for f_{i+1} in $c_i(\mathbf{x},\varepsilon_i) = f_{i+1}(\mathbf{x}) - \varepsilon_i \leq 0$. Nevertheless, it is necessary to take particular care in this process. Unless it is certain that the optimal value of a given variable is the same for all objectives, i.e., $\arg\min f_i(x) = \arg\min f_j(x)$ for any i and j, optimizing the variable out would result in a model that does not describe the entire Pareto set. When a globally active constraint can be identified, the bound objectives can always be partially minimized. This is relatively straightforward to do in situations where the condition $\arg\min f_i(x) = \arg\min f_j(x)$ for any i and j is upheld by definition. Following MP1, harmonious variables and critically constrained variables [36] will always meet this condition. As will variables that are bound by constraints that only depend on harmonious variables or on variables that only influence one objective, because constraint activity will be unaffected by the values of ε.

4.1.2 Impact of constraint activity in multiobjective problems

With these definitions, we can apply MA to multiobjective problems and, in doing so, identify trade-off variables that may be *hidden* in constraints. Here, it is beneficial to note the impact on the objective functions. There are two situations of relevance to trade-off analysis; when an objective changes monotonicity w.r.t a variable, or when it becomes dependant on new variables. Consider an example:

$$\begin{aligned}
\text{min.} \quad & f_1(x_1,x_2,x_3) = x_1^2 - x_2 + x_3 && (12)\\
& f_2(x_2,x_4,x_5) = \frac{1}{x_2} - x_4^2 + 2x_5 && (13)\\
\text{s.j.t} \quad & 2x_4 - x_1 \leq 0 && (14)\\
& x_2^2 + 4x_2 - 2x_3 \leq 0 && (15)\\
& x_2^3 + 2x_4 \leq P_1 && (16)\\
& 10 - 3x_5 \leq x_5^2 && (17)\\
& x \in \mathbb{P} && (18)
\end{aligned}$$

Without inspection of the influence of the constraints, it would seem there is no trade-off between f_1 and f_2, as they are both monotonically decreasing w.r.t the only shared variable, x_2. Yet, when converted into an upper-bound formulation, monotonicity analysis reveals hidden dependencies:

$$\begin{aligned}
\text{min.} \quad & f_1(x_1^+,x_2^-,x_3^+) = x_1^2 - x_2 + x_3 && (19)\\
\text{s.j.t} \quad & c_1(x_2^-,x_4^-,x_5^+;\varepsilon_1^-) = \frac{1}{x_2} - x_4^2 + 2x_5 - \varepsilon_1 \leq 0 && (20)\\
& g_1(x_1^-,x_4^+) = 2x_4 - x_1 \leq 0 && (21)\\
& g_2(x_2^+,x_3^-) = x_2^2 + 4x_2 - 2x_3 \leq 0 && (22)\\
& g_3(x_2^+,x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 && (23)\\
& g_4(x_5^-) = 10 - x_5^2 - 3x_5 \leq 0 && (24)
\end{aligned}$$

where f_2 has been converted into a bound objective $c_1(\mathbf{x},\varepsilon_1)$. Following MP1, it is clear that g_1 and g_2 are critical w.r.t. x_1 and x_3, respectively, for any value of ε_1, and are therefore active. Following Definition 3, we also conclude that g_4 is active as it is critical for x_5, meaning we partially minimize f_2 in c_1 by optimizing x_5 out. Solving for the minimizers yields $x_1^* = 2x_4$, $x_3^* = \frac{1}{2}x_2^2 + 2x_2$, and $x_5^* = 2$. With back-substitution, a reduced problem is reached:

$$\begin{aligned}
\text{min.} \quad & f_1(x_2^+,x_4^+) = 4x_4^2 + \frac{1}{2}x_2^2 + x_2 && (25)\\
\text{s.j.t} \quad & c_1(x_2^-,x_4^-;\varepsilon_1^-) = \frac{1}{x_2} - x_4^2 + 4 - \varepsilon_1 \leq 0 && (26)
\end{aligned}$$

$$g_3(x_2^+, x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 \quad (27)$$

Here, f_1 has changed monotonicity w.r.t. x_2 and now depends on x_4, being oppositely monotonic to the bound objective c_1. Following Theorem 1, both x_2 and x_4 are trade-off variables, meaning that there is no single solution to the optimization problem but rather a Pareto set. Considering Corollary 1.1 the problem is, in fact, asymptotically bounded, as x_2 and x_4 are unbounded from below unless a well defined upper limit is imposed on ε_1. Hence, c_1 is globally active.

While this example may seem simplistic, it demonstrates the shifts in dependency between objectives that occur in the presence of active constraints. Such relationships are not necessarily easy to spot in non-reduced optimization models, nor is it given that the designer is aware of them. As such, monotonicity analysis can be used to identify trade-off variables, and in doing so, reveal what constraints in a design cause a lack of objective alignment - in this case, g_1 and g_2, as they introduce trade-off variables into the problem.

4.2 ε-Monotonicity Analysis

With the theoretical developments introduced so far, one can apply monotonicity analysis to systematically reduce multiobjective models, gradually converging towards an explicit description of the Pareto set while identifying trade-off variables in the process. When all globally active constraints have been identified, one can optimize the active bound objectives out of the model. If one determines that $c_j(\mathbf{x};\varepsilon_j) \equiv 0$, and subsequently optimizes a trade-off variable $\underline{x_i}$ out, then $f(\mathbf{x})$ and $g(\mathbf{x}), c_i(\mathbf{x};\varepsilon) \in D_s(x_i), i \neq j$ become dependent on ε_j through back-substitution. A parameter from an eliminated bound objective will be denoted as $\tilde{\varepsilon}_j$ and treated as a variable, referred to as the *reduced-objective variable*.

The reasoning behind treating $\tilde{\varepsilon_j}$ as a variable is twofold. Firstly, the primary objective function has been transformed into a bi-objective function, $f(\mathbf{x},\tilde{\varepsilon}_j)$, describing the trade-off between the primary objective, $f(\mathbf{x})$ and $\tilde{\varepsilon}_j$. Secondly, the feasible values of $\tilde{\varepsilon}_j$ are now determined by a set of constraints $\mathbf{g}(\mathbf{x},\tilde{\varepsilon}_j)$. The bi-objective Pareto front between f_1 and f_{j+1} will thus be defined by $f(\mathbf{x},\tilde{\varepsilon}_j)$ and $\mathbf{g}(\mathbf{x},\tilde{\varepsilon}_j)$. Meanwhile, the trade-offs amongst the eliminated objectives themselves are expressed through $\mathbf{g}(\mathbf{x},\tilde{\varepsilon})$, henceforth referred to as *Pareto-constraints*. This means that if we treat $\tilde{\varepsilon}_j$ as a variable, identifying the constraints that bound it can be used to better understand the cause of the shape of the Pareto set.

In principle, all active bound objectives can be eliminated from the model. This will result in a multiobjective expression $f(\mathbf{x},\tilde{\varepsilon})$ describing the trade-off between the primary objective and all others, while all the Pareto-constraints $\mathbf{g}(\mathbf{x},\tilde{\varepsilon})$ describe the trade-offs between the eliminated objectives. However, it may not always be beneficial to do so, for instance, when elimination results in a loss of monotonic properties or when explicit elimination becomes too time-consuming. To allow the furthest reduction of the model, it is beneficial to attempt to eliminate the trade-off variables that are shared between the largest number of constraints.

What remains after objective reduction is:

$$\begin{aligned} \text{min.} \quad & f_1(\mathbf{x},\tilde{\varepsilon_1},...,\tilde{\varepsilon}_{k-1}) && (28)\\ \text{s.j.t} \quad & \mathbf{g}(\mathbf{x},\tilde{\varepsilon}) \leq 0 && (29)\\ & \mathbf{h}(\mathbf{x}) = 0 && (30) \end{aligned}$$

where $f_1(\mathbf{x},\tilde{\varepsilon_i}^+)$ or when $\tilde{\varepsilon_i}$ is a maximisation objective, and $f_1(\mathbf{x},\tilde{\varepsilon_i}^-)$ when $\tilde{\varepsilon_i}$ is a minimisation objective. Applying monotonicity analysis to this formulation thus allows the identification of active Pareto-constraints at the single objective optimum, f_1^*. Solving for $\tilde{\varepsilon_i}^*$ would then yield an explicit description of the relationship between the remaining design variables, and $\tilde{\varepsilon_i}$ at a single Pareto point. Subsequent back-substitution reveals how influential the trade-off with $\tilde{\varepsilon_i}$ is upon f_1^*. To study the whole Pareto set, however, a symbolic cost function $U(f_1,\tilde{\varepsilon})$ is introduced; $U(f_1,\tilde{\varepsilon})$ is monotonically increasing w.r.t. minimization objectives and decreasing w.r.t. maximization objectives:

$$\begin{aligned} \text{min.} \quad & \mathrm{U}(f_1^+,\tilde{\varepsilon_1}^+,...,\tilde{\varepsilon}_{k-1}^-) && (31)\\ & f_1(\mathbf{x},\tilde{\varepsilon_1}^-,...,\tilde{\varepsilon}_{k-1}^+) && (32)\\ \text{s.j.t} \quad & \mathbf{g}(\mathbf{x},\tilde{\varepsilon}) \leq 0 && (33)\\ & \mathbf{h}(\mathbf{x}) = 0 && (34) \end{aligned}$$

In minimizing cost, we can exploit its inherent monotonicity w.r.t. the objectives to identify the constraints that *bound* $\tilde{\varepsilon}$, and hence affect the topology of the Pareto set. Thus MP1 can be employed to derive the following theorem:

Theorem 3 Boundedness of $\tilde{\varepsilon_i}$
In a reduced multiobjective problem, the single objective optimum of a minimisation objective, $\tilde{\varepsilon_i}$, is determined by its greatest lower bound. Correspondingly, the lowest upper bound determines the nadir of $\tilde{\varepsilon_i}$. As such, the span of the Pareto set is in part determined by $\mathcal{X}(\tilde{\varepsilon})$.

