ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Bernreuther, Marco; Müller, Georg; Volkwein, Stefan

Article — Published Version

Efficient scalarization in multiobjective optimal control of a nonsmooth PDE

Computational Optimization and Applications

Provided in Cooperation with:
Springer Nature

Suggested Citation: Bernreuther, Marco; Müller, Georg; Volkwein, Stefan (2022) : Efficient scalarization in multiobjective optimal control of a nonsmooth PDE, Computational Optimization and Applications, ISSN 1573-2894, Springer US, New York, NY, Vol. 83, Iss. 2, pp. 435-464, https://doi.org/10.1007/s10589-022-00390-y

This Version is available at:
https://hdl.handle.net/10419/307013

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Computational Optimization and Applications (2022) 83:435–464
https://doi.org/10.1007/s10589-022-00390-y

Efficient scalarization in multiobjective optimal control of a nonsmooth PDE

Marco Bernreuther[1] · Georg Müller[2] · Stefan Volkwein[1]

Received: 9 April 2021 / Accepted: 17 June 2022 / Published online: 28 August 2022
© The Author(s) 2022

Abstract

This work deals with the efficient numerical characterization of Pareto stationary fronts for multiobjective optimal control problems with a moderate number of cost functionals and a mildly nonsmooth, elliptic, semilinear PDE-constraint. When "ample" controls are considered, strong stationarity conditions that can be used to numerically characterize the Pareto stationary fronts are known for our problem. We show that for finite dimensional controls, a sufficient adjoint-based stationarity system remains obtainable. It turns out that these stationarity conditions remain useful when numerically characterizing the fronts, because they correspond to strong stationarity systems for problems obtained by application of weighted-sum and reference point techniques to the multiobjective problem. We compare the performance of both scalarization techniques using quantifiable measures for the approximation quality. The subproblems of either method are solved with a line-search globalized pseudo-semismooth Newton method that appears to remove the degenerate behavior of the local version of the method employed previously. We apply a matrix-free, iterative approach to deal with the memory and complexity requirements when solving the subproblems of the reference point method and compare several preconditioning approaches.

Keywords Multiobjective optimal control · Nonsmooth optimization · Stationarity conditions · Pareto optimality · Scalarization methods · Pseudo-semismooth Newton method

✉ Georg Müller
georg.mueller@uni-heidelberg.de

Marco Bernreuther
marco.bernreuther@uni-konstanz.de

Stefan Volkwein
stefan.volkwein@uni-konstanz.de

[1] University of Konstanz Department of Mathematics and Statistics, D-78457 Konstanz, Germany

[2] University of Heidelberg Interdisciplinary center for scientific computing (IWR), D-69120 Heidelberg, Germany

1 Problem formulation

The prevailing notion of optimality in multiobjective optimization is that of optimal compromises or Pareto optimal points, see Definition 1. Evaluating the quality of Pareto optimal points typically involves interpreting the set of function values of all Pareto optimal points – the Pareto front. The aim of this paper is the efficient numerical characterization of the Pareto fronts of nonsmooth multiobjective optimal control problems with few objectives and an elliptic PDE-constraint with max-type nonsmoothness using generalized stationarity conditions, and the comparison of the numerical performance of a weighted-sum approach and a reference point method in terms of efficiency and discretization quality. For ease of presentation, we restrict this exposition to the case of two objectives only – though the scope of this paper can readily be extended to moderately many objective functions using hierarchical approaches, see [3, 4, 15]. Hence, we consider bicriterial problems of the form

$$\min_{(y,u)} \mathcal{J}(y,u) = \begin{pmatrix} \mathcal{J}_1(y,u) \\ \mathcal{J}_2(y,u) \end{pmatrix} = \begin{pmatrix} j_1(y) + \frac{\sigma_1}{2}\|u\|_U^2 \\ j_2(y) + \frac{\sigma_2}{2}\|u\|_U^2 \end{pmatrix} \tag{P}$$
$$\text{s.t. } (y,u) \in V \times U \text{ satisfies } -\Delta y + \kappa \max\{0, y\} = \mathcal{B}(u) \text{ in } V'.$$

In (**P**), the symbols y and u denote the state and control variables in the corresponding state space V and (possibly finite dimensional) control space U, respectively, the j_i denote suitably well behaved scalar cost functionals, the σ_i are nonnegative regularization parameters and $\mathcal{B} : U \to L^2(\Omega)$ denotes a control-to-right-hand-side mapping. For the detailed assumptions on the problem, we refer to Assumption 1.

Multicriterial optimization problems with nonsmooth PDE-constraints like (**P**) arise in various physical applications with conflicting objectives, see, e.g., [13, 16, 19, 21]. The combination of generally only (Hadamard) directionally differentiable Nemytski operators in the constraint and the inherently nonsmooth structure of multiobjective optimization makes sensitivity and stationarity analysis as well as the numerical solution of these problems rather delicate. Their treatment typically requires specialized stationarity concepts and approaches that do not follow standard procedure for Gâteaux differentiable problems. The particular case of the PDE-constraint in (**P**) is rather well understood in terms of existence and regularity of solutions and differentiability properties of the solution operator and has previously been addressed as a constraint in optimization problems in, e.g., [6, 7]. The specific structure of the PDE even allows for the derivation of strong stationarity systems when the control space is sufficiently rich, which has been considered in both scalar and multiobjective optimization with an arbitrary number of objectives, see [6, 7]. Stationarity conditions of intermediate strength based on the characterization of the subdifferentials of the solution operator to the constraining PDE have been addressed in [6] and the considerations in [7] show that C-stationarity and strong stationarity in fact coincide for ample controls.

Numerically, the Pareto stationary front of problem (**P**) has been characterized for two and three cost functionals and $L^2(\Omega)$ controls [7] by combining a

first-optimize-then-discretize approach and a pseudo-semismooth Newton (PSN) method, and by using a regularization approach. With multiple objectives, computation times can quickly become large with increased fineness of discretizations of the domain and the Pareto front, making efficient numerics a vital component to these simulations. In [5], the authors discussed a standard offline/online greedy *reduced-basis* approach for (**P**) with a single scalar objective function and both low and high dimensional control/parameter space and compared the results to an adaptive way of generating the reduced basis along the solution process of the PSN.

In this paper, our goal is the efficient characterization of the Pareto stationary front of the bicriterial problem (**P**) for both $L^2(\Omega)$-controls as well as finite-dimensional controls with specific structure to allow for a combination with the reduced-basis approaches from [5] mentioned above. In the case of the $L^2(\Omega)$-controls, where strong stationarity conditions were derived in [7], this is essentially a refinement of the numerical techniques employed in the same paper. For finite dimensional controls, which are typically insufficiently "rich" to obtain strong stationarity conditions, however, this requires revisiting the optimality conditions and how they can be used to characterize the Pareto stationary fronts. Following the approaches in [7], we show that, in the case of finite dimensional controls, a sufficient stationarity system can still be obtained. Though a sufficient system for a necessary optimality condition is unfavorable in general, it turns out that the system we obtain is essentially a strong stationarity system for the optimization problems corresponding to the weighted-sum method and the Euclidean reference point method and therefore a somewhat reasonable system to use for characterization of the Pareto stationary fronts. The nonlinear structure in the sets that are involved there, introduced by the nonsmoothness of the PDE, however, does not lend itself to use for characterization of the front directly. Instead, a linearization of the same system will be solved by a line-search stabilized PSN approach and combined with a sign condition that is checked a-posteriori. The line-search essentially eliminates the rare, mesh-dependent non-convergence issue observed for the undamped PSN in [5, 7]. We will compare the efficiency and the approximation quality of the front of the weighted-sum and the reference point approach using quantifiable quality measures.

The structure of this paper is as follows: We will shortly comment on the assumptions for this paper in Subsect. 1.1. Then, we recall the required notions of Pareto optimality and Pareto stationarity and state the respective first order systems for (non-)ample controls in Sect. 2. In the case $U = L^2(\Omega)$, the analytical results are analogous to those in [7] and the strong stationarity results are simply restated, whereas similar conditions in the case $U = \mathbb{R}^p$ are shown to remain sufficient only. In Sect. 3, we will shortly recall the weighted-sum method and the reference point method, and show that the sufficient system from the multiobjective setting is equivalent to strong stationarity systems for the scalarized systems. We explain how these methods can therefore be used to characterize the Pareto stationary front. The numerical implementation is explained in detail in Sect. 4, where we present a matrix-free preconditioned limited-memory generalized minimal residual (L-GMRES) method for the line-search globalized pseudo-semismooth Newton (gPSN) method that will be used to handle the density of the discretization matrices in the reference point method and the convergence issues arising from the

nonsmoothness of the PDE-constraint, respectively. We further present two numerical examples – one with (FE-discretized) $L^2(\Omega)$-controls and one with inherently finite dimensional controls in Sect. 5. The interpretation of the numerical results is specifically focused on the effects of different preconditioning strategies for the reference point method. We introduce two quantities to measure the approximation quality of the two methods and use them as a basis for a performance comparison of the two scalarization approaches.

1.1 Notation and assumptions on the data

We endow $V = H_0^1(\Omega)$ with the inner product $\langle \varphi, \phi \rangle_V = \int_\Omega \nabla\varphi \cdot \nabla\phi + \varphi\phi \, \mathrm{d}\boldsymbol{x}$ for $\varphi, \phi \in V$ and the induced norm $\| \cdot \|_V = \langle \cdot, \cdot \rangle_V^{1/2}$. Its topological dual space is written as $V' = H^{-1}(\Omega)$. The space Y denotes $V \cap H^2(\Omega)$ with topological dual space Y'. We also set $H = L^2(\Omega)$. For functions in Y, the Laplacian is understood in the non-variational sense and the Dirichlet Laplacian $\Delta : H \to Y'$ is understood in the very weak sense (see [12], Section 1.9). Our assumptions on the data are as follows:

Assumption 1

1) $\Omega \subset \mathbb{R}^d$ for $d \in \mathbb{N} \setminus \{0\}$ is a bounded domain that is convex or possesses a $C^{1,1}$ -boundary (cf. [10, Section 6.2]),
2) $j_1, j_2 : Y \to \mathbb{R}$ are weakly lower semicontinuous, twice continuously Fréchet-differentiable and bounded from below,
3) $\sigma_1 \geq 0, \sigma_2 > 0, \kappa \geq 0$,
4) $U = \mathbb{R}^p$ for $p \in \mathbb{N} \setminus \{0\}$ and $\| \cdot \|_U$ denotes the Euclidean norm or $U = H$ and $\| \cdot \|_U$ denotes the L^2-norm,
5) $\mathcal{B} : U \to H$ possesses the following property:

5a) If $U = \mathbb{R}^p$ the operator $\mathcal{B} : U \to H$ is linear (and therefore automatically bounded) and the pairwise intersection of the sets $\{b_i \neq 0\} \subset \Omega$ of $b_i := \mathcal{B}(e_i) \in H$, where e_i, $i = 1, \ldots, p$ denote the unit vectors in U, are Lebesgue nullsets and none of the b_i are zero.
5b) If $U = H$ the operator $\mathcal{B} : U \to H$ is unitary.