Essentially, each reduced-objective variable is bounded by one or more Pareto-constraints across the objective space. Beyond simple optimization models, they are not necessarily critically constrained. Rather, the optimization of one $\tilde{\varepsilon}_i$ will affect the constraints of another, $\tilde{\varepsilon}_j$, if their respective glb/lub share variables, or depend on multiple $\tilde{\varepsilon}$.

Theorem 4 Conditional Activity of Pareto Constraints
In a set of Pareto-constraints that are conditionally critical for $\tilde{\varepsilon_i}$, any constraint, $g_i(\mathbf{x},\tilde{\varepsilon})$, will at least be semi-active w.r.t. $\tilde{\varepsilon_i}$ somewhere in the objective space, if it is dependant on $\underline{\bar{x}}$ or more than one reduced-objective variable. That is, unless there exists a Pareto constraint g_j such that $g_i(\mathbf{x},\tilde{\varepsilon}) < g_j(\mathbf{x},\tilde{\varepsilon}) \leq 0$ for any feasible value of $\tilde{\varepsilon}$.

The implication here is that changes in constraint activity can occur across the Pareto set if no $\tilde{\varepsilon_i}$ is critically constrained, and no Pareto-constraint is dominant. Identifying these changes in activity reveals how the objectives interact, as exemplified in Figure 2. Pareto-constraints can take on several forms, that shape the Pareto set in different ways:

- *Bound shift*: A Pareto constraint can for example shift the extremum of a monotonic variable, in effect making it a trade-off variable. Consider a problem where $f_1(x_1^+, x_2^-, \tilde{\varepsilon}_1^-)$, and one of the constraints is $g_i(x_2^+, \tilde{\varepsilon}_1^-) \equiv 0$. As $\tilde{\varepsilon}_1 \to 0$, the lub of x_1 shifts downward, worsening the optimum of f_1. Thus, g_i makes x_1 a trade-off variable w.r.t. f_1 and $\tilde{\varepsilon_1}$, with $\text{argmin}\{\tilde{\varepsilon_1}, x_2 \in \mathcal{X}\} = \underline{x_2}$.
- *Inconsistency by* ε: Pareto constraints can narrow the feasible domain of design variables that are bounded from above and below. Consider a problem with $U(f_1^+, \tilde{\varepsilon_1}^+, \tilde{\varepsilon_2}^-)$ where a variable x_1 is bounded from above by $g_1(x_1^+, \tilde{\varepsilon}_1^-) \leq 0$ and from below by $g_2(x_1^-, \tilde{\varepsilon}_2^+) \leq 0$. As $\tilde{\varepsilon} \to \tilde{\varepsilon}^*$, the feasible domain for x is reduced, meaning g_1 and g_2 become inconsistent beyond the Pareto set. Hence, g_1 and g_2 reduce objective alignment between $\tilde{\varepsilon_1}$ and $\tilde{\varepsilon_2}$, with one becoming semi-active at the resulting bi-objective Pareto frontier.
- *Multiple objectives*: Pareto constraints that depend on multiple $\tilde{\varepsilon_i}$ drastically reduce objective alignment, for instance if a constraint takes the form $g_1(\mathbf{x}, \tilde{\varepsilon}_1^+, \tilde{\varepsilon}_2^-)$.

Hence, trade-offs between the reduced-objectives are apparent in the Pareto-constraints themselves. An objective pair, $\tilde{\varepsilon_i}$ and $\tilde{\varepsilon_j}$, is in trade-off if they share a constraint of the form $g(\mathbf{x}, \tilde{\varepsilon_i}, \tilde{\varepsilon_j})$ or if their constraints become inconsistent w.r.t. to a shared variable, x, when $\tilde{\varepsilon} \to \tilde{\varepsilon}^*$. Such constraints therefore require special attention.

4.3 Analysis Procedure

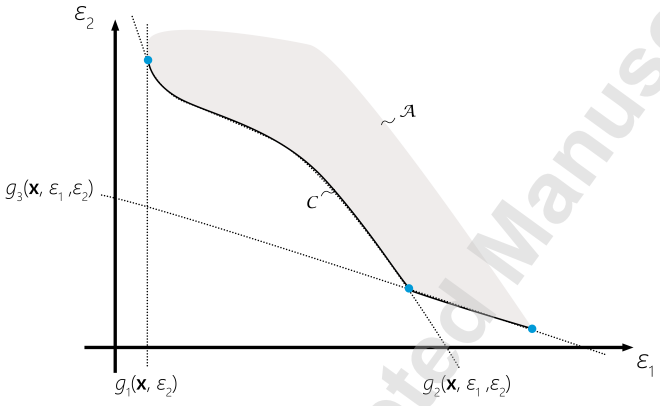

Fig. 2: An example of how the topology of a Pareto set is affected by constraints. Here the optima of ε_1 and ε_2 are determined by g_1 and g_3 respectively, with the multiobjective Pareto constraint, g_2 further reducing objective alignment

Applying the MOMA and ε-monotonicity theorems to multiobjective optimization problems allows systematic reduction down to a point where the dependencies that exist in the Pareto set are revealed. The root causes of these dependencies are, from a design perspective, the constraints and shared variables that create said dependencies. Thus, if we systematically reduce multiobjective problems and make a note of trade-off variables, the constraints that introduce them, and the constraints that bound the Pareto set, we find the relationships that in effect create, shape, and position the Pareto set. The steps in the required analysis process, which builds upon monotonicity analysis as developed by Papalambros and Wilde [36], are as follows:

1. Model the multiobjective problem as an upper-bound formulation in negative-null form.
2. Set up a monotonicity table (see e.g. [36, 41]) and assess the monotonicity of the objectives and constraints w.r.t. their variables. Make a note of any trade-off variables.
3. Use monotonicity analysis procedures to assess whether the model is well bounded [36], with the addition of the special case of the well-boundedness of trade-off variables. If the model is not well bounded, add constraints.
4. Identify constraints that are active w.r.t the primary objective and use them to reduce the model. Make a note of constraints that introduce new trade-off variables. If possible, identify the conditions under which the bound objectives become active, following Theorems 1 and 2.
5. Partially minimize the bound objectives when no further reductions to the primary objective can be made. Take care not to use constraints that potentially bound other variables regionally in the objective space. Make a note of constraints that introduce new trade-off variables.
6. When the remaining variables are either trade-off variables, non-monotonic or bounded by a conditionally critical set of constraints, run the optimization model.
7. If the numerical results reveal further globally active constraints, make further model reductions.
8. If any bound objectives are globally active, optimize said objectives out, eliminating trade-off variables in the process. The ε parameters will now appear in the remaining constraints and objective functions.
9. Treat ε parameters of the eliminated bound objectives as variables and identify the constraints that bound them. In the presence of conditional critical Pareto constraints, decompose the problem into *Pareto-Optimal Activity Cases* (see Table 1). Identify the values of ε that cause change in constraint activity or make specific constraints inconsistent. Verify this against the numerical results.

Following Theorem 4, the bounds of $\tilde{\varepsilon}$ can be interdependent, meaning that the minimisation of $\tilde{\varepsilon_i}$ affects the bounds of the remaining $\tilde{\varepsilon_j}, \forall j \neq i$, and $\underline{x}$, causing changes in activity across the Pareto set. Each change in activity implies regional dependencies between the objectives in regions of the Pareto set, as illustrated in Fig. 2. Each potential combination of active Pareto constraints hence represents a unique *Pareto Efficient Activity Case*. One can either exhaustively study all cases or focus the analysis procedure upon cases of interest. The case analysis procedure, demonstrated on a problem with minimisation objectives, is shown in Table 1. It closely resembles the parametric solution procedure developed by Wilde [36], albeit for objectives instead of design parameters. The analysis of three Pareto optimal activity case is demonstrated in Section 5.4 as a part of the case study.

Journal of Mechanical Design. Received December 17, 2020;
Accepted manuscript posted August 31, 2021. doi:10.1115/1.4052444

Table 1 - Analysis of Pareto-optimal Activity Cases

Step 1 - Case identification

To minimise $U(f_1^+, \tilde{\boldsymbol{\varepsilon}})$, identify the conditionally critical set of Pareto constraints for each $\tilde{\varepsilon}_i$. For each $g_j(\mathbf{x}, \tilde{\varepsilon}_i)$ that is conditionally critical w.r.t. $\tilde{\varepsilon}_i$:

1. Assume $g_j(\mathbf{x}, \tilde{\varepsilon}_i) \equiv 0$, and solve w.r.t. $\tilde{\varepsilon}_i$.
2. Identify the constraints that become active as a consequence of $\tilde{\varepsilon}_i \to \tilde{\varepsilon}_i{}^* \wedge g_j(\mathbf{x}, \tilde{\varepsilon}_i) \equiv 0$ and use this to reduce the expression $\tilde{\varepsilon}_i{}^*(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}})$.
3. Back substitute the eliminated variables into the remaining constraints, including the Pareto constraints that bound other reduced objective variables. If possible, identify the glb and lub of $\tilde{\varepsilon}_l, \forall l \neq i$ and use it to solve for $\underline{\tilde{\varepsilon}_l}$.