2 A sufficient condition for pareto stationarity

Structurally, this section follows [7] closely, where the case $U = H = L^2(\Omega)$ is dealt with and carries over immediately, yielding strong stationarity conditions for the Pareto stationary points. For the case where $U = \mathbb{R}^p$, however, the same reasoning does not hold up – specifically, the system obtained by similar arguments will only be sufficient in general. The main issue in the analysis is the well-known fact that strong stationarity conditions typically require ample controls. Note that the results presented in this section can readily be generalized to an arbitrary finite number of objective functionals.

We start by summarizing the main properties of the solution operator $\mathcal{S}$ to the PDE-constraint in the next lemma as a minor extension to [7, Lemma 4.2]. Note that, since we will mostly focus on the case $U = \mathbb{R}^p$, we do not obtain analogous results to all the results stated in [7, Lemma 4.2] for the case $U = H$.

Lemma 1 (Properties of the solution operator $\mathcal{S}$) *Let $u \in U$ be a control with associated state $y = \mathcal{S}(u)$. Then*:

1) *There is a solution operator $\mathcal{S} : U \to Y$ that is Lipschitz continuous and Hadamard directionally differentiable, where the derivative $\mathcal{S}'(u;h) = w \in Y$ for given direction $h \in U$ is the unique solution to*

$$-\Delta w + \kappa \mathbb{1}_{\{y=0\}} \max\{0, w\} + \kappa \mathbb{1}_{\{y>0\}} w = \mathcal{B}(h) \quad \text{in } V'. \tag{1}$$

This especially implies the Y-regularity of the state variable y.

2) *If $U = \mathbb{R}^p$, then the map*

$$\tilde{\mathcal{B}^\dagger} : H \to U, \quad v \mapsto \nu = (\nu_i)_{1 \le i \le p} \quad \text{with } \nu_i = \frac{\langle v, b_i \rangle_H}{\|b_i\|_H^2} \tag{2}$$

is a linear and bounded left inverse of $\mathcal{B} : U \to H$.

3) *The map $\mathcal{S}'(u;\cdot) : U \to Y$ is Lipschitz continuous and allows for a Lipschitz continuous left inverse given by*

$$\mathcal{S}'(u;\cdot)^\dagger : Y \to U, \quad w \mapsto \mathcal{B}^\dagger\Big(\underbrace{-\Delta w + \kappa \mathbb{1}_{\{y=0\}} \max\{0, w\} + \kappa \mathbb{1}_{\{y>0\}} w}_{\in H} \Big),$$

where $\mathcal{B}^\dagger$ is any linear, bounded left inverse of $\mathcal{B}$.

4) *There exists a linear and bounded left inverse $\mathcal{B}^\dagger$ of $\mathcal{B}$ that does not depend on u such that*

$$\left\langle u, \mathcal{S}'(u;\cdot)^\dagger(w) \right\rangle_U = \left\langle (-\Delta + \kappa \mathbb{1}_{\{y>0\}}) \mathcal{B}^{\dagger *}(u), w \right\rangle_{Y',Y}$$

for every $w \in Y$, where $\mathcal{B}^{\dagger *} : U \to H$ is the Hilbert adjoint of the left inverse $\mathcal{B}^\dagger$.

Proof Let $u \in U$ be arbitrarily given and $y = \mathcal{S}(u) \in Y$. Notice that the Y-regularity also follows from Proposition 2.1 in [6].

1) The linearity and boundedness of $\mathcal{B}$ imply Lipschitz continuity and Hadamard differentiability analogously to Proposition 2.1 and Theorem 2.2 in [6] and due to the chain rule, the directional derivative of the solution operator $w = \mathcal{S}'(u;h)$ solves

$$-\Delta w + \kappa \mathbb{1}_{\{y=0\}} \max\{0, w\} + \kappa \mathbb{1}_{\{y>0\}} w = \mathcal{B}(h) \quad \text{in } V'.$$

2) Note that the operator $\tilde{\mathcal{B}^\dagger}$ is well defined due to Assumption 1-5). Clearly, the operator is linear and bounded. The left inverse quality remains to be proved. Let $\tilde{u} \in U = \mathbb{R}^p$ and $v = \mathcal{B}\tilde{u} = \sum_{j=1}^p b_j \tilde{u}_j \in H$. For every $i \in \{1, \dots, p\}$, we have that

$$\left(\tilde{\mathcal{B}^\dagger}(v)\right)_i = \frac{\langle v, b_i\rangle_H}{\|b_i\|_H^2} = \sum_{j=1}^p \tilde{u}_j \frac{\langle b_j, b_i\rangle_H}{\|b_i\|_H^2} = \tilde{u}_i \frac{\langle b_i, b_i\rangle_H}{\|b_i\|_H^2} = \tilde{u}_i,$$

where the second to last equality holds due to the H-orthogonality of the b_i's induced by Assumption 1-5).

3) The Lipschitz continuity of the linearized solution operator is implied by the form of the linearization (1). Existence of a left inverse is clear due to part *1* if $U = \mathbb{R}^p$ and since $\mathcal{B}$ is unitary if $U = H$. Thus existence of a left inverse of $\mathcal{S}'(u;\cdot)$ and its Lipschitz continuity are obvious from the explicit definition.

4) First assume that $U = H$ and that $\mathcal{B}$ is a unitary operator. Then, by definition, we obtain that $\mathcal{B}^{\dagger *} = \mathcal{B}$. Consequently, for $u \in U$ and $y = \mathcal{S}(u)$,

$$\begin{aligned}\langle \mathcal{B}^{\dagger *}(u), \kappa \mathbb{1}_{\{y=0\}} \max\{0, w\}\rangle_H &= \langle \mathcal{B}(u), \kappa \mathbb{1}_{\{y=0\}} \max\{0, w\}\rangle_H \\ &= \langle \underbrace{-\Delta y + \kappa \max\{0, y\}}_{=0 \text{ a.e. on } \{y=0\}}, \kappa \mathbb{1}_{\{y=0\}} \max\{0, w\}\rangle_H = 0, \end{aligned} \tag{3}$$

where $\Delta y = 0$ a.e. on $\{y = 0\}$ is a consequence of [7, Lemma 4.1]. Thus, using part 3) and (3), we find that

$$\begin{aligned}\left\langle u, \mathcal{S}'(u;\cdot)^\dagger(w)\right\rangle_U &= \left\langle u, \mathcal{B}^\dagger\big(-\Delta w + \kappa \mathbb{1}_{\{y=0\}} \max\{0, w\} + \kappa \mathbb{1}_{\{y>0\}} w\big)\right\rangle_U \\ &= \left\langle \mathcal{B}^{\dagger *}(u), \big(-\Delta + \kappa \mathbb{1}_{\{y>0\}}\big) w\right\rangle_H + \left\langle \mathcal{B}^{\dagger *}(u), \kappa \mathbb{1}_{\{y=0\}} \max\{0, w\}\right\rangle_H \\ &= \left\langle \mathcal{B}^{\dagger *}(u), \big(-\Delta + \kappa \mathbb{1}_{\{y>0\}}\big) w\right\rangle_H \\ &= \left\langle \big(-\Delta + \kappa \mathbb{1}_{\{y>0\}}\big) \mathcal{B}^{\dagger *}(u), w\right\rangle_{Y',Y}\end{aligned}$$

for every $w \in Y$, where the last line follows due to the definition of the very weak Dirichlet Laplacian.Now assume that $U = \mathbb{R}^p$. We show that $\tilde{\mathcal{B}^\dagger}$ is the desired left inverse. For $u \in U$, $y = \mathcal{S}(u)$ and any $i \in \{1, \dots, p\}$, we have that $\{y = 0\} \cap \{b_i u_i \neq 0\}$ is a nullset, because

$$\{b_i u_i \neq 0\} \cap \{y = 0\} \subset \{b_i u_i \neq 0\} \cap \{\mathcal{B}(u) = 0\} \subset \{b_i u_i \neq 0\} \cap \{b_i u_i = 0\},$$

where the first inclusion is again a consequence of [7, Lemma 4.1] and the second inclusion is due to Assumption 1-5). Thus, for any $i \in \{1, \dots, p\}$, we infer

$$\frac{\langle \mathbb{1}_{\{y=0\}} \max\{0, w\}, b_i u_i\rangle_H}{\|b_i\|_H^2} = 0.$$

Due to part 2), we obtain that

$$\left\langle u, \tilde{\mathcal{B}}^{\dagger}\left(\kappa \mathbb{1}_{\{y=0\}} \max\{0, w\}\right)\right\rangle_U = \sum_{i=1}^{p} u_i \frac{\langle \kappa \mathbb{1}_{\{y=0\}} \max\{0, w\}, b_i\rangle_H}{\|b_i\|_H^2} = \kappa \sum_{i=1}^{p} \frac{\langle \mathbb{1}_{\{y=0\}} \max\{0, w\}, u_i b_i\rangle_H}{\|b_i\|_H^2} = 0$$

for every $w \in Y$. Consequently,

$$\begin{aligned}\langle u, \mathcal{S}'(u;\cdot)^{\dagger}(w)\rangle_U &= \left\langle u, \tilde{\mathcal{B}}^{\dagger}\left(-\Delta w + \kappa \mathbb{1}_{\{y=0\}} \max\{0, w\} + \kappa \mathbb{1}_{\{y>0\}} w\right)\right\rangle_U \\ &= \left\langle \left(-\Delta + \kappa \mathbb{1}_{\{y>0\}}\right) \tilde{\mathcal{B}}^{\dagger *}(u), w\right\rangle_{Y',Y},\end{aligned}$$

for every $w \in Y$.

□

As usual, we denote the well-defined reduced cost functional as $\hat{\mathcal{J}} : U \to \mathbb{R}^2, \hat{\mathcal{J}}(u) = \mathcal{J}(\mathcal{S}(u), u)$. Having established the properties of the solution operator, we are ready to review the different notions of Pareto optimality and the optimality conditions that will play a role later on.