Step 2 - Case elimination

Compare the terms for $\tilde{\varepsilon}_i{}^*$ from each case:

1. If any case j is dominant, i.e. $\tilde{\varepsilon_{i,j}}{}^* > \tilde{\varepsilon_{i,k}}{}^*$ for any feasible value of $\mathbf{x}$ and $\tilde{\varepsilon}_l \, , \forall l \neq i$, then $g_k(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}})$ is either inactive or bounds another variable.
2. If any variable is revealed to be unbounded as a consequence of $g_j(\mathbf{x}, \tilde{\varepsilon}_i) \equiv 0$, then the problem is either not well-constrained, or g_j is never critical w.r.t. $\tilde{\varepsilon}_i$, meaning the case can be disregarded.
3. Identify the conditions under which the remaining cases become active. If feasible values of $\mathbf{x}$ and $\tilde{\varepsilon}_l \, , \forall l \neq i$ exist such that two cases become equivalent, i.e. $\tilde{\varepsilon_{i,j}}{}^* = \tilde{\varepsilon_{i,k}}{}^*$ then g_j and g_k are regionally active in the objective space, with a change in activity occurring at $\tilde{\varepsilon_{i,j}}{}^* = \tilde{\varepsilon_{i,k}}{}^*$. Such points are vertices of the Pareto set.

Step 3 - Case reduction

Reduce the remaining cases further to identify the extrema of the Pareto set:

1. Further minimise $\tilde{\varepsilon}_i{}^*(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}})$ by optimizing trade-off variables out, letting $\underline{x} \to \{\underline{x} \text{ if } \tilde{\varepsilon}_i(x^+), \underline{x} \text{ if } \tilde{\varepsilon}_i(x^-)\}$. If the glb and lub of $\underline{x}$ cannot be determined, the problem case can be split into sub-cases.
2. If possible, identify the cases that yield utopia and nadir points for each objective

Beyond potentially deriving single-objective optima, this procedure can be used to explicitly derive trade-off functions of the forms $f_1(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}})$ and $\tilde{\varepsilon}_i^*(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}})$. As a consequence of Theorem 1, these equations actually describe the Pareto set prior to the elimination of $\underline{\mathbf{x}}$, as any feasible value of a monotonic trade-off variable yields a Pareto point. If an objective pair $\tilde{\varepsilon}_i$ and $\tilde{\varepsilon}_j$ is in trade-off, then these reduction steps will inevitably yield minima of the form $\tilde{\varepsilon}_i{}^*(\mathbf{x}, \tilde{\varepsilon}_j{}^-)$ and $\tilde{\varepsilon}_j{}^*(\mathbf{x}, \tilde{\varepsilon}_i{}^-)$, or of the form $\tilde{\varepsilon}_i{}^*(\mathbf{x})$ and $\tilde{\varepsilon}_j{}^*(\mathbf{x})$, where $\underline{\mathbf{x}} \subset \mathbf{x}$.

Pareto-constraints that become inconsistent beyond the Pareto set are revealed as the bound objectives are optimized out. In simple problems, this degree of reduction might be reached through algebraic manipulations alone. For complex problems, however, full reduction might not be worthwhile due to the algebraic effort. Here, one can employ a more pragmatic approach by utilizing numerical results to identify additional active constraints that can be used to reduce the model further post optimality. If numerical solution reveals constraints that fulfill the Global Activity criterion from Definition 2, then such constraints can essentially be dealt with exactly as with constraints that are found to be active through MA. The globally active constraint is used to back-substitute variables post-optimality, thereby reducing the model further and giving a clearer picture of the relationships that exist at the Pareto set. As outlined in Section 3.1, this does require that the Lagrange multipliers are stored for each Pareto point to allow the evaluation of the activity of each constraint across the Pareto set.

5 Case - The Self-orienting Millimeter-scale Applicator

First published by Abramson et al. [42], the SOMA device (**S**elf-**O**rienting **M**illimeter-scale **A**pplicator) is a drug delivery device currently in development. The SOMA is designed for oral delivery of large proteins such as insulin, which cannot otherwise be administered orally, as the stomach breaks them down, and as they have poor permeability across the intestinal barrier. This substantially reduces the efficacy of such drugs, meaning they are mostly delivered via subcutaneous injections today.

Essentially, the SOMA is a pill-sized device designed to be swallowed by the user. Once in the stomach, the SOMA self-orients to a stable position due to a low center of mass and an outer shape inspired by that of leopard tortoises (*S. pardalis*) [42]. Once oriented, the device injects a biodegradable needle loaded with active pharmaceutical ingredient (API) into the *submucosa* tissue-layer of the stomach, which has a high density of blood vessels, allowing systemic uptake. This functionality is currently embodied with a linear spring actuator, held in place by a triggering mechanism (see Fig.3). The API mixture is shaped into a needle-like geometry (6) and is attached to a hub component (2) which is pre-loaded by a compression spring (4). The hub is held in place by two snap features, which are press-fit against the housing (1) by a plug (3), made out of isomalt, a dissoluble solid poly-alcohol. Once in the stomach, the device is submerged in stomach fluid, causing the plug to start dissolving to a point where the spring force pushes the snap features out of engagement. This triggers the device, with the spring pushing the needle into the stomach lining through a hole in the base (8) of the device. Until injection, the needle is kept dry in the hostile environment of the stomach by a silicone O-ring (5) and valve (7) that seal the needle inside the SOMA. The position of the centre of mass is low, as the base (8) is denser than the other parts, which aids self-orientation.

At the time of this study, the SOMA device was in the preliminary phases of design, still in the process of configuration, prototyping, and testing [42], and was yet to be tested on humans. Numerous configurations have been designed and built, with one, shown in Fig.3, showing the most

	Part name	Suffix
1	Top Housing	*H*
2	Hub	*NH*
3	Plug	*P*
4	Power Spring	*PS*
5	Sealing ring	*S*
6	Needle	*N*
7	Valve	*V*
8	Base housing	*B*

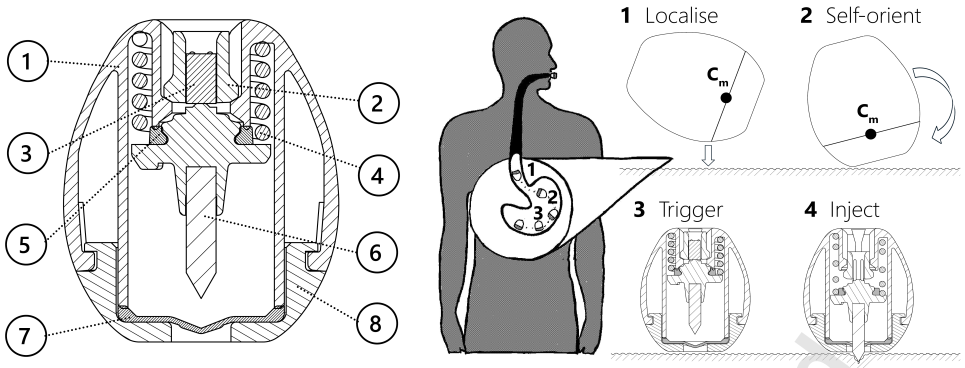

Fig. 3: An overview of the SOMA device (in part) adapted from [42]. The patient swallows the device, which self-orients inside the stomach and injects a needle of pure API into gastric tissue, detaching the needle from the device. Here the needle dissolves, resulting in systemic uptake while the device passes through the gastrointestinal tract and out of the body.

promise. Its outer shape was originally derived through optimization [42], but the inner configuration was iteratively developed, within the limits defined by the outer shape. This study focuses on the design of these internal components.

5.1 Optimization Model

The internal configuration design of the SOMA presents several design trade-off challenges. For an oral device to be viable, it needs to deliver an amount of API comparable to dosing with injection devices (e.g., insulin pens). This implies a dose of at least 80 units of insulin, which equates to a payload of approximately 2.8 mg of pure crystalline insulin. At the same time, the needle needs to be delivered reliably into a tissue layer deep enough to enable systemic uptake. The properties of the stomach lining are such that a large injection force is required to deliver the needle at the right depth. Hence, the challenge is to design a device that is small enough to be swallowable while reliably self orienting and injecting a sufficient amount of API deep enough. Furthermore, low cost and robust performance is essential. If only 1% of the world's 400M+ diabetics were treated with long-acting once-daily insulin from a SOMA, the annual production volume would be over 1.46bn devices. Given the potential volume, even slight improvements to the configuration may have a vast financial and societal impact. Understanding what causes trade-offs is hence highly valuable.

Four objectives were modelled for this study: swallowability, the height of the center of mass, API capacity, and injection depth. Given the early stage of development, the goal was to develop the simplest meaningful model, leading to the following simplifications:

1. As the focus is on the internal configuration, the outer shape is kept constant. Hence it is sufficient to optimize the vertical position of the center of mass to improve self-orientation.
2. Swallowability is proportional with the minor diameter [43], which is equivalent to d_{t1} illustrated in Fig. 4.
3. Injection depth is dependant on the mechanical properties of gastric tissue, the velocity at which the needle impacts gastric tissue, the diameter of the needle, and the sharpness of its tip. The needle is made from compacted protein, meaning that the sharper the tip, the more costly and sensitive the production process. Hence it is preferable to achieve a sufficient depth with a large velocity and not rely on sharpness. The impact velocity is therefore modelled as a maximizing objective since the injection depth increases monotonically with it.