Definition 1 (Pareto Optimality) Let $\bar{y}, \bar{u}$ with $\bar{y} = \mathcal{S}(\bar{u})$ and $\bar{u} \in U$. The control $\bar{u}$ is called:

1) a *local weak Pareto optimal point* of (**P**) if an $r > 0$ exists such that there is no $u \in U$ satisfying

$$\|u - \bar{u}\|_U < r, \quad \mathcal{J}_i(\mathcal{S}(u), u) < \mathcal{J}_i(\bar{y}, \bar{u}) \text{ for } i = 1, 2;$$

2) a *local Pareto optimal point* of (**P**) if an $r > 0$ exists such that there is no $u \in U$ satisfying

$$\|u - \bar{u}\|_U < r, \quad \mathcal{J}_i(\mathcal{S}(u), u) \leq \mathcal{J}_i(\bar{y}, \bar{u}) \text{ for } i = 1, 2,$$

where the latter inequality is strict for at least one i;

3) a *local proper Pareto optimal point* of (**P**) if there are $r, C > 0$ such that for every $u \in U$ satisfying $\|u - \bar{u}\|_U < r$ and $\mathcal{J}_i(\mathcal{S}(u), u) \leq \mathcal{J}_i(\bar{y}, \bar{u})$ for some index $i \in \{1, 2\}$, there exists an index $m \in \{1, 2\} \setminus \{i\}$ with

$$\mathcal{J}_i(\bar{y}, \bar{u}) - \mathcal{J}_i(\mathcal{S}(u), u) \leq C(\mathcal{J}_m(\mathcal{S}(u), u) - \mathcal{J}_m(\bar{y}, \bar{u}));$$

4) a *global (weak/proper) Pareto optimal point* of (**P**) if the previous conditions hold with $r = \infty$;

The image sets of all controls that are (local/global) (weak/proper) Pareto optimal under the cost functional are called the Pareto fronts corresponding to the respective sense of optimality.

Analogously to [7], we obtain the following corresponding primal optimality conditions.

Theorem 1 (Optimality Conditions – Primal Form)

1) *If $\bar{u} \in U$ with associated state $\bar{y} = \mathcal{S}(\bar{u})$ is a local weak Pareto optimal point of* (**P**), *then there exists no direction $h \in U$ satisfying*

$$\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_i \langle \bar{u}, h\rangle_U < 0 \quad \text{for } i = 1,2. \tag{4}$$

2) *If $\bar{u} \in U$ with associated state $\bar{y} = \mathcal{S}(\bar{u})$ is a local proper Pareto optimal point of* (**P**) *with constants $r, C > 0$, then for every $h \in U$ with $\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_i\langle \bar{u}, h\rangle_U < 0$ for some $i \in \{1,2\}$, there exists an $m \in \{1,2\} \setminus \{i\}$ with*

$$\begin{aligned} &- \big(\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_i \langle \bar{u}, h\rangle_U\big) \\ &\quad \le C\big(\langle j_m'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_m \langle \bar{u}, h\rangle_U\big). \end{aligned}$$

Proof See [7, Theorem 3.1]. □

Accordingly, we obtain the corresponding notions of Pareto stationarity.

Definition 2 (Pareto Stationarity) Let $\bar{u} \in U$ and $\bar{y} = \mathcal{S}(\bar{u})$. The control $\bar{u}$ is called:

1) a *weak Pareto stationary point* of (**P**) if there is no $h \in U$ satisfying

$$\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_i \langle \bar{u}, h\rangle_U < 0 \quad \text{for } i = 1,2;$$

2) a *Pareto stationary point* of (**P**) if there is no $h \in U$ satisfying

$$\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_i \langle \bar{u}, h\rangle_U \le 0 \quad \text{for } i = 1,2,$$

where the latter inequality is strict for at least one i;

3) a *proper Pareto stationary point* of (**P**) if there is a $C > 0$ such that for all $h \in U$ with $\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_i\langle \bar{u}, h\rangle_U < 0$ for some $i \in \{1,2\}$, there exists an $m \in \{1,2\} \setminus \{i\}$ with

$$\begin{aligned} &- \big(\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_i \langle \bar{u}, h\rangle_U\big) \\ &\quad \le C\big(\langle j_m'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_m \langle \bar{u}, h\rangle_U\big). \end{aligned}$$

Remark 1 By definition, all proper Pareto optima are Pareto optima, which in turn all are weak Pareto optima. The same holds locally. As a consequence of Theorem 1, all local weak Pareto optima are weakly Pareto stationary and all local proper Pareto optima are properly Pareto stationary. However, Pareto stationarity is generally not necessary for local Pareto optimality. ◊

In the case of $L^2(\Omega)$-controls, the version of Tucker's/Motzkin's theorem of the alternative in infinite dimensions in [7, Lemma 4.4] provides the existence of the multipliers appearing in the system of strong stationarity conditions. In the case of finite dimensional controls, we unfortunately have to deal with generally nonlinear

subsets of the spaces involved, and, as it turns out, an extension of the existence results to arbitrary subsets does not hold. Hence, we can only recover one of the implications when the result is generalized to potentially nonlinear subsets of Hilbert spaces.

Lemma 2 *Suppose $\mathcal{W}$ is a nonempty subset of a real Hilbert space $\mathcal{V}$ and $v_1', \dots, v_N' \in \mathcal{V}'$ are given. Assume that there exists $\lambda \in \mathbb{R}^N$ with $\lambda_i \geq 0$ for $i = 1, \dots, N$ such that*

$$\sum_{i=1}^{N} \lambda_i = 1 \quad \text{and} \quad \sum_{i=1}^{N} \lambda_i \langle v_i', w \rangle_{\mathcal{V}} \geq 0 \quad \text{for all } w \in \mathcal{W}.$$

Then there exists no $z \in \mathcal{W}$ such that

$$\langle v_i', z \rangle_{\mathcal{V}} < 0 \quad \text{for all } i = 1, \dots, N.$$

Furthermore, if $\lambda_i > 0$ for all $i = 1, \dots, N$, then there exists no $z \in \mathcal{W}$ such that

$$\langle v_i', z \rangle_{\mathcal{V}} \leq 0 \quad \text{for all } i = 1, \dots, N,$$

with the inequality holding strictly for at least one i.

Proof Assume that there exists a $z \in \mathcal{W}$ with $\langle v_i', z \rangle_{\mathcal{V}} < 0$ for all $i = 1, \dots, N$. This would imply that

$$\sum_{i=1}^{N} \lambda_i \langle v_i', z \rangle_{\mathcal{V}} < 0,$$

which is a contradiction. Analogously, if $\lambda_i > 0$ for all $i = 1, \dots, N$, then there exists no $z \in \mathcal{W}$ such that $\langle v_i', z \rangle_{\mathcal{V}} \leq 0$ for all $i = 1, \dots, N$ with the inequality holding strictly for at least one i, which shows the claim. □

It follows that the technique used to show [7, Theorem 4.5] now yields a sufficient system only, i.e., we obtain the following sufficient adjoint-based optimality system.

Theorem 2 (Sufficient adjoint-based system)

1. *Assume that there exists an adjoint state $\bar{p}$ and a multiplier $\bar{\alpha}$ such that $\bar{u}, \bar{y}, \bar{p}, \bar{\alpha}$ satisfy the coupled system*

$$\bar{u} \in U, \quad \bar{y} \in Y, \quad \bar{p} \in H, \quad \bar{\alpha} \in \mathbb{R}^2, \tag{5a}$$

$$\bar{\alpha}_i \geq 0 \quad \text{for } i = 1, 2, \quad \sum_{i=1}^{2} \bar{\alpha}_i = 1, \tag{5b}$$

$$-\Delta\bar{y} + \kappa \max\{0, \bar{y}\} = \mathcal{B}(\bar{u}) \quad \text{in } V',$$
$$\left\langle -\Delta\bar{p} + \kappa \mathbb{1}_{\{\bar{y}>0\}}\bar{p}, w \right\rangle_{Y',Y} \leq \sum_{i=1}^{2} \bar{\alpha}_i \langle j_i'(\bar{y}), w\rangle_{Y',Y} \text{ for all } w \in \mathrm{Im}(\mathcal{S}'(\bar{u};\cdot)), \tag{5c}$$

$$\bar{p} + \sum_{i=1}^{2} \bar{\alpha}_i \sigma_i \mathcal{B}^{\dagger *}(\bar{u}) = 0 \quad \text{in } H. \tag{5d}$$

Then $\bar{u} \in U$ is a weak Pareto stationary point of (**P**).

2. *Assume that $\bar{u}, \bar{y}, \bar{p}, \bar{\alpha}$ satisfy the system* (5), *where the inequality in* (5b) *is strict, i.e. $\bar{\alpha}_i > 0$ for $i = 1, 2$. Then $\bar{u}$ is a proper Pareto stationary point of* (**P**) (*and thus also a Pareto stationary point*).

Proof Let $\bar{u}, \bar{y}, \bar{p}, \bar{\alpha}$ that solve (5) be given. Inserting (5d) into (5c), yields

$$\sum_{i=1}^{2} \bar{\alpha}_i \langle j_i'(\bar{y}) + \sigma_i(-\Delta + \kappa \mathbb{1}_{\{\bar{y}>0\}})\mathcal{B}^{\dagger *}(\bar{u}), w\rangle_{Y',Y} \geq 0 \quad \text{for all } w \in \mathrm{Im}(\mathcal{S}'(\bar{u};\cdot)).$$

According to Lemma 2, this implies that

$$\langle j_i'(\bar{y}), w\rangle_{Y',Y} + \sigma_i \langle(-\Delta + \kappa \mathbb{1}_{\{\bar{y}>0\}})\mathcal{B}^{\dagger *}(\bar{u}), w\rangle_{Y',Y} < 0 \quad \text{for } i = 1, 2$$

is valid for no $w \in \mathrm{Im}(\mathcal{S}'(\bar{u};\cdot))$. Since $\mathcal{S}'(\bar{u};\mathcal{S}'(\bar{u};\cdot)^{\dagger}(w)) = w$ for all $w \in \mathrm{Im}(\mathcal{S}'(\bar{u};\cdot))$ and due to the explicit form of $\mathcal{S}'(\bar{u};\cdot)^{\dagger}$ provided by Lemma 1-*1*, we have that

$$\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};\mathcal{S}'(\bar{u};\cdot)^{\dagger}(w))\rangle_{Y',Y} + \sigma_i \langle \bar{u}, \mathcal{S}'(\bar{u};\cdot)^{\dagger}(w)\rangle_U < 0 \quad \text{for } i = 1, 2$$

is valid for no $w \in \mathrm{Im}(\mathcal{S}'(\bar{u};\cdot)) \subset Y$. Since $\mathrm{Im}(\mathcal{S}'(\bar{u};\cdot))$ under the map $\mathcal{S}'(\bar{u};\cdot)^{\dagger}$ is U, this is equivalent to weak Pareto stationarity. For part *2*, ordinary Pareto stationarity immediately follows from Lemma 2 and we will show that this implies proper Pareto stationarity analogously to [7, Theorem 4.5 iii)]. To that end, assume that there is $h \in U$ such that $\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_i\langle \bar{u}, h\rangle_U < 0$ for some $i \in \{1, 2\}$. Just like in the proof of part 1, we obtain that

$$\sum_{i=1}^{2} \bar{\alpha}_i \langle j_i'(\bar{y}) + \sigma_i(-\Delta + \kappa \mathbb{1}_{\{\bar{y}>0\}})\mathcal{B}^{\dagger *}(\bar{u}), w\rangle_{Y',Y} \geq 0 \quad \text{for all } w \in \mathrm{Im}(\mathcal{S}'(\bar{u};\cdot)).$$

and applying the form of $\mathcal{S}'(\bar{u};\cdot)^{\dagger}$ provided by Lemma 1-4) again, this implies that

$$\sum_{i=1}^{2} \bar{\alpha}_i \left(\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};\mathcal{S}'(\bar{u};\cdot)^{\dagger}(w))\rangle_{Y',Y} + \sigma_i \langle \bar{u}, \mathcal{S}'(\bar{u};\cdot)^{\dagger} w\rangle_{Y',Y}\right) \geq 0$$

for all $w \in \mathrm{Im}(\mathcal{S}'(\bar{u};\cdot))$ and therefore

$$\sum_{i=1}^{2} \bar{\alpha}_i \left(\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};v) \rangle_{Y',Y} + \sigma_i \langle \bar{u}, v \rangle_{Y',Y} \right) \geq 0$$

for all $v \in U$. Hence

$$\begin{aligned} 0 &< -\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h) \rangle_{Y',Y} + \sigma_i \langle \bar{u}, h \rangle_U \\ &\leq \frac{1}{\min\limits_{l=1,2} \alpha_l} \langle j_k'(\bar{y}), \mathcal{S}'(\bar{u};h) \rangle_{Y',Y} + \sigma_k \langle \bar{u}, h \rangle_U \end{aligned}$$

for $k \in \{1,2\} \setminus \{i\}$. □

Additionally, we make the following observations.

Corollary 1 *Consider the setting of Theorem 2 with $U = \mathbb{R}^p$. If (5d) is replaced by*

$$\mathcal{B}^*(\bar{p}) + \sum_{i=1}^{2} \bar{\alpha}_i \sigma_i \bar{u} = 0 \quad \text{in } U. \tag{6}$$

in Theorem 2-1) or -2) and the sign condition

$$\langle \mathbb{1}_{\{\bar{y}=0\}} \max\{0, w\}, \bar{p} \rangle_H \leq 0 \quad \text{for all } w \in \mathrm{Im}(\mathcal{S}'(\bar{u};\cdot)), \tag{7}$$

is added, then the resulting system is sufficient (but generally not necessary) for weak/proper Pareto stationarity.

Proof Assume that $\bar{u}, \bar{y}, \bar{p}, \bar{\alpha}$ satisfy the system (5a)-(5c), (6) and (7) with $\bar{\alpha}_i \geq 0$ for $i = 1, 2$. For arbitrary $h \in U$ set $w = \mathcal{S}'(\bar{u};h)$. It follows that

$$\begin{aligned} -\sum_{i=1}^{2} \bar{\alpha}_i \sigma_i \langle \bar{u}, h \rangle_U &= \langle \mathcal{B}^*(\bar{p}), h \rangle_U = \langle \bar{p}, \mathcal{B}(h) \rangle_H \\ &= \langle -\Delta w + \kappa \mathbb{1}_{\{\bar{y}>0\}} w + \kappa \mathbb{1}_{\{\bar{y}=0\}} \max\{0, w\}, \bar{p} \rangle_H \\ &\leq \langle -\Delta w + \kappa \mathbb{1}_{\{\bar{y}>0\}} w, \bar{p} \rangle_H = \langle -\Delta \bar{p} + \kappa \mathbb{1}_{\{\bar{y}>0\}} \bar{p}, w \rangle_{Y',Y} \\ &\leq \sum_{i=1}^{2} \bar{\alpha}_i \langle j_i'(\bar{y}), w \rangle_{Y',Y}. \end{aligned}$$

Thus since $\bar{\alpha}_i \geq 0$ and $\sum\limits_{i=1}^{2} \bar{\alpha}_i = 1$ the inequality

$$\langle j_i'(\bar{y}), w \rangle_{Y',Y} + \sigma_i \langle \bar{u}, h \rangle_U < 0$$

cannot be true for all $i = 1, 2$. This implies the desired weak Pareto stationarity. (Proper) Pareto stationarity can be shown analogously. □

3 Connection to scalarization methods

As shown in Sect. 2, we can obtain an adjoint based system that is sufficient for Pareto stationarity. When we base numerical characterizations of the Pareto stationary front on this system, since it is potentially not necessary for Pareto stationarity, we may lose out on points on the front. In this section, we want to show that the adjoint system from Theorem 2 is a strong stationarity system for problems arising in two well-known scalarization methods – i.e., using the adjoint multiobjective stationarity system is not a worse approach than straight forward scalarization. We will briefly explain the scalarization methods – the weighted-sum method (cf., e.g., [8]) and the reference point method (cf., e.g., [14, 17]) – and how we intend to use them to characterize the Pareto stationary front.

3.1 Weighted-sum method (WSM)

For weights $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = 1$, the optimization problem

$$\begin{aligned} &\min_{(y,u)} \alpha_1 \mathcal{J}_1(y,u) + \alpha_2 \mathcal{J}_2(y,u) \\ &\text{s.t. } (y,u) \in V \times U \text{ satisfies } -\Delta y + \kappa \max\{0, y\} = \mathcal{B}(u) \text{ in } V', \end{aligned} \tag{$\mathbf{P}_\alpha$}$$

is called the *weighted-sum problem* (with non-negative weights α_1, α_2) corresponding to ($\mathbf{P}$). The weighted-sum method is based on solving ($\mathbf{P}_\alpha$) for varying α. The primal optimality conditions for the WSM are given in the following theorem.

Theorem 3 *Let $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = 1$ and denote $\alpha = (\alpha_1, \alpha_2)$. Let the control $\bar{u} \in U$ be locally optimal for ($\mathbf{P}_\alpha$) and let $\bar{y} = \mathcal{S}(\bar{u}) \in Y$ be the associated state. Then*

$$\sum_{i=1}^{2} \alpha_i \left(\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h) \rangle_{Y',Y} + \sigma_i \langle u, h \rangle_U \right) \geq 0 \quad \text{for all } h \in U. \tag{8}$$

Proof The claim follows analogously to [7, Theorem 3.1]. □

A control $\bar{u} \in U$ with associated $\bar{y} = \mathcal{S}(\bar{u})$ is called a *stationary point of* ($\mathbf{P}_\alpha$) if (8) is satisfied.

Corollary 2 *Let $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = 1$ and denote $\alpha = (\alpha_1, \alpha_2)$. Then the following statements are equivalent*:

1. *A control $\bar{u} \in U$ with associated state $\bar{y} = \mathcal{S}(\bar{u}) \in Y$ is a stationary point of* ($\mathbf{P}_\alpha$).
2. *There exists $\bar{p}$ such that $\bar{u}, \bar{y}, \bar{p}$ satisfy the system* (5) *with $\bar{\alpha} = \alpha$.*

Proof The corollary follows analogously to the proof of Theorem 2. However, since the problem is inherently scalar, the reverse implication of Lemma 2 is obtained for free and we obtain equivalence. □

Corollary 2 especially implies that the adjoint system (5) is a strong stationarity system for the weighted-sum problem ($\mathbf{P}_\alpha$), i.e., that it is equivalent to the primal necessary stationarity conditions (8) of the weighted-sum problem. Accordingly, it is reasonable to characterize the stationary points of the weighted-sum problems using the system (5). However, system (5) may have multiple solutions, and (in the case of the finite dimensional controls) we cannot solve it numerically because of the variational inequality (5c) on the possibly unknown and nonlinear set $\mathrm{Im}(\mathcal{S}'(\bar{u};\cdot))$. In practice, we therefore modify (5c)-(5d) and instead consider the system

$$\bar{u} \in U, \quad \bar{y} \in Y, \qquad \bar{p} \in H, \quad \bar{\alpha} \in \mathbb{R}^2 \tag{9a}$$

$$\bar{\alpha}_i \geq 0 \quad \text{for } i = 1, 2, \qquad \sum_{i=1}^{2} \bar{\alpha}_i = 1, \tag{9b}$$

$$-\Delta \bar{y} + \kappa \max\{0, \bar{y}\} = \mathcal{B}(\bar{u}) \quad \text{in } V', \tag{9c}$$

$$-\Delta \bar{p} + \kappa \mathbb{1}_{\{\bar{y}>0\}} \bar{p} = \sum_{i=1}^{2} \bar{\alpha}_i j_i'(\bar{y}) \quad \text{in } V', \tag{9d}$$

$$\mathcal{B}^*(\bar{p}) + \sum_{i=1}^{2} \bar{\alpha}_i \sigma_i \bar{u} = 0 \quad \text{in } U, \tag{9e}$$

which coincides with the strong stationarity system in the case $U = H$. If $U = \mathbb{R}^p$, we additionally check the sign condition

$$\bar{p} \leq 0 \quad \text{a.e. in } \{\bar{y} = 0\} \tag{10}$$

for $\bar{p}$ a posteriori. If the condition is satisfied, then the solution is still a solution to (5) and therefore a weak Pareto stationary point, see Corollary 2. In the weighted-sum algorithm, we can now set $\alpha_1 = 1 - \alpha_2$ and solve the stationarity system of the WSM for varying $\alpha_2 \in [0, 1]$, where $\alpha_2 \neq 0$ is required to guarantee well-posedness of the problems. Specifically, we introduce an additional small parameter $\alpha_{\mathrm{tol}} > 0$ and choose α_2 in $[\alpha_{\mathrm{tol}}, 1 - \alpha_{\mathrm{tol}}]$. The final procedure of the WSM is summarized in Algorithm 1.

Algorithm 1: Weighted-sum method (WSM)

Require: Number $k_{\max} \in \mathbb{N}$ of discretization points, $\alpha_{\text{tol}} > 0$;
Return : Discrete approximations $\tilde{\mathcal{P}}_s^{sw}$ and $\tilde{\mathcal{P}}_f^{sw}$ of weakly Pareto stationary points and front;
Set $\tilde{\mathcal{P}}_s^{sw} = \tilde{\mathcal{P}}_f^{sw} = \emptyset$;
for $i = 1, \ldots, k_{\max}$ **do**
  Set $\alpha_2 = \alpha_{\text{tol}} + (1 - 2\alpha_{\text{tol}}) \frac{i-1}{k_{\max}-1}$;
  Solve (9) with weight $(1 - \alpha_2, \alpha_2)$, check (10) and save a solution as u^i;
  Set $\tilde{\mathcal{P}}_s^{sw} = \tilde{\mathcal{P}}_s^{sw} \cup \{u^i\}$, $\tilde{\mathcal{P}}_f^{sw} = \tilde{\mathcal{P}}_f^{sw} \cup \{\hat{\mathcal{J}}(u^i)\}$;

The result of Algorithm 1 is a discrete approximation of the set of weakly Pareto stationary points and the corresponding front.