The resulting initial optimization model is

$$\min \quad f_1(\mathbf{x}) = -\frac{\sum_{p=1}^{p=8} m_p \cdot (C_p + Z_p)}{(l_{t1}+l_{t2}+l_{b1}) \cdot \sum_{p=1}^{p=8} m_p} \tag{35}$$

$$\text{s.j.t} \quad c_1(\mathbf{x};\boldsymbol{\varepsilon}_1) = d_{t1} - \varepsilon_1 \leq 0 \tag{36}$$

$$c_2(\mathbf{x};\boldsymbol{\varepsilon}_2) = \varepsilon_2 - \rho\frac{\pi}{4} d_{n1}^2 \left(l_{n1} + \frac{1}{3}\cdot l_{n2} \right) \leq 0 \tag{37}$$

$$c_3(\mathbf{x};\boldsymbol{\varepsilon}_3) = \varepsilon_3 - \sqrt{2\left(g + \frac{F_s}{m_{acc}}\right) z_{acc}} \leq 0 \tag{38}$$

$$\mathbf{g}(\mathbf{x}) \leq 0 \tag{39}$$

$$\mathbf{h}(\mathbf{x}) = 0 \tag{40}$$

$$\mathbf{x}, \boldsymbol{\varepsilon} \in \mathbb{P} \tag{41}$$

where:

- f_1 is the self-orientation objective, which maximises the distance, Z_{cm}, between the top of the device and the system centre of mass, C_m, relative to the total height of the device, $l_{t1}+l_{t2}+l_{b1}$. Here, m_p, C_p, and Z_p are intermediate functions, with m_p describing the mass of each part in the device, C_p the centre of mass in each part, and Z_p the axial distance of each part from the top of the device. Expressions for m_p and C_p were derived explicitly using geometric idealisations (e.g. ellipsoids and cylinders) to describe the shape of the parts while accounting for fea-

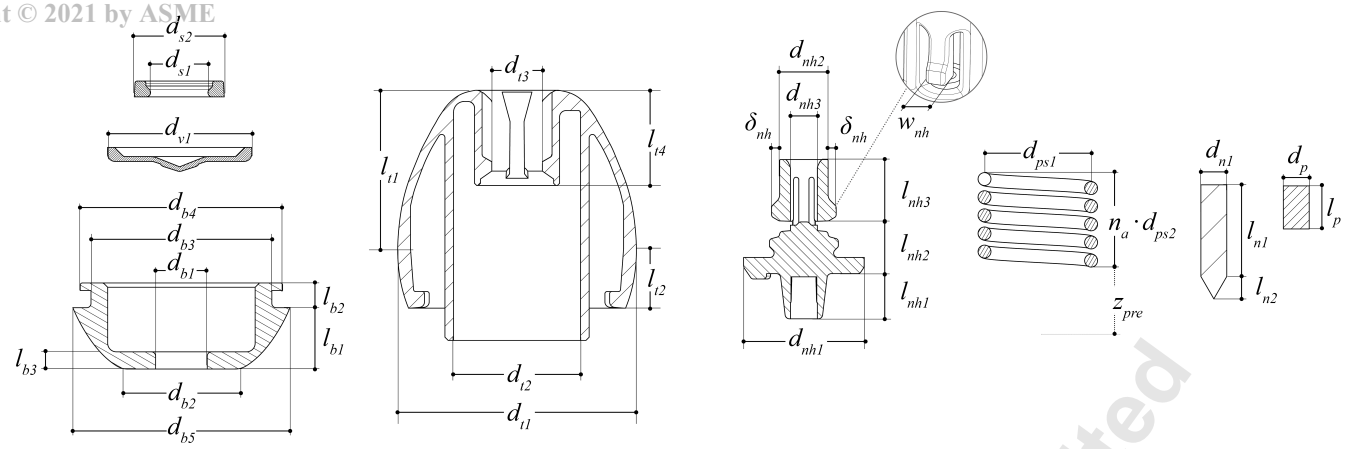

Fig. 4: Design variables of relevance to the constraints used to demonstrate the trade-off root cause analysis. l denotes length variables, d -diameters, δ -deflections, and z denotes vertical positions

tures such as rounds, draft, and snap interfaces. The objective function was then verified against CAD models of the SOMA in several sizes, with a max. deviation of 0.83% compared to the mass distribution analysis done in the CAD system (PTC Creo 4.0). This was deemed to be acceptable for the purposes of the study.

- c_1 is the bound objective for size minimization, where d_{t1} is the diameter of the device.
- c_2 is the bound API capacity objective, which maximises the mass of the needle, and is hence a simple volumetric expression. The design variables are illustrated in fig. 4, while ρ is the density of the compacted API.
- c_3 is the bound impact velocity objective, maximising the velocity of impact between needle and tissue. Here, F_s is the spring force, m_{acc} the mass that is accelerated during injection, z_{acc} the acceleration stroke between the initial position of the needle tip and the gastric tissue), and g the gravitational acceleration. Non-linearities such as device recoil (which counteracts z_{acc},), self-accelerating spring mass, and the reduction of F_s as a function of displacement, are also accounted for.

The inequality constraints, $\mathbf{g}(\mathbf{x})$, encompass criteria such as geometric fits (e.g., radial and axial part fits in use and assembly), manufacturability (e.g. min. wall thicknesses and feature size), and structural load cases (e.g., spring stress, interface stresses, needle buckling). The equality constraints, $\mathbf{h}(\mathbf{x})$, mostly account for the shape of the device. In total, there are 52 inequality constraints, 7 equality constraints, 45 design variables, and multiple parameters. For the present demonstration, further model details are not necessary and are omitted for brevity.

ods described in Section 4. These constraints ensure that the parts fit together radially - i.e., that the plug fits into the hub, the hub into the top housing, the spring around the trigger system, and the top housing in the base:

$$h_1(l_{t1}^+, d_{t1}^-) = l_{t1} - d_{t1} C_T = 0 \tag{42}$$
$$h_2(d_{nh3}^+, d_p^-) = d_{nh3} - d_p - 2R_{cl} = 0 \tag{43}$$
$$h_3(d_{b4}^+, d_{b5}^-) = d_{b4} + 2R_{wt} + 2R_{cl} - d_{b5} = 0 \tag{44}$$
$$g_1(d_{t1}^-, l_{t1}^-, l_{t2}^+, d_{b5}^+) = d_{b5} - \sqrt{\frac{2(l_{t1}-l_{t2})d_{t1}^2}{l_{t1}} - \frac{(l_{t1}-l_{t2})^2 d_{t1}^2}{l_{t1}^2}} \leq 0 \tag{45}$$
$$g_2(d_{b3}^+, d_{b4}^-) = d_{b3} + 2R_{ov} - d_{b4} \leq 0 \tag{46}$$
$$g_3(d_{t2}^+, d_{b3}^-) = d_{t2} + 4R_{wt} + 2R_{cl} - d_{b3} \leq 0 \tag{47}$$
$$g_4(d_{t2}^-, d_{ps1}^+, d_{ps2}^+) = d_{ps1} + d_{ps2} + 2R_{cl} - d_{t2} \leq 0 \tag{48}$$
$$g_5(d_{t3}^+, \delta_{nh}^+, d_{ps2}^+, d_{ps1}^-) = d_{t3} + 2(\delta_{nh} + R_{cl} + R_{wt}) + d_{ps2} - d_{ps1} \leq 0 \tag{49}$$
$$g_6(d_{nh2}^+, d_{t3}^-) = d_{nh2} - d_{t3} + 2R_{cl} \leq 0 \tag{50}$$
$$g_7(d_{nh2}^-, d_{nh3}^+) = d_{nh3} + 2R_{wt} - d_{nh2} \leq 0 \tag{51}$$
$$g_8(\delta_{nh}^-) = 0.3\text{mm} - \delta_{nh} \leq 0 \tag{52}$$
$$g_9(d_p^-) = 1\text{mm} - d_p \leq 0 \tag{53}$$
$$g_{10}(\delta_{nh}^+, d_{nh3}^-) = 2\delta_{nh} + R_{cl} - d_{nh3} \leq 0 \tag{54}$$
$$g_{11}(d_{ps1}^-, d_{ps2}^+, n_a^-, z_{pre}^+, \delta_{nh}^-, w_{nh}^-) = \frac{z_{pre}\cos(\Theta_{nh}) G_{st} d_{ps2}^4}{16 d_{ps1}^3 n_a \delta_{nh} w_{nh}} - \sigma_c \leq 0 \tag{55}$$

5.2 Monotonicity Analysis

Using MOMA, the model was reduced prior to computation down to 18 design variables, 28 inequality constraints, 2 equality constraints and 4 objectives, 3 of them bound. An important set of constraints - the radial fits between the parts - will be used to demonstrate the application of the method

where C_T is the aspect ratio between the diameter and height of the top housing, R_{wt} is the min. wall thickness, R_{ov} the min. radial interface overlap, R_{cl} the min. radial clearance, Θ_{nh} the contact angle in the trigger interface, G_{st} the shear modulus of the spring steel, and σ_c the allowable stress in the trigger interface. The variables are illustrated in fig. 4.