3.2 Reference point method (RPM)

The *reference point problem* with Euclidean norm $\|\cdot\|_2$ for a reference point $z \in \mathbb{R}^2$ is given by

$$\begin{aligned} &\min_{(y,u)} \mathcal{F}_z(y,u) = \frac{1}{2}\|\mathcal{J}(y,u) - z\|_2^2 \\ &\text{s.t. } (y,u) \in V \times U \text{ satisfies } -\Delta y + \kappa \max\{0, y\} = \mathcal{B}(u) \text{ in } V'. \end{aligned} \tag{$\mathbf{P}_z$}$$

The reference point method is based on solving ($\mathbf{P}_z$) for varying reference points. As the next theorem shows, optimizers to the reference point problems are Pareto optimal.

Theorem 4 *Every local* (*global*) *solution to* ($\mathbf{P}_z$) *such that* $z \in \mathbb{R}^2$ *with* $\hat{\mathcal{J}}(u) - z > 0$ *holds*, *is also a local* (*global*) *Pareto optimal point of* (**P**).

Proof We assume that $\bar{u} \in U$ with $\bar{y} = \mathcal{S}(\bar{u})$ is a local solution to ($\mathbf{P}_z$), i.e., there exists an $r_1 > 0$ such that for all $u \in U$ with $\|\bar{u} - u\|_U < r_1$ the inequality $\mathcal{F}_z(\bar{y}, \bar{u}) \leq \mathcal{F}_z(\mathcal{S}(u), u)$ is satisfied. Now, we assume that $\bar{u}$ is not locally Pareto optimal, which implies that for every $r_2 > 0$ there exists $u_{\text{end}} \in U$ with $\|u_{\text{end}} - \bar{u}\| < r_2$ and $\mathcal{J}_i(\mathcal{S}(u_{\text{end}}), u_{\text{end}}) \leq \mathcal{J}_i(\bar{y}, \bar{u})$ for $i = 1, 2$, where the latter inequality is strict for at least one i. Since we can choose r_2 arbitrarily small, $\mathcal{S}$ and $\mathcal{J}_i$ are continuous and $\mathcal{J}_i(\bar{y}, \bar{u}) - z_i > 0$ holds for $i = 1, 2$ by assumption, this implies that

$$z_i \leq \mathcal{J}_i(\mathcal{S}(u_{\text{end}}), u_{\text{end}}) \leq \mathcal{J}_i(\bar{y}, \bar{u}), \quad i = 1, 2,$$

where the second inequality is strict for at least one i. Since the Euclidean norm is strictly monotone [8, Definition 4.19], this implies the contradiction $\mathcal{F}_z(\mathcal{S}(u_{\text{end}}), u_{\text{end}}) < \mathcal{F}_z(\bar{y}, \bar{u})$ and proves the first inclusion. The statement for global solutions follows immediately from the first statement by choosing $r_1 = \infty$. □

Primal stationarity conditions for $(\mathbf{P}_z)$ are shown in the next theorem.

Theorem 5 *Let the control $\bar{u} \in U$ with associated state $\bar{y} = \mathcal{S}(\bar{u}) \in Y$ be locally optimal for $(\mathbf{P}_z)$. Then, we have for all $h \in U$*

$$\sum_{i=1}^{2} \big(\mathcal{J}_i(\bar{y}, \bar{u}) - z_i\big)\big(\langle j_i'(\bar{y}), \mathcal{S}'(\bar{u};h)\rangle_{Y',Y} + \sigma_i \langle \bar{u}, h\rangle_U\big) \geq 0. \tag{11}$$

Proof Due to the chain rule for Hadamard differentiable functions, this follows analogously to [7, Theorem 3.1]. □

A control $\bar{u} \in U$ with associated state $\bar{y} = \mathcal{S}(\bar{u})$ is called a *stationary point of* $(\mathbf{P}_z)$ for $z \in \mathbb{R}^2$ if (11) is satisfied.

Corollary 3 *Let the control $\bar{u} \in U$ with associated state $\bar{y} = \mathcal{S}(\bar{u})$ be given.*

1) *The following are equivalent*:

(a) *The control $\bar{u}$ is a stationary point of $(\mathbf{P}_z)$.*
(b) *There exists an adjoint state $\bar{p}$ such that $\bar{u}, \bar{y}, \bar{p}$ satisfy*

$$\bar{u} \in U, \quad \bar{y} \in Y, \quad \bar{p} \in H \tag{12a}$$

$$-\Delta \bar{y} + \kappa \max\{0, \bar{y}\} = \mathcal{B}(\bar{u}) \quad \text{in } V', \tag{12b}$$

$$\langle -\Delta \bar{p} + \kappa \mathbb{1}_{\{\bar{y}>0\}} \bar{p}, w\rangle_{Y',Y} \leq \sum_{i=1}^{2} \big(\mathcal{J}_i(\bar{y}, \bar{u}) - z_i\big)\langle j_i'(\bar{y}), w\rangle_{Y',Y} \quad \text{for all } w \in \operatorname{Im}(\mathcal{S}'(\bar{u};\cdot)), \tag{12c}$$

$$\bar{p} + \sum_{i=1}^{2} \big(\mathcal{J}_i(\bar{y}, \bar{u}) - z_i\big)\sigma_i B^{\dagger *}(\bar{u}) = 0 \quad \text{in } H. \tag{12d}$$

2) *The following are equivalent*:

(a) *There exists $z \in \mathbb{R}^2$ such that the control $\bar{u}$ is a stationary point of $(\mathbf{P}_z)$ with $0 \neq \mathcal{J}(\bar{y}, \bar{u}) - z \geq 0$ (or $\mathcal{J}(\bar{y}, \bar{u}) - z > 0$).*
(b) *There exists $\alpha \in \mathbb{R}^2$ with $\alpha \geq 0$ for $i = 1, 2$ (or $\alpha > 0$) and $\alpha_1 + \alpha_2 = 1$ such that the control $\bar{u}$ is a stationary point of $(\mathbf{P}_\alpha)$.*

Proof Part 1) follows analogously to the proof of Corollary 2.

To show part 2), let $\bar{u}$ be a stationary point of $(\mathbf{P}_z)$ with associated state $\bar{y} = \mathcal{S}(\bar{u})$ such that $0 \neq \mathcal{J}(\bar{y}, \bar{u}) - z \geq 0$. Then part 1) implies that there exists an adjoint state

$\bar{p}$ such that $\bar{u}, \bar{y}, \bar{p}$ solve (12). Accordingly, with normalized weight α and adjoint $\tilde{p}$ given by

$$\alpha_i = \tilde{\alpha}_i \Big/ \sum_{j=1}^{2} \tilde{\alpha}_j \text{ with } \tilde{\alpha}_i = \mathcal{J}_i(\bar{y}, \bar{u}) - z_i \geq 0, \quad \tilde{p} = \bar{p} \Big/ \sum_{i=1}^{2} \tilde{\alpha}_i,$$

system (5) is also satisfied. Thus Corollary 2 implies that $\bar{u}$ is a stationary point of $(\mathbf{P}_\alpha)$ with $\alpha \geq 0$ and $\alpha_1 + \alpha_2 = 1$. The other implication follows analogously without normalization by choosing the reference point $z = \mathcal{J}(\bar{y}, \bar{u}) - \alpha$, since then $0 \neq \mathcal{J}(\bar{y}, \bar{u}) - z = \alpha \geq 0$. The cases with strict inequalities follow analogously. □

Corollary 3 shows that primal stationarity in the two scalarization methods is essentially equivalent. Also note that, except for rather degenerate choices of reference points, system (12) is equivalent to the adjoint multiobjective system (5), i.e., (5) is a strong stationarity condition for the reference point problem $(\mathbf{P}_z)$. Again, we generally cannot solve (12) directly in the implementation when $U = \mathbb{R}^p$ because of the variational inequality on a possibly unknown and nonlinear image set. We proceed analogously to the modifications in the WSM, cf. (9), and instead solve

$$\bar{u} \in U, \quad \bar{y} \in Y, \quad \bar{p} \in H, \tag{13a}$$

$$-\Delta \bar{y} + \kappa \max\{0, \bar{y}\} = \mathcal{B}(\bar{u}) \qquad \text{in } V', \tag{13b}$$

$$-\Delta \bar{p} + \kappa \mathbb{1}_{\{\bar{y}>0\}} \bar{p} = \sum_{i=1}^{2} \big(\mathcal{J}_i(\bar{y}, \bar{u}) - z_i\big) j_i'(\bar{y}) \qquad \text{in } V', \tag{13c}$$

$$\mathcal{B}^*(\bar{p}) + \sum_{i=1}^{2} \big(\mathcal{J}_i(\bar{y}, \bar{u}) - z_i\big) \sigma_i \bar{u} = 0 \qquad \text{in } U. \tag{13d}$$

For $U = L^2(\Omega)$, this again coincides with the strong stationarity system from [7]. For $U = \mathbb{R}^p$ we test for the sign condition (10) a posteriori. If it is satisfied as well, $\bar{u}$ is a weak Pareto stationary point of $(\mathbf{P})$.

Another central question for the RPM is how suitable reference points can be chosen in the numerical implementation. To this end, we follow the approach presented in [2]. Let $k_{\max}$ denote the maximal number of Pareto stationary points in the numerical implementation and let (y^1, u^1) denote an initial starting point with u^1 being a stationary point of the weighted-sum problem with weights $\alpha_1 = 1 - \alpha_{\text{tol}}$ and $\alpha_2 = \alpha_{\text{tol}} \ll 1$. Then the first reference point z^2 (corresponding to the second point on the front) is chosen as

$$z^2 = \mathcal{J}(y^1, u^1) - \begin{pmatrix} h^\perp \\ h^\parallel \end{pmatrix}, \tag{14}$$

where $h^\perp, h^\parallel > 0$ are scaling parameters. For $i = 2, \ldots, k_{\max} - 2$ the reference point z^{i+1} is chosen as

$$z^{i+1} = \mathcal{J}(y^i, u^i) + h^{\parallel} \cdot \frac{\varphi^{\parallel}}{\|\varphi^{\parallel}\|} + h^{\perp} \cdot \frac{\varphi^{\perp}}{\|\varphi^{\perp}\|}, \tag{15}$$

with $\varphi^{\perp} = z^i - \mathcal{J}(y^i, u^i)$ and $\varphi^{\parallel} = (-\varphi_2^{\perp}, \varphi_1^{\perp})^T$. Note that due to the strong weighting of $\mathcal{J}_1$ at (y^1, u^1), the Pareto front is approximately vertical in the area of the first reference point. This motivates the initial choice $\varphi^{\parallel} = (0, -1)^T$ and $\varphi^{\perp} = (-1, 0)^T$.