The monotonicity of the objectives w.r.t these variables is:

$$f(l_{t1}^-, l_{t2}^-, d_{t1}^-, d_{t2}^+, d_{t3}^+, d_{b3}^+, d_{b4}^+, d_{ps1}^+, d_{ps2}^+, d_{nh2}^+, d_{nh3}^-, \delta_{nh}^+, d_p^+) \tag{56}$$

$$c_1(d_{t1}^+) \tag{57}$$

$$c_3(l_{t1}^-, l_{t2}^-, d_{ps1}^+, d_{ps2}^-, d_{nh2}^+, \delta_{nh}^+) \tag{58}$$

c_2 is independent of these variables in the initial model. The only trade-off variables that are visible so far, are the device diameter, d_{t1}, and the spring wire diameter d_{ps2}, as exhibited by their opposite monotonicity in the objectives. To begin reducing the problem, we first use h_1 to eliminate l_{t1}, h_2 to eliminate d_{nh3}, and h_3 to eliminate d_{b5}. Furthermore, g_2, g_3, g_4, and g_6 are critical w.r.t. d_{b4}, d_{b3}, d_{t2} and d_{t3} respectively, meaning MP1 can be applied to eliminate them. After back-substitution, the objectives and constraints have changed:

$$\begin{aligned}
\text{min.} \quad & f(l_{t2}^-, d_{t1}^-, d_{t3}^+, d_{ps1}^+, d_{ps2}^+, d_{nh2}^+, \delta_{nh}^+, d_p^+) && (59)\\
\text{s.j.t.} \quad & c_1(d_{t1}^+; \varepsilon_1) && (60)\\
& c_3(d_{t1}^-, l_{t2}^-, d_{ps1}^+, d_{ps2}^-, d_{nh2}^+, \delta_{nh}^+; \varepsilon_3) && (61)\\
& g_1(d_{t1}^-, l_{t2}^+, d_{ps1}^+, d_{ps2}^+) = d_{ps1} + d_{ps2} \\
& + 6R_{wt]} + 6R_{cl} + 2R_{ov} \\
& - \sqrt{\frac{2(C_T d_{t1} - l_{t2}) d_{t1}}{C_T} - \frac{(C_T d_{t1} - l_{t2})^2}{C_T^2}} \leq 0 && (62)\\
& g_5(d_{nh2}^+, \delta_{nh}^+, d_{ps2}^+, d_{ps1}^-) = d_{nh2} + 2\delta_{nh} + d_{ps2} \\
& - d_{ps1} + 4R_{cl} + 2R_{wt} \leq 0 && (63)\\
& g_7(d_{nh2}^-, d_p^+) = d_p + 2R_{wt} + 2R_{cl} - d_{nh2} \leq 0 && (64)\\
& g_8(\delta_{nh}^-) = 0.3\text{mm} - \delta_{nh} \leq 0 && (65)\\
& g_9(d_p^-) = 1\text{mm} - d_p \leq 0 && (66)\\
& g_{10}(\delta_{nh}^+, d_p^-) = 2\delta_{nh} - R_{cl} - d_p \leq 0 && (67)\\
& g_{11}(d_{ps1}^-, d_{ps2}^+, n_a^-, z_{pre}^+, \delta_{nh}^-, w_{nh}^-) \\
& = \frac{z_{pre} \cos(\Theta_{nh}) G_{st} d_{ps2}^4}{16 d_{ps1}^3 n_a \delta_{nh} w_{nh}} - \sigma_c \leq 0 && (68)
\end{aligned}$$

So far, trade-offs have been revealed between size, c_1, and both impact velocity, c_3, and self-orientation, f_1, through d_{t1}, and between impact velocity and self-orientation through the spring wire diameter d_{ps2}. Increasing the wire diameter increases spring force and hence velocity, but it also increases the spring mass, shifting the system centre of mass upward.

Following Theorems 1 and 2, the boundedness of d_{t1} reveals important information about the SOMA device. Firstly, $c_1(\mathbf{x}; \varepsilon_1) \equiv 0$ for any $\varepsilon_L(1) < \varepsilon_1 < \varepsilon_U(1)$. Secondly, the only *non-objective lower bound* for d_{t1} is g_1, the constraint that ensures that the top and base housings fit together radially. This means that g_1 will be active at the single-objective minimum. Looking at eq. 62, it is evident that all the objectives cannot be minimised simultaneously, without reaching a point where $c_1(d_{t1}^+) < g_1(d_{t1}^-)$ meaning that $\mathcal{X}(d_{t1}) = \emptyset$. Hence, g_1 is at least semi-active in any bi-objective Pareto-front involving the size objective, c_1. As a consequence, l_{t2} is a trade-off variable when g_1 is active, and d_{ps2} also becomes a trade-off variable w.r.t. size. The implication for design is, that the further the mating surface between top and base is moved downward, the less space there is available for the spring mechanism. The only harmonious variable left in g_1, is d_{ps1}; identifying its' glb may reveal additional variables that contribute to the trade-offs between f_1, c_1 and c_3.

The remaining variables, including d_{ps1} have a conditionally critical set of constraints. Specifically, the spring, hub, and plug variables are potentially bound by inequality constraints relating to the yield stress of different parts, while the top housing variables are also involved in the axial fit constraints. Hence they cannot be eliminated without substantial algebraic manipulation to identify the glb or lub of each variable, meaning it is more efficient to identify the remaining active constraints numerically.

5.3 Numerical Results

The upper bound problem was solved 200,000 times using the SQP fmincon routine in MATLAB2019R [44] for different values of ε sampled from a quasi-random set (a leaped Halton set) distributed between $\varepsilon_{\mathbf{L}} = [8.5\text{mm}; 1.5\text{mg}; 10\text{m/s}]$ and $\varepsilon_{\mathbf{U}} = [11.5\text{mm}; 4.5\text{mg}; 30\text{m/s}]$. These values were set based on input from the SOMA team in Novo Nordisk. The results are shown in Fig. 5 and Table 1.

Objective	Optimum	Nadir	λ_{min}	λ_{max}
f_1	-0.78h	-0.64h	-	-
ε_1	8.67 mm	11.50 mm	0.0131	2.7708
ε_2	4.50 mg	1.50 mg	0.0016	0.4355
ε_3	28.34 m/s	10 m/s	0.0008	0.3795

Table 2: Numerical results

As the minimum λ values of each bound objectives are positive, they are active in the entire sampling region, and all feasible solutions are Pareto-optimal. As seen in Fig. 5, all four objectives are in trade-off with each other. Furthermore, g_5 and g_7 are globally active. Interestingly, g_1 and g_5 were violated in every infeasible iteration, pointing to inconsistent constraints beyond the Pareto set.

While 42% of the iterations yielded feasible, optimal solutions, the other 58% failed to identify a feasible solution. A few measures were taken to verify the model that led to these results. Firstly, the validity of the MA was assessed by running the original unreduced model over a narrower range of ε values. This led to the same results as with the reduced model. Secondly, a constraint satisfaction problem was run for all the failed iterations. This was used to search for a feasible solution to use as a new initial guess in a re-run with the

Journal of Mechanical Design. Received December 17, 2020;
Accepted manuscript posted August 31, 2021. doi:10.1115/1.4052444

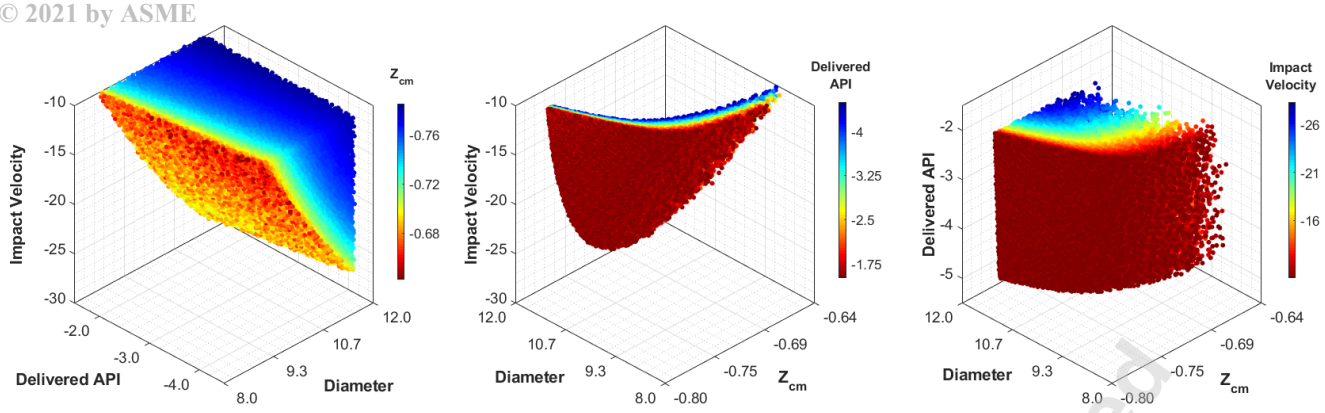

Fig. 5: Different projections of the 4D-Pareto set, where the 4th objective is visualised with a color map

same values of ε. Only 1.8% of these cases identified a new feasible initial guess, and these all yielded a Pareto point subsequently. This indicates that the remaining iterations indeed failed due to a lack of feasible domain caused by inconsistent constraints, meaning that the approximate Pareto-frontier of the sampled objectives had indeed been identified.

We note that the US-FDA generally recommends pills and capsules stay below a standard 00-size [43], which has a 8.35mm diameter, while the largest standard size, 000 capsules, are 9.91mm in diameter. Complications from swallowing pills start at about 8mm dia. and grows substantially beyond a 11mm dia. [45]. Initial work in the SOMA project has revealed that the impact velocity is critically important to the bioavailability of the delivered API (the % of the administered drug that reaches systemic circulation). It is also critical to the robustness and cost of the shaping of the needle geometry, as a low velocity results in a need for a sharper tip. Thus, the trade-off between size and velocity ultimately affects the amount of drug that can be delivered in a swallowable device and the cost of treatment. Model reduction using the numerical results reveals the cause of this trade-off.