Using this update technique, we end up with the reference point method stated in Algorithm 2.

Algorithm 2: Reference point method (RPM)

Require: Maximal number $k_{\max} \in \mathbb{N}$ of Pareto stationary points, recursive parameters $h^{\parallel}, h^{\perp} > 0$, weighted-sum parameter $0 < \alpha_{\text{tol}} \ll 1$;

Return : Discrete approximations $\tilde{\mathcal{P}}_s^{sw}$ and $\tilde{\mathcal{P}}_f^{sw}$ of weakly Pareto stationary points and front;

Compute solution (y^1, u^1) to (9) with $(1 - \alpha_{\text{tol}}, \alpha_{\text{tol}})$;

Compute solution $(y_{\text{end}}, u_{\text{end}})$ to (9) with $(\alpha_{\text{tol}}, 1 - \alpha_{\text{tol}})$;

Set $\tilde{\mathcal{P}}_s^{sw} = \{u^1\}$, $\tilde{\mathcal{P}}_f^{sw} = \{\hat{\mathcal{J}}(u^1)\}$ and $i = 2$;

Compute reference point z^i using (14);

while $z_1^{i+1} < \mathcal{J}_1(y_{\text{end}}, u_{\text{end}})$ **and** $i \leq k_{\max} - 1$ **do**

  Compute solution (y^i, u^i) to (13) with reference point z^i and check (10);

  Set $i = i + 1$, $\tilde{\mathcal{P}}_s^{sw} = \tilde{\mathcal{P}}_s^{sw} \cup \{u^{i-1}\}$, $\tilde{\mathcal{P}}_f^{sw} = \tilde{\mathcal{P}}_f^{sw} \cup \{\hat{\mathcal{J}}(u^{i-1})\}$;

  Compute reference point z^i using (15);

Set $\tilde{\mathcal{P}}_s^{sw} = \tilde{\mathcal{P}}_s^{sw} \cup \{u_{\text{end}}\}$ and $\tilde{\mathcal{P}}_f^{sw} = \tilde{\mathcal{P}}_f^{sw} \cup \{\hat{\mathcal{J}}(u_{\text{end}})\}$;

Note that the stopping criterion implies that if $k_{\max}$ is large enough, then the upper left as well as the lower right corner points of the Pareto front coincide with those of the WSM. If $0 \neq \mathcal{J}(\mathcal{S}(\bar{u}), \bar{u}) - z \geq 0$ holds for all $\bar{u} \in \tilde{\mathcal{P}}_s^{sw}$, then the result of Algorithm 2 is a discrete approximation of the set of weak Pareto stationary points and the corresponding Pareto front. If one wants to ensure this condition a priori, it is possible to, e.g., choose fixed reference points on shifted coordinate axes. The shift has to be performed such that all reference points are below the lower bounds on $\mathcal{J}_i$, cf. [3].

4 Numerical implementation

For the numerical realization and tests of the algorithms, we will assume that $j_1(y) = \frac{1}{2}\|y - y^d\|_H^2$, $j_2(y) = 0$ and $\sigma_1 = 0$. We fix the domain $\Omega = (0, 1)^2$ and consider P_1-type finite elements (FE) on a Friedrichs-Keller triangulation of the domain. The measure of fineness of the grids will be $h > 0$, which denotes the inverse number of square cells per dimension – i.e., the grid will have $2/h^2$ triangles. We write the coefficient vector of the piecewise linear interpolant of a function $w : \Omega \to \mathbb{R}$ on the grid vertices in sans-serif font (i.e., $\mathsf{w} \in \mathbb{R}^N$) and use the same font for the matrices in the

discretized settings. We resort to mass lumping for the nonlinear max-term in order to be able to evaluate it componentwisely. Inevitably, this introduces a numerical discretization error. Its effects decrease with increasing fineness of the discretization but increase with the coefficient κ that scales the nonlinearity. The corresponding stiffness matrix $\mathsf{K} \in \mathbb{R}^{N\times N}$, mass matrix $\mathsf{M} \in \mathbb{R}^{N\times N}$ and lumped mass matrix $\tilde{\mathsf{M}} \in \mathbb{R}^{N\times N}$ are given from the FE ansatz functions φ_i, $i = 1, \dots, N$, as

$$\mathsf{K}_{ij} = \langle \nabla\varphi_i, \nabla\varphi_j \rangle_H,\ \mathsf{M}_{ij} = \langle \varphi_i, \varphi_j \rangle_H,\ \tilde{\mathsf{M}} = \operatorname{diag}\left(\frac{|\operatorname{supp}(\varphi_i)|}{3} \,:\, i = 1, \dots, N \right).$$

Thus the FE approximation of (9c, e) introduced for the WSM is

$$\begin{pmatrix} \mathsf{K}\bar{\mathsf{y}}_h + \kappa\tilde{\mathsf{M}} \max\{0, \bar{\mathsf{y}}_h\} - \mathsf{B}\bar{u} \\ \mathsf{K}\bar{\mathsf{p}}_h + \kappa\tilde{\mathsf{M}}\Theta(\bar{\mathsf{y}}_h)\bar{\mathsf{p}}_h - \alpha_1\mathsf{M}(\bar{\mathsf{y}}_h - \mathsf{y}^d) \\ \mathsf{B}^T\bar{\mathsf{p}}_h + \alpha_2\sigma_2\mathsf{A}\bar{u} \end{pmatrix} = 0 \tag{16}$$

for some given $\alpha \in \mathbb{R}^2$ that satisfies (9a, b), where B is the FE-discretized version of the linear operator $\mathcal{B}$ and $\Theta := \Theta_0$ where $\Theta_x : \mathbb{R}^N \to \mathbb{R}^{N\times N}$ maps a vector to the diagonal matrix that takes the Heaviside function with functional value x at 0 evaluated for each entry of the vector as its diagonal entries. The matrix $\mathsf{A} = \mathsf{I}_p \in \mathbb{R}^{p\times p}$ is the identity if $U = \mathbb{R}^p$, and $\mathsf{A} = \mathsf{M}$ is the mass matrix if $U = H$. Note that this means that, depending on the space U, sometimes sans-serif notation would be appropriate for the (discretized) control u. To avoid any misunderstandings, we will always denote u without sans-serif style. These finite dimensional systems are solved with a globalized version of a pseudo-semismooth Newton (PSN) method (which essentially ignores the indicator functions' dependence on the state in the continuous system, i.e. the Heaviside functions' dependence in the discretized system, when the linearization of the systems is computed). For more details on the PSN without globalization, we refer to [5, 7]. The FE system matrix at iterates $(\mathsf{y}_h, \mathsf{p}_h, u)$ reads as:

$$\begin{pmatrix} \mathsf{K} + \kappa\tilde{\mathsf{M}}\Theta(\mathsf{y}_h) & 0 & -\mathsf{B} \\ -\alpha_1\mathsf{M} & \mathsf{K} + \kappa\tilde{\mathsf{M}}\Theta(\mathsf{y}_h) & 0 \\ 0 & \mathsf{B}^T & \alpha_2\sigma_2\mathsf{A} \end{pmatrix}. \tag{17}$$

We proceed analogously for the RPM and discretize (13 b, d) using finite elements, which yields

$$\begin{pmatrix} \mathsf{K}\bar{\mathsf{y}}_h + \kappa\tilde{\mathsf{M}} \max\{0, \bar{\mathsf{y}}_h\} - \mathsf{B}\bar{u} \\ \mathsf{K}\bar{\mathsf{p}}_h + \kappa\tilde{\mathsf{M}}\Theta(\bar{\mathsf{y}}_h)\bar{\mathsf{p}}_h - \left(\frac{1}{2}(\bar{\mathsf{y}}_h - \mathsf{y}^d)^T\mathsf{M}(\bar{\mathsf{y}}_h - \mathsf{y}^d) - z_1 \right)\mathsf{M}(\bar{\mathsf{y}}_h - \mathsf{y}^d) \\ \mathsf{B}^T\bar{\mathsf{p}}_h + \left(\frac{\sigma_2}{2}\bar{u}^T\mathsf{A}\bar{u} - z_2 \right)\sigma_2\mathsf{A}\bar{u} \end{pmatrix} = 0. \tag{18}$$

These discretized systems are solved with a PSN method as well. The FE system matrix at iterates $(\mathsf{y}_h, \mathsf{p}_h, u)$ reads as:

$$\begin{pmatrix} \mathsf{K} + \kappa\tilde{\mathsf{M}}\Theta(\mathsf{y}_h) & 0 & -\mathsf{B} \\ \mathsf{C}(\mathsf{y}_h) & \mathsf{K} + \kappa\tilde{\mathsf{M}}\Theta(\mathsf{y}_h) & 0 \\ 0 & \mathsf{B}^T & \mathsf{D}(u) \end{pmatrix} \tag{19}$$

with

$$\mathsf{C}(\mathsf{y}_h) := (\mathsf{M}(\mathsf{y}^d - \mathsf{y}_h))(\mathsf{M}(\mathsf{y}_h - \mathsf{y}^d))^T - \left(\tfrac{1}{2}(\mathsf{y}_h - \mathsf{y}^d)^T\mathsf{M}(\mathsf{y}_h - \mathsf{y}^d) - z_1\right)\mathsf{M}, \tag{20}$$

$$\mathsf{D}(u) := (\sigma_2\mathsf{A}u)(\sigma_2\mathsf{A}u)^T + \left(\frac{\sigma_2}{2}u^T\mathsf{A}u - z_2\right)\sigma_2\mathsf{A}. \tag{21}$$

Remark 2 For both methods, the sign condition (10) is not added into the discretized stationarity system and instead verified a posteriori if $U = \mathbb{R}^p$. ♢

Compared to the system matrix of the single objective case presented in [5], especially the system matrix of the RPM is more complicated and possesses a non-sparse substructure. This is due to the matrices $\mathsf{C}(\mathsf{y}_h)$ and $\mathsf{D}(u)$, which possess the dense terms $\mathsf{M}(\mathsf{y}_h - \mathsf{y}^d)(\mathsf{M}(\mathsf{y}_h - \mathsf{y}^d))^T$ and $(\sigma_2\mathsf{A}u)(\sigma_2\mathsf{A}u)^T$. These can cause severe memory and runtime problems when the reference point problem is solved on fine finite element grids with a linear solver. Due to these restrictions, the reference point method's subproblems will generally take longer to solve than the WSM, also on coarser grids.

One thing to keep in mind when applying the PSN method is that there is no guarantee of convergence as seen in the numerical examples in [7]. The method in fact shows failure to converge in practice, with the rate of failed attempts over the subproblems decaying as the grid discretization's fineness is increased. This suggest some sort of degeneration of the undamped search directions that could be countered with a globalization mechanism.

Accordingly, the two questions that we will address in the remainder of this section are the following:

1) Can the non-convergence issue of the PSN method itself be removed?
2) Is it possible to reduce computation times and memory problems of the (linear) PSN steps in the reference point method to make it competitive in terms of computation times?

4.1 Globalized PSN

The numerical experiments in [7] indicate that non-convergence of the PSN is an issue that strongly depends on (insufficiently fine) discretizations. As a stabilization approach independently of the grid fineness, we will present and test a line-search globalization of the PSN based on results for semismooth Newton methods as in, e.g., [9] and [11], which will be referred to as the gPSN method. Let us assume that we want to find a root of a function $F : \mathbb{R}^{2N+p} \to \mathbb{R}^{2N+p}$ with system matrix $G : \mathbb{R}^{2N+p} \to \mathbb{R}^{(2N+p)\times(2N+p)}$.

This will either be (16) with system matrix (17) for the subproblems in the WSM or (18) with system matrix (19) for the subproblems in the RPM. We will employ a line-search globalization with the merit function $\Lambda : \mathbb{R}^{2N+p} \to \mathbb{R}$, $x \mapsto \frac{1}{2}\|F(x)\|^2$.

Remark 3 Note that the norm in the merit function Λ is the discrete equivalent of the norm in $V' \times V' \times U$. Hence it is generally expensive to evaluate, since we need to compute a Riesz representative. However, we will precompute the necessary factorizations to speed-up the computations to some extent. ◊

The gPSN method is summarized in the following algorithm. Note that we cannot conclude – e.g., from theory on globalized semismooth Newton methods – that the algorithm converges without introducing a maximum number $k_{\max}$ of PSN steps and a minimum step length $\epsilon_2 > 0$.

Algorithm 3: Globalized PSN Method (gPSN)

Require: Initial point $x^0 \in \mathbb{R}^{2N+p}$, tolerances $\epsilon_1, \epsilon_2 > 0$, maximum number of iterations $k_{\max} \in \mathbb{N}$ and line-search parameters $\beta \in (0,1)$, $\gamma \in (0, \frac{1}{2})$;
Return : Approximated root $\bar{x} \in \mathbb{R}^{2N+p}$;
Set $i = 0$;
while $\sqrt{2\Lambda(x^i)} > \epsilon_1$ **and** $i < k_{\max}$ **do**
  Compute line-search direction d^i by solving $G(x^i)d^i = -F(x^i)$;
  Set $k_i = 0$;
  while $\Lambda(x^i + \beta^{k_i} d^i) > (1 - 2\gamma\beta^{k_i})\Lambda(x^i)$ **and** $\beta^{k_i+1} > \epsilon_2$ **do**
    Set $k_i = k_i + 1$;
  Set $x^{i+1} = x^i + \beta^{k_i} d^i$ and $i = i + 1$;

It turns out that there are examples, where Algorithm 3 converges while a refinement of the grid does not yield convergence, see Sect. 5. However, it can still happen that the gPSN method does not converge. Obviously, it would not be a good idea to add the final iterate of the gPSN method to the Pareto set nonetheless. Instead, if within the RPM the PSN does not converge, we update the reference point as follows: If no previous solution to the reference point problem is available, we choose

$$z^{i+1} = z^i - \begin{pmatrix} 0 \\ h^{\|} \end{pmatrix}.$$

If previous solutions to the reference point problem are available, we choose

$$z^{i+1} = z^i + h^{\|} \frac{\varphi^{\|}}{\|\varphi^{\|}\|}.$$

Essentially, previous information is used repeatedly to find a new reference point by going into the same parallel direction. If the gPSN is used within the WSM and does not converge, we can simply proceed to the next discretized weight.

4.2 A Preconditioned Matrix-Free L-GMRES Method

We will now focus on how to speed-up the computation and how to overcome the difficulties arising from the dense terms in the RPM. Notice that both dense terms are rank-1-matrices. Therefore it is easy to implement the matrix-vector-product for some $w \in \mathbb{R}^N$ and $v \in \mathbb{R}^N$ with $v = \mathsf{M}(\mathsf{y}_h - \mathsf{y}^d)$ or $v = \sigma_2 \mathsf{A}u$:

$$\left(vv^T\right)w = \left(v^T w\right)v.$$

This motivates the use of an iterative solver that only relies on matrix-vector-products in each PSN step for solving the reference point subproblem. Since the system is not symmetric positive definite, the CG method is not an alternative and we will use L-GMRES instead. As the performance heavily depends on the condition number of the system matrix (see [18]), which might be very large, especially for very small values of the regularization parameter σ_2, we will precondition the method with one of the following preconditioners:

1. **a:** The dense terms $\mathsf{M}(\mathsf{y}_h - \mathsf{y}^d)(\mathsf{M}(\mathsf{y}_h - \mathsf{y}^d))^T$ and $(\sigma_2 \mathsf{A}u)(\sigma_2 \mathsf{A}u)^T$ are omitted in an approximated system matrix that is used as a preconditioner.
2. **aBJ:** A block Jacobi preconditioner is applied with the approximation described for the preconditioner **a**.
3. **aBGS:** A block Gauss-Seidel preconditioner is applied with the approximation described for the preconditioner **a**.
4. **aILU:** An incomplete LU factorization together with the approximation described for the preconditioner **a** is applied.

Of course the same iterative, preconditioned approach can be implemented for the WSM, with the only difference being that no approximation – by ignoring dense terms – is necessary for the preconditioner. Note that also standard Jacobi, Gauss-Seidel, block Jacobi and block Gauss-Seidel and incomplete LU factorization preconditioners were tested. But the first two did not give any speed-up and the last four were still significantly less effective than the preconditioners above due to the dense terms still remaining. As expected, it is quite important to use the block structure of the problem as much as possible and to avoid the dense terms.

Remark 4 As long as $\sigma_2 u^T \mathsf{A}u/2 - z \neq 0$, the invertibility of the **aBJ** preconditioner is ensured for the RPM, since the first two block diagonal elements are symmetric positive definite and the last block diagonal element is symmetric and either positive or negative definite (see (19)). Analogously in case of the WSM, the invertibility is always ensured. ◊

For the four different preconditioners above, we propose three different update strategies. Those strategies are:

1. **Never update:** Only one preconditioner is generated for the first iteration of the first subproblem and then this preconditioner is used for all subproblems and all gPSN iterations.
2. **Update once:** One preconditioner is generated for each subproblem and then used for all gPSN iterations.
3. **Always update:** The preconditioner is generated for each gPSN iteration of each subproblem.

5 Numerical examples

In this section, we present numerical results for two examples – one with finite and one with infinite dimensional control space. First, the focus of our exposition will be on the performance of the RPM and the different preconditioning strategies. After the best update strategy is identified, we will compare RPM and WSM method.

In order to reasonably quantify the quality of the approximation of the respective Pareto (stationary) fronts, we employ two quality measures. The maximal distance between neighboring points on the Pareto front

$$\Delta_{\max} := \max_{a \in \tilde{\mathscr{P}}_f} \min_{b \in \tilde{\mathscr{P}}_f \setminus \{a\}} \|a - b\|_2 \tag{22}$$

will be our first measure. As a second measure of approximation quality, we will consider

$$\Delta_{\text{clust}} = \frac{|\tilde{\mathscr{P}}_f| \Delta_{\max}}{\sum_{a \in \tilde{\mathscr{P}}_f} \min_{b \in \tilde{\mathscr{P}}_f \setminus \{a\}} \|a - b\|_2}, \tag{23}$$

which is the maximum shortest distance between points on the front divided by the average shortest distance and therefore bounded from below by one. If this quantity is small, this indicates that the approximation quality is somewhat uniform across the entire Pareto front, while a large value indicates that some parts of the Pareto front are approximated better than others are, i.e., a localized clustering.

Note that in all results presented here, the sign condition (see Remark 2) is satisfied and the gPSN method always converges.

Our code is implemented in Python3 and uses FEniCS [1] for the matrix assembly. Sparse memory management and computations (especially L-GMRES) are implemented with SciPy [20]. All computations below were run on an Ubuntu 20.04 notebook with 32 GB main memory and an Intel Core i7-8565U CPU.

5.1 The numerical examples

First we introduce the two examples. The parameters listed in Table 1 are fixed for the rest of this work.

5.1.1 Example 1 – Infinite dimensional controls

For the first numerical example the desired state is chosen as $y^d = \mathbb{1}_{\Omega_1} - \mathbb{1}_{\Omega_2}$ with $\Omega_1 = \{(x_1, x_2) \in \Omega : x_1, x_2 > 1/3\}$ and $\Omega_2 = \{(x_1, x_2) \in \Omega : x_1, x_2 < 2/3\}$. This desired state is chosen to promote nonsmoothness. The control space U is chosen as $H = L^2(\Omega)$, the operator $\mathcal{B}$ as the identity on U and A is the mass matrix.

5.1.2 Example 2 – finite dimensional controls

For the second numerical example, we choose $y^d(x) = \left(\frac{1}{2} - x_1\right) \sin(\pi x_1) \sin(\pi x_2)$. The space U is chosen as $\mathbb{R}^2$, $\mathsf{A} = \mathsf{I}_2$ is the identity matrix in $\mathbb{R}^{2\times 2}$ and the operator $\mathcal{B}$ is set to

$$(\mathcal{B}(u))(\boldsymbol{x}) = 10 \cdot \begin{cases} u_1 x_1 x_2, & \text{for } \boldsymbol{x} = (x_1, x_2) \text{ and } x_1 \leq \frac{1}{2}, \\ u_2 x_1^2 x_2^2, & \text{otherwise,} \end{cases}$$

where the definition is to be understood as L^2-functions mapping $\boldsymbol{x} \in \Omega$ to $\mathbb{R}$ that are embedded into V'. For plots of the operator $\mathcal{B}$ and the desired state we refer to [5, Fig. 1 and 4].