5.4 εMA and Pareto Optimal Case Analysis

The global activity of g_5 and g_7 is used to eliminate $d^*_{ps1} = d_p + 2\delta_{nh} + d_{ps2} + 4R_{cl} + 2R_{wt}$ and $d^*_{nh2} = d_p + 2(R_{wt} + R_{cl})$ further reducing equations 59-68. Globally active axial fit and mechanical yield constraints allow the elimination of n_a and w_{nh}, the back substitution of which results in the elimination of z_{pre} from g_{11}, which will be handled implicitly from here on. Subsequently, the globally active bound size objective $c_1(d^+_{t1};\varepsilon_1)$ is used to eliminate d_{t1}, introducing $\tilde{\varepsilon_1}$ into g_1 and two objectives, f_1 and c_3:

$$\text{min.} \quad f(l^-_{t2}, d^+_{t3}, d^+_{ps2}, \delta^+_{nh}, d^+_p, \tilde{\varepsilon_1}^-) \tag{69}$$

$$\text{s.j.t.} \quad c_3(l^-_{t2}, d^-_{ps2}, d^+_p, \delta^+_{nh}, \tilde{\varepsilon_1}^-; \varepsilon_3) \tag{70}$$

$$g_1(\tilde{\varepsilon_1}^-, l^+_{t2}, d^+_{ps2}, d^+_p, \delta^+_{nh}) = 2d_{ps2} + 2\delta_{nh} + d_p + 10R_{wt} + 13R_{cl} + 2R_{ov} - \sqrt{\frac{2(C_T\tilde{\varepsilon_1} - l_{t2})\tilde{\varepsilon_1}}{C_T} - \frac{(C_T\tilde{\varepsilon_1} - l_{t2})^2}{C_T^2}} \leq 0 \tag{71}$$

$$g_8(\delta^-_{nh}) = 0.3\text{mm} - \delta_{nh} \leq 0 \tag{72}$$

$$g_9(d^-_p) = 1\text{mm} - d_p \leq 0 \tag{73}$$

$$g_{10}(\delta^+_{nh}, d^-_p) = 2\delta_{nh} - R_{cl} - d_p \leq 0 \tag{74}$$

$$g_{11}(d^+_{ps2}, d^-_p, \delta^-_{nh}) \leq 0 \tag{75}$$

As expected, g_1 makes l_{t2} a trade-off variable, as $\overline{l_{t2}} \to 0$ as $\tilde{\varepsilon_1} \to 0$ when $g_1 \equiv 0$,and given that $f(l^-_{t2})$ and $c_3(l^-_{t2})$. The velocity objective c_3 has not been optimized out, as there is no closed form solution to $c_3(\mathbf{x}, \tilde{\varepsilon_1}; \varepsilon_3) \equiv 0$ w.r.t any $\underline{x}$. Its elimination would involve solving for d_{ps2}, as it is critically constrained from below by c_3 and is shared with the largest number of constraint functions that remain in the model. This would make g_1 a multiobjective Pareto constraint. Therefore, g_1 is involved in three Pareto-optimal activity cases; when g_1 bounds $\tilde{\varepsilon_1}$, d_{ps2}, and l_{t2}. Looking at these cases in detail using the procedure from Table 1, reveals the root cause of the shape and position of the bi-objective Pareto front between size and velocity.

Activity case 1: Smallest Possible Device, $\text{U}(\tilde{\varepsilon_1}^+)$
Here g_1 determines $\underline{\tilde{\varepsilon_1}}^*$ and yields the optimal size. Eliminating l_{t2} allows a closed form solution for $\tilde{\varepsilon_1}$ using Eq. 71. Letting $l_{t2} \to 0$, implying that the mating surface between top and base is located at the widest point of the device, allows the smallest $\tilde{\varepsilon_1}$. Inserting this, and the parameter values, $C_t = 0.68$, $R_{wt} = 0.45$mm, $R_{cl} = 0.1$mm, $R_{ov} = 0.6$mm yields a reduced expression:

$$g_1(\tilde{\varepsilon_1}^-, d^+_{ps2}, d^+_p, \delta^+_{nh}) = 2(d_{ps2} + \delta_{nh}) + d_p + 7\text{mm} - \tilde{\varepsilon_1} \tag{76}$$

$$\Rightarrow \underline{\tilde{\varepsilon_1}}(d^+_{ps2}, \delta^+_{nh}, d^+_p) = 2d_{ps2} + 2\delta_{nh} + d_p + 7\text{mm} \tag{77}$$

$$\Rightarrow g_8(\delta_{nh^-}) \equiv 0 \wedge g_9(d_p^-) \equiv 0 \quad (78)$$

$$\Leftrightarrow \underline{\delta_{nh}} = 0.3\text{mm}, \underline{d_p} = 1\text{mm} \quad (79)$$

$$\Leftrightarrow \underline{\tilde{\varepsilon}_1}^* = 2d_{ps2} + 8.6\text{mm} \quad (80)$$

As d_{ps2} is a trade-off variable, minimising $\tilde{\varepsilon}_1$ will lead to a point where $g_{11} < g_8$ and $g_{11} < g_9$ w.r.t. d_p and δ_{nh}. This results in g_8 and g_9 becoming active, leading to the back-substitution performed in eqs. 77-80. As a result, $\mathcal{X}(d_{p2})$ is narrowed at the Pareto frontier between size and velocity, given that these reductions leave $g_{11}(d_{ps2}^+)$ and $c_3(d_{ps2}^-, \tilde{\varepsilon_1}^-; \varepsilon_3)$. Further, c_3 is critical w.r.t. d_{ps2}, meaning that $\varepsilon_L(3)$ ultimately determines the lowest feasible value of d_{ps2}, and hence $\underline{\tilde{\varepsilon}_1}^*$.

Activity case 2: Maximum Impact Velocity, $\text{U}(\tilde{\varepsilon_3}^-)$
Here g_1 determines $\overline{d_{ps2}}$. As c_3 is monotonically decreasing w.r.t. d_{ps2} to the power of 4, its supremum yields the single-objective optimal impact velocity. Thus, the same parameter values and value of l_{t2} can be inserted as in Case 1. Yet, as opposed to Case 1, g_8 and g_9 are inactive, as pushing d_{ps2} to its upper limit makes $g_{11}(d_{ps2}^+, d_p^-, \delta_{nh}^-) \leq 0$ active. g_{11} is a interface stress criterion, and because the spring force grows with the wire diameter, the dimensions that determine the area - d_p and δ_{nh} increase correspondingly. This in turn makes $g_{10}(\delta_{nh}^+, d_p^-)$ active, as $g_9 < g_{10}$ for any $\delta_{nh} > 0.45$mm. These activities yield:

$$g_1(\tilde{\varepsilon_1}^-, d_{ps2}^+, d_p^+, \delta_{nh}^+) = 2(d_{ps2} + \delta_{nh}) + d_p + 7\text{mm} - \tilde{\varepsilon_1} \quad (81)$$

$$\Rightarrow \overline{d_{ps2}}(\tilde{\varepsilon}_1^+, \delta_{nh}^-, d_p^-) = 0.5(\tilde{\varepsilon_1} - d_p - 2\delta_{nh} - 7\text{mm}) \quad (82)$$

$$g_8 \equiv 0 \Rightarrow d_p^* = 2\delta_{nh} - R_{cl} \quad (83)$$

$$g_{11} \equiv 0 \Rightarrow \delta_{nh} = \delta_{nh}(d_{ps2}^+; \sigma_c, \sigma_y) \quad (84)$$

$$\Rightarrow \overline{d_{ps2}} = 0.5(\tilde{\varepsilon_1} - 4\delta_{nh}(d_{ps2}^+; \sigma_c, \sigma_y) - 6.9mm) \quad (85)$$

where g_{11} has been used to implicitly eliminate the overlap δ_{nh} between hub and top housing in the load bearing trigger interface as no closed form solution exists; σ_y is the yield stress of the spring, and σ_c is the allowable static stress in the trigger interface. This substitution reveals a feedback coupling; as the wire diameter d_{ps2} is increased, so does the required load bearing area, reducing the space available for the spring wire in a device of a given size, $\tilde{\varepsilon_1}$.

Activity case 3: Lowest Possible Center of Mass, $f_1(l_{t2}^-)$
Here g_1 determines $\overline{l_{t2}}$. As $f_1(l_{t2}^-)$, this case occurs at the single objective optimal self-orientation. Given the non-linearity of Eq. 71 w.r.t. l_{t2}, the variable is best eliminated implicitly, yielding $\overline{l_{t2}} = l_{t2}(\tilde{\varepsilon_1}^+, d_p^-, d_{ps2}^-, \delta_{nh}^-)$. As a consequence d_{ps2} is bounded by c_3, d_p by either g_9 or g_{10}, and δ_{nh} by either g_8 or g_{11}. Furthermore, $\tilde{\varepsilon_1} = \varepsilon_U(1)$, as no constraint bounds $\tilde{\varepsilon_1}$ from above. As discussed in Section 5.2, g_1 reduces objective alignment between self-orientation and size and impact velocity respectively.

5.5 Design Implications

These Pareto-optimal activity cases demonstrate the root cause of the position and shape of the bi-objective Pareto front between size and impact velocity. The smaller the coiling diameter d_{ps1} of the spring, the more spring force (and hence impact velocity) and the smaller a device. Given that the spring needs to fit inside the diameter of the guiding cylinder d_{t2} and around the trigger system, g_4 and g_5 are active, meaning a harmonious variable is minimised out, $d_{ps1}^* = d_p + d_{ps2} + 2\delta_{nh} + 2R_{wt} + 4R_{cl}$ introducing a trade-off variable, d_{ps2}, and δ_{nh} into g_1.

We can see from g_1 that as the coiling diameter is reduced and the wire diameter increased, the available space left for the trigger system is reduced. The trigger system distributes the spring force over an area equal to $A = 2\delta_{nh} w_{nh} = 2\delta_{nh}(d_p - R_{cl} - R_{wt})$, and stiffening the spring increases the spring force but also reduces the load-bearing area, see Equation 85. With d_{ps2} being the variable with the largest influence on impact velocity (to the power of 4), and d_{ps1} being the second most (to the power of 3), activities of g_5, g_7 and g_{10} are ultimately the main driver of the trade-off. Had the spring and the trigger geometry existed in different cross-sections, the alignment between the two objectives might be drastically improved. This would correspond to d_{ps1}^* being independent of δ_{nh}, not only shifting the glb of d_{ps1} downward, but also removing the contribution of δ_{nh} to the constraint that determines $\underline{\tilde{\varepsilon}_1}$, improving size and impact velocity simultaneously. This also increases $\overline{\delta_{nh}}$. Furthermore, the design of the snap-interface between the top- and base housings also influences objective alignment between self-orientation, size, and impact velocity.

6 Discussion

Pareto set dependency analysis presents an optimization-focused alternative to current techniques for dependency analysis (e.g. DSM and Axiomatic Design [7]), with MOMA allowing the analysis of dependencies unique to the optimum by addressing the impact of constraints directly, and εMA revealing regional dependencies that shape the Pareto set. As an added benefit, the procedure for systematic reduction of multiobjective problems helps reduce computational cost due to the elimination of constraints and variables. In computationally expensive problems, this pre- and post optimality analysis procedure may also help reveal insights that would be too costly to reach computationally, e.g., describing certain relationships that would otherwise only come to light if the Pareto set is exhaustively identified.

From a design optimization perspective, Pareto set dependency analysis is a rigorous approach to exploring the limitations of a given configuration. The definitions and theorems presented allow the systematic identification of trade-off variables, active bound objectives and Pareto-constraints, and the constraints that introduce new trade-off variables. In doing so, one determines what objectives are in trade-off and, even more importantly, the underlying root causes of these trade-offs, clearly exposing the weaknesses in the configura-

tion design. An example from the SOMA case is the trade-off between size and impact velocity, which is in part caused by the spring needing to fit around the trigger.

If ε-monotonicity analysis is applied exhaustively to all Pareto-optimal activity cases and all active Pareto-constraints are be identified, the Pareto set is essentially derived explicitly. This is similar to the approaches developed by Gobbi et al. [30] and Mastinu et al. [31] for the explicit derivation of the Pareto set of lower-dimensional problems using back-substitution of ε. However, ε-monotonicity analysis goes beyond this to allow the identification of the drivers of trade-offs. Furthermore, the analysis is opportunistic; it can be performed partially and still provide useful insights. It is neither necessary that every case is studied nor that all bound objectives are eliminated in order for the analyst to identify some of the dependencies that reduce objective alignment and may guide redesign. From this, it follows that one might view the presented analysis as a way of *checking the design* ahead of computation or *explaining the results* after computation. For example, one can use the theorems to reduce a multiobjective model after computation using numerical activity information, should the model at hand be too large or complex to reduce through algebraic analysis alone. Alternatively, one could skip computation should initial MOMA reveal drivers of trade-off that might be eliminated through a change in configuration design.

The opportunistic nature of monotonicity analysis (MA) also reveals the key limitations of the methodology we have developed. Firstly, not all problems are monotonic or even differentiable. This might be dealt with using techniques for local [46] and regional MA [36] if the expressions are regionally differentiable. This comes at the cost of increased analysis effort, which might be offset using sampling-based computational experiments (e.g., DoE) to reveal regional properties in non-monotonic or non-algebraic problems.

Secondly, MA mostly relies on algebraic manipulations, and some design problems are too complex to be expressed algebraically. Yet, that certain aspects of a design's behaviour can only be expressed numerically does not necessarily imply a lack of monotonicity, e.g. as is often the case with stiffness and deflection. In such situations, implicit MA [36] procedures, and meta-models might be used. This would reveal the variables and constraints that cause trade-offs, albeit without the derivation of explicit expressions of the relationships that exist in the Pareto set.

It is also well accepted that purely algebraic models can play a substantial role in practice [47], in both conceptual and configuration design. These phases are often characterized by a lack of sophisticated quantitative models to support decision making due to requirement uncertainty and the modelling effort involved, compared to how quickly and often the design changes [48]. Configuration design also often involves the combination and arrangement of well-known types of parts and modules, which might be described algebraically, for example as seen in the machine elements, engines, hydraulics, and thermal systems.

Finally, the effort in analysis is proportional to the number of objectives, constraints, and variables in the problem. This effort is amplified by non-monotonicity and by regionally active constraints. Thus the bookkeeping and algebraic effort required to reduce a multiobjective model systematically may be prohibitive if the problem is large.. Here, the use of symbolic solvers can help reduce the effort in back-substitution and model reduction. In that regard, quite some work was done (with some success) on automating MA in the 1980s [46, 49]. In the view of the authors, there is potential in attempting to improve automation of MA (and thus MOMA and εMA) by leveraging the achievements made in computational techniques such as machine learning, AI, and data analysis and clustering, since MA methods where last in vogue. Given the advances in meta-modelling since then, it is also not unlikely that more complicated non-algebraic models might be analysed using the methods described in this paper. Hence the (partial) automation of MOMA and εMA is possible future work.

Ultimately, the value of this methodology comes down to the cost involved in analysis vs. the expected benefit in discovering better configurations. As discussed in the introduction, trade-off knowledge and decision-making are largely experience-driven in early stage design. Finger and Dixon [50] highlighted the dearth of quantitative design analysis and evaluation methods for the early stages, especially those which allow multiobjective analysis and support the identification of alternative configurations and concepts. The presented methodology addresses some of these unmet needs.

When the cost vs. benefit estimate noted above is favorable, the methodology might be used to target iterative configuration redesign efforts, to guide morphological studies to identify alternative solutions, or simply to explain the results of an optimization model from a design perspective. For small, tightly coupled systems such as the SOMA device, the value in discovering the non-obvious influence of certain variables and constraints on design trade-offs, amply justifies the analysis effort. For a larger system, the methodology can be worthwhile if the system is obviously monotonic or if the optimization model is constructed at an architectural level of abstraction that limits the number of design variables and expressions to analyze.

7 Conclusions

Trade-offs between objectives are an inevitable challenge in mechanical design. In multiobjective optimization, most prior work focused on quantifying these trade-offs, but there has been little prior work on their causes. Understanding this causality provides insights for improvements in proportional and, most importantly, in configuration design.

We demonstrated extensions to monotonicity analysis specific to multiobjective problems that allow rigorous identification of the constraints and variables contributing to trade-offs. Using the upper bound formulation for multiobjective problems, we extended monotonicity analysis and its application, proposing a novel procedure, ε-monotonicity analysis, to identify and study the constraints bounding the Pareto set. The methodology leads to deeper insights into the strengths and weaknesses of a design configura-

tion. We demonstrated the methodology on the early-stage design of the SOMA device, finding trade-offs that are in part caused by a load-bearing interface needing to fit inside the spring that exerts said load. Such insights may guide redesign resulting in improvements in performance beyond what is achievable through proportional design, i.e., beyond the Pareto set for the particular embodiment. A systematic redesign procedure to identify such improvements utilizing the output of the presented analysis method will be treated in a subsequent publication.

Acknowledgements

The authors would like to thank the Danish Innovations Fund and the Novo Nordisk STAR-programme for funding this research project (grant nr. 7038-00221B), Novo Nordisk for sharing design information and data, Asst. Prof. Giovanni Traverso of MIT and his colleagues for their helpful comments and input, and the University of Michigan Donald C. Graham Endowed Chair for providing visiting scholar support. The opinions presented here are solely those of the authors.

References

[1] W Brian Arthur. "Why Do Things Become More Complex?" In: *Scientific American* 268.5 (1993), pp. 144–144. DOI: 10.1038/scientificamerican0593-144.

[2] Durward K Sobek, Allen C Ward, and Jeffrey K Liker. "Toyota ' s Principles of Set-Based Concurrent Engineering Toyota ' s Principles of Set-Based Concurrent Engineering". In: *Sloan Management Review* 40.2 (1999), pp. 67–83.

[3] Saeema Ahmed, Ken M. Wallace, and Luciënne T.M. Blessing. "Understanding the differences between how novice and experienced designers approach design tasks". In: *Research in Engineering Design* 14.1 (2003), pp. 1–11. DOI: 10.1007/s00163-002-0023-z.

[4] G Pahl and W Beitz. *Engineering design — A systematic approach*. 1999. DOI: 10.1016/0261-3069(96)84970-3.

[5] Panos Y Papalambros and Kristina Shea. "Creating Structural Configurations". In: *Formal Engineering Design Synthesis*. Cambridge University Press, Nov. 2001, pp. 93–125. DOI: 10.1017/CBO9780511529627.007.

[6] David G. Ullman, Thomas G. Dietterich, and Larry A. Stauffer. "A model of the mechanical design process based on empirical data". In: *Artificial Intelligence for Engineering, Design, Analysis and Manufacturing* 2.1 (1988), pp. 33–52. DOI: 10.1017/S0890060400000536.

[7] Nam P. Suh. "Axiomatic Design Theory for Systems". In: *Research in Engineering Design - Theory, Applications, and Concurrent Engineering* 10.4 (1998), pp. 189–209. DOI: 10.1007/s001639870001.

[8] Hillary G. Sillitto. "On systems architects and systems architecting: Some thoughts on explaining and improving the art and science of systems architecting". In: *19th Annual International Symposium of the International Council on Systems Engineering, INCOSE 2009*. 2009. DOI: 10.1002/j.2334-5837.2009.tb00995.x.

[9] M M Andreasen and T J Howard. "Is Engineering Design Disappearing from Design Research ?" In: *The Future of Design Methodology*. Ed. by H Birkhofer. London: Springer Verlag, 2011. Chap. 2, pp. 21–34. DOI: 10.1007/978-0-85729-615-3.

[10] Ching-Shin Norman Shiau and Jeremy J Michalek. "Should Designers Worry About Market Systems?" In: *Journal of Mechanical Design* 131.1 (Dec. 2008). DOI: 10.1115/1.3013848.

[11] Robin C Purshouse and Peter J Fleming. "Conflict, Harmony, and Independence: Relationships in Evolutionary Multi-criterion Optimisation". In: *Evolutionary Multi-Criterion Optimization*. Ed. by Carlos M Fonseca et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 16–30.

[12] Indraneel Das. "A Preference Ordering Among Various Pareto Optimal Alternatives". In: *Structural Optimization* 18 (1999), pp. 30–35. DOI: 10.1007/BF01210689.

[13] Jarod C Kelly et al. "Incorporating user shape preference in engineering design optimisation". In: *Journal of Engineering Design* 22.9 (2011), pp. 627–650. DOI: 10.1080/09544821003662601.

[14] R. T. Marler and J. S. Arora. "Survey of multi-objective optimization methods for engineering". In: *Structural and Multidisciplinary Optimization* 26.6 (2004), pp. 369–395. DOI: 10.1007/s00158-003-0368-6.

[15] E. M. Kasprzak and K. E. Lewis. "Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method". In: *Structural and Multidisciplinary Optimization* 22.3 (Oct. 2001), pp. 208–218. DOI: 10.1007/s001580100138.

[16] Kevin N. Otto and Erik K. Antonsson. "Trade-off strategies in engineering design". In: *Research in Engineering Design* 3.2 (1991), pp. 87–103. DOI: 10.1007/BF01581342.

[17] S Gunawan and S Azarm. "Multi-objective robust optimization using a sensitivity region concept". In: *Structural and Multidisciplinary Optimization* 29.1 (2005), pp. 50–60. DOI: 10.1007/s00158-004-0450-8.

[18] Christopher A Mattson and Achille Messac. *Pareto Frontier Based Concept Selection Under Uncertainty, with Visualization*. Tech. rep. 2005, pp. 85–115.

[19] Carlos M. Fonseca and Peter J. Fleming. "Multiobjective optimization and multiple constraint handling with evolutionary algorithms - Part I: A unified formulation". In: *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.* (1998). DOI: 10.1109/3468.650319.

[20] Mehmet Unal, Gordon P. Warn, and Timothy W. Simpson. "Quantifying tradeoffs to reduce the dimensionality of complex design optimization problems and expedite trade space exploration". In: *Structural and Multidisciplinary Optimization* 54.2 (2016), pp. 233–248. DOI: 10.1007/s00158-015-1389-7.

[21] Martin Philip Bendsøe and Noboru Kikuchi. "Generating optimal topologies in structural design using a homogenization method". In: *Computer Methods in Applied Mechanics and Engineering* (1988). DOI: 10.1016/0045-7825(88)90086-2.

[22] Mehmet Unal, Gordon P. Warn, and Timothy W. Simpson. "Quantifying the shape of pareto fronts during multi-objective trade space exploration". In: *Journal of Mechanical Design, Transactions of the ASME* 140.2 (2018), pp. 1–13. DOI: 10.1115/1.4038005.

[23] Indraneel Das. "On characterizing the "knee" of the Pareto curve based on Normal-Boundary Intersection". In: *Structural Optimization* 18 (1999), pp. 107–115. DOI: 10.1007/BF01195985.

[24] Bart Frischknecht and Panos Papalambros. "A Pareto Approach To Aligning Public and Private Objectives in Vehicle Design". In: 2008.

[25] Bart D. Frischknecht, Diane L. Peters, and Panos Y. Papalambros. "Pareto set analysis: Local measures of objective coupling in multiobjective design optimization". In: *Structural and Multidisciplinary Optimization* 43.5 (May 2011), pp. 617–630. DOI: 10.1007/s00158-010-0599-2.

[26] Jin Wu and Shapour Azarm. "Metrics for quality assessment of a multiobjective design optimization solution set". In: *Journal of Mechanical Design, Transactions of the ASME* (2001). DOI: 10.1115/1.1329875.

[27] Timothy Ward Athan and Panos Y. Papalambros. "A quasi-Monte Carlo method for multicriteria design optimization". In: *Engineering Optimization* 27.3 (1996), pp. 177–198. DOI: 10.1080/03052159608941405.

[28] P. Papalambros and D. J. Wilde. "Global Non-iterative Design Optimization Using Monotonicity Analysis". In: *Journal of Mechanical Design, Transactions of the ASME* 78 -WA/DE-17 (1978).

[29] Nestor F Michelena and Alice M Agogino. "Multiobjective Hydraulic Cylinder Design". In: *Journal of Mechanisms, Transmissions, and Automation in Design* 110.1 (Mar. 1988), pp. 81–87. DOI: 10.1115/1.3258910.

[30] M. Gobbi et al. "On the analytical derivation of the Pareto-optimal set with applications to structural design". In: *Structural and Multidisciplinary Optimization* 51.3 (2015), pp. 645–657. DOI: 10.1007/s00158-014-1152-5.

[31] Giampiero Mastinu, Massimiliano Gobbi, and Carlo Miano. *Optimal design of complex mechanical systems: With applications to vehicle engineering*. 2006, pp. 1–359. DOI: 10.1007/978-3-540-34355-4.

[32] Pramod Jain and Alice M. Agogino. "Theory of design: An optimization perspective". In: *Mechanism and Machine Theory* 25.3 (1990), pp. 287–303. DOI: 10.1016/0094-114X(90)90030-N.

[33] K. Ishii and B. Parkan. "Active Constraint Deduction - A Framework for Expert Systems in Mechanical Systems Design". In: *Advances in Design Automation - ASME Design Technology Conferences - The Design Automation Conference*. Vol. 10. 1987.

[34] Jonathan Cagan and Alice M. Agogino. *Innovative design of mechanical structures from first principles*. 1987. DOI: 10.1017/S0890060400000275.

[35] Kalyanmoy Deb and Aravind Srinivasan. "Innovization: Innovating design principles through optimization". In: *GECCO 2006 - Genetic and Evolutionary Computation Conference* 2 (2006), pp. 1629–1636.

[36] Panos Y. Papalambros and Douglass J. Wilde. *Principles of Optimal Design*. Cambridge University Press, Jan. 2017. DOI: 10.1017/9781316451038.

[37] D.G. Carmichael. "Computation of Pareto Optima in Structural Design". In: *International Journal for Numerical Methods in Engineering* 15 (1980), pp. 925–952. DOI: 10.1017/S0022029900029393.

[38] J G Lin. "Maximal Vectors and Multi-Objective Optimization". In: *Journal of Optimization Theory and Applications* 18.01 (1976).

[39] George Mavrotas. "Effective implementation of the ε-constraint method in Multi-Objective Mathematical Programming problems". In: *Applied Mathematics and Computation* 213.2 (2009), pp. 455–465. DOI: https://doi.org/10.1016/j.amc.2009.03.037.

[40] Yacov Y. Haimes and Warren A. Hall. "Multiobjectives in water resource systems analysis: The Surrogate Worth Trade Off Method". In: *Water Resources Research* 10.4 (1974), pp. 615–624. DOI: 10.1029/WR010i004p00615.

[41] Panos Y. Papalambros. "Model Reduction and Verification Techniques". In: *Advances in Design Optimization*. Ed. by H Adeli. New York: Chapman and Hall, 1994, pp. 109–138.

[42] Alex Abramson et al. "An ingestible self-orienting system for oral delivery of macromolecules". In: *Science* 363.6427 (2019). DOI: 10.1126/science.aau2277.

[43] U.S. Department of Health and Human Services Food and Drug Administration (CDER). "Guidance for Industry: Size, Shape and Other Physical Attributes of Generic Tablets and Capsules". In: *Pharmaceutical Quality/CMC* December (2013), pp. 1–11.

[44] Mathworks. *Optimization Toolbox ™ Users Guide R2020b, retrieved November 27, 2020*. 2020.

[45] K. S. Channer and J. P. Virjee. "The effect of size and shape of tablets on their esophageal transit". In: *Journal of Clinical Pharmacology* 26.2 (1986), pp. 141–

146. DOI: 10.1002/j.1552-4604.1986.tb02922.x.

[46] S Azarm and P Papalambros. "An Automated Procedure for Local Monotonicity Analysis". In: *Journal of Mechanisms, Transmissions, and Automation in Design* 106.1 (Mar. 1984), pp. 82–89. DOI: 10.1115/1.3258566.

[47] G A Hazelrigg. "On the Role and Use of Mathematical Models in Engineering Design". In: *Journal of Mechanical Design* 121.3 (Sept. 1999), pp. 336–341. DOI: 10.1115/1.2829465.

[48] Rajesh Radhakrishnan and Daniel A McAdams. "A Methodology for Model Selection in Engineering Design". In: *Journal of Mechanical Design* 127.3 (2005), pp. 378–387. DOI: 10.1115/1.1830048.

[49] J Zhou and R W Mayne. "Interactive Computing in the Application of Monotonicity Analysis to Design Optimization". In: *Journal of Mechanisms, Transmissions, and Automation in Design* 105.2 (June 1983), pp. 181–186. DOI: 10.1115/1.3258506.

[50] Susan Finger and John R Dixon. "A review of research in mechanical engineering design. Part II: Representations, analysis, and design for the life cycle". In: *Research in Engineering Design* 1.2 (1989), pp. 121–137. DOI: 10.1007/BF01580205.