5.2 Preconditioning the reference point method

In this section, we consider the different preconditioning approaches for the RPM. Therefore, we additionally choose the parameter $\sigma_2 = 5 \cdot 10^{-3}$ and the parameters $h^{\perp} = 10$, $h^{\parallel} = 0.1$ in the RPM and $\epsilon_1 = 1 \cdot 10^{-4}$ for the gPSN. Note that the arguably large value of ϵ_1 is necessary for the L-GMRES method to converge without preconditioner, because the (sometimes badly conditioned) problem is numerically difficult to solve. The results for Example 1 are given in Table 2.

First of all, we can see that the computation time decreases for all preconditioning approaches. Also the average number of L-GMRES iterations is very small (between 2 and 3.5) and increases only slightly for smaller step sizes h. The latter observation is in contrast to the performance of the non-preconditioned solving, which starts out with a large number of average L-GMRES iterations that nonetheless significantly increases for smaller step sizes h. Furthermore, we can see that cheaper preconditioners lead to a larger speed-up if the preconditioner is updated more often, i.e. with the preconditioning strategy **always update** the preconditioner **aBJ** still gives a significant speed-up of about 35, but all other preconditioners cannot give a speed-up above 9. Nonetheless, the best preconditioning approach surprisingly is the **never update** strategy

Table 1 Fixed parameters for the two numerical examples

gPSN			RPM		PDE	
β	γ	ϵ_2	$k_{\max}$	α_{tol}	Ω	κ
$5 \cdot 10^{-1}$	10^{-1}	10^{-3}	10^3	10^{-2}	$(0,1)^2$	10

Table 2 Example 1. Comparison of average L-GMRES iterations (av. it.) and speed-up (s.-up) of the RPM for different preconditioning approaches and different step sizes h for $\sigma_2 = 5 \cdot 10^{-3}$

$1/h$	a		aBJ		aBGS		aILU		none	
	av. it.	s.-up	av. it.	s.-up	av. it.	s.-up	av. it.	s.-up	av. it.	time [s]
Never update										
50	3.61	8.82	3.17	23.51	3.24	16.49	3.57	9.19	83.65	$2.55 \cdot 10^2$
100	3.53	14.64	3.36	42.04	3.21	26.53	3.44	17.62	257.46	$2.02 \cdot 10^3$
200	3.51	20.65	3.38	63.46	3.23	35.28	3.53	23.07	380.42	$1.14 \cdot 10^4$
Update once										
50	2.55	10.37	2.92	17.71	2.88	10.01	2.10	7.92		
100	2.67	12.01	3.02	30.51	2.92	11.10	2.12	10.27		
200	2.85	9.40	3.19	39.75	2.98	8.04	2.34	8.20		
Always update										
50	2.35	7.75	2.93	15.45	2.95	7.43	2.00	5.58		
100	2.43	9.04	3.02	26.94	2.99	8.29	2.03	7.32		
200	2.62	7.20	3.15	35.54	3.05	6.29	2.32	6.28		

combined with the preconditioner **aBJ**. This indicates that the problem structure does not change significantly with respect to the current reference point and thus a computationally expensive update of the preconditioner is unnecessary.

Next, we consider the results for Example 2, which can be found in Table 3.

We basically observe the same behavior as previously and again **never update** and **aBJ** is the best preconditioning approach. Note that this finite dimensional example is inherently better conditioned and thus combinations of more expensive preconditioners such as **a**, **aBGS** and **aILU** and expensive preconditioning strategies such as **update once** and **always update** often lead to larger computation times compared to the performance in the absence of a preconditioner. Furthermore, the preconditioner **aILU** seems to behave unstably, since the number of average L-GMRES iterations increases significantly for smaller step sizes h. Note that **aILU** includes some parameters which could be varied and might improve this behavior, but we will not go into details here.

Next we consider a fixed step size $h = 1/100$ and investigate the behavior of the different preconditioning approaches for varying σ_2. We expect larger condition numbers for smaller values of σ_2 and therefore problems which are harder to solve numerically. Since the strategies **update once** and **always update** and the preconditioner **aILU** did not prove useful, they are excluded from this considerations. The results can be found in Table 4 for Example 1 and in Table 5 for Example 2.

In both examples the average number of L-GMRES iterations increases slightly for smaller values of σ_2. This increase is stronger for the first example. If, on the other hand, no preconditioner is used, there is a significant increase for smaller values of σ_2. This is especially true for the first example, which starts with an average number of 18.75 iterations for $\sigma_2 = 1$ and ends with an average number of 501.07 iterations for $\sigma_2 = 10^{-3}$. In the second example there is only an increase from 10.48

Table 3 Example 2. Comparison of average L-GMRES iterations (av. it.) and speed-up (s.-up) of the RPM for different preconditioning approaches and different step sizes h for $\sigma_2 = 5 \cdot 10^{-3}$

$1/h$	a		aBJ		aBGS		aILU		none	
	av. it.	s.-up	av. it.	s.-up	av. it.	s.-up	av. it.	s.-up	av. it.	time [s]
Never update										
50	2.79	3.43	2.78	4.16	2.95	3.21	2.79	3.37	9.97	$1.79 \cdot 10^1$
100	2.78	6.46	2.62	11.18	2.81	5.54	7.00	0.55	40.50	$1.91 \cdot 10^2$
200	2.72	14.73	2.38	59.41	2.67	12.83	15.23	0.98	197.35	$3.85 \cdot 10^3$
Update once										
50	2.10	0.94	2.19	2.87	2.05	1.20	2.04	0.70		
100	2.11	0.67	2.20	7.19	2.08	1.00	6.01	0.40		
200	2.18	0.53	2.16	30.25	2.12	0.70	12.95	0.74		
Always update										
50	2.05	0.67	2.17	2.36	2.05	0.88	2.17	0.48		
100	2.09	0.52	2.20	6.08	2.07	0.74	5.99	0.35		
200	2.15	0.41	2.16	25.44	2.11	0.53	12.95	0.65		

Table 4 Example 1. Comparison of average L-GMRES iterations (av. it.) and speed-up (s.-up) of the RPM for different preconditioning approaches and different values of σ_2 for fixed step size $h = 1/100$

σ_2	a		aBJ		aBGS		none	
	av. it.	s.-up	av. it.	s.-up	av. it.	s.-up	av. it.	time [s]
10^0	2.50	2.95	2.50	6.60	2.86	3.57	18.75	$6.44 \cdot 10^{-1}$
10^{-1}	3.47	5.82	2.98	14.46	3.12	8.59	57.65	$2.34 \cdot 10^0$
10^{-2}	3.37	15.08	3.27	40.73	3.12	25.18	201.42	$8.92 \cdot 10^0$
10^{-3}	3.53	19.46	3.46	62.84	3.28	40.75	501.07	$1.73 \cdot 10^1$

Table 5 Example 2. Comparison of average L-GMRES iterations (av. it.) and speed-up (s.-up) of the RPM for different preconditioning approaches and different values of σ_2 for fixed step size $h = 1/100$

σ_2	a		aBJ		aBGS		none	
	av. it.	s.-up	av. it.	s.-up	av. it.	s.-up	av. it.	time [s]
10^0	2.53	1.35	2.30	3.71	2.63	1.61	10.48	$2.46 \cdot 10^{-1}$
10^{-1}	2.61	2.03	2.21	5.30	2.11	2.90	13.41	$4.50 \cdot 10^{-1}$
10^{-2}	2.75	5.13	2.59	9.24	2.71	4.60	28.12	$8.91 \cdot 10^{-1}$
10^{-3}	2.68	9.68	2.88	12.64	2.93	6.78	52.10	$1.41 \cdot 10^0$

to 52.10. Nonetheless in both examples preconditioning pays off and again the preconditioner **aBJ** is the best. It results in a speed-up of 62.84 in the first example and a speed-up of 12.64 in the second example for $\sigma_2 = 10^{-3}$.

5.3 Comparison of the RPM and the WSM

In this section, we want to compare the performance of the reference point method and the weighted-sum method both in terms of computation times and discretization quality. We choose a step size of $h = 1/100$, a tolerance $\epsilon_1 = 10^{-5}$ in the gPSN and $h^{\perp} = 1$, $h^{\parallel} = 0.2$ in the RPM. We will consider $\sigma_2 = 1$ for both examples. This means that the problems are relatively well conditioned, but the Pareto fronts are harder to approximate than for smaller values of σ_2.

In order to make the results of the RPM and the WSM qualitatively comparable, we first run the RPM, which yields a number of discretization points on the front. Afterwards we run the WSM with $k_{\max}$ (the number of Pareto points) chosen as the number of discretization points generated by the RPM. At this point we have the same number of discretization points on the respective approximated fronts. However, the WSM tends to cluster the discretization points. In order to obtain comparable approximation quality, we then double the number of points in the WSM until the maximal distance for points on the Pareto front (see (22)) is below the maximal distance for the RPM. Afterwards, the parameter $h^{\parallel}$ is halved as often as $k_{\max}$ in the WSM was doubled before to compare the evolution of the quality measures $\Delta_{\max}$ and Δ_{clust} for an increasing size of the approximated Pareto front.

The Pareto fronts are shown in Fig. 1.

We can see that both methods approximate the same curve. But the WSM shows a clustering behavior in the lower right corner whilst giving only a poor approximation in the remainder of the Pareto front. This already indicates that some refinement of the weights' distribution is generally well advised for the WSM. In Fig. 2, the evolution of the quality measures for the WSM and RPM for the procedure described above is shown.

In the left figures, we can see that for the WSM the number of points on the Pareto front needs to be doubled seven times in order to reach a maximal distance of points on the Pareto front that is smaller than that of the RPM. Furthermore the approximation quality decreases every time the size of the Pareto front is doubled. This is due to the clustering behavior in the lower right corner. As a result, an unnecessarily large number of points on the Pareto front is needed to reach a desired maximal distance $\Delta_{\max}$. On the other hand, with the RPM, the approximation quality

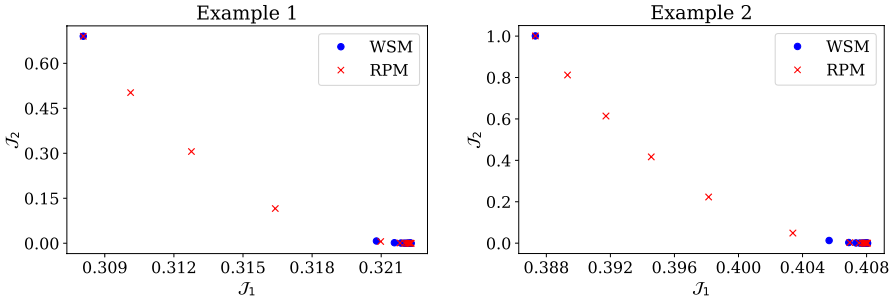

Fig. 1 Pareto fronts from WSM and RPM for step size $1/h = 100$ and $\sigma_2 = 1$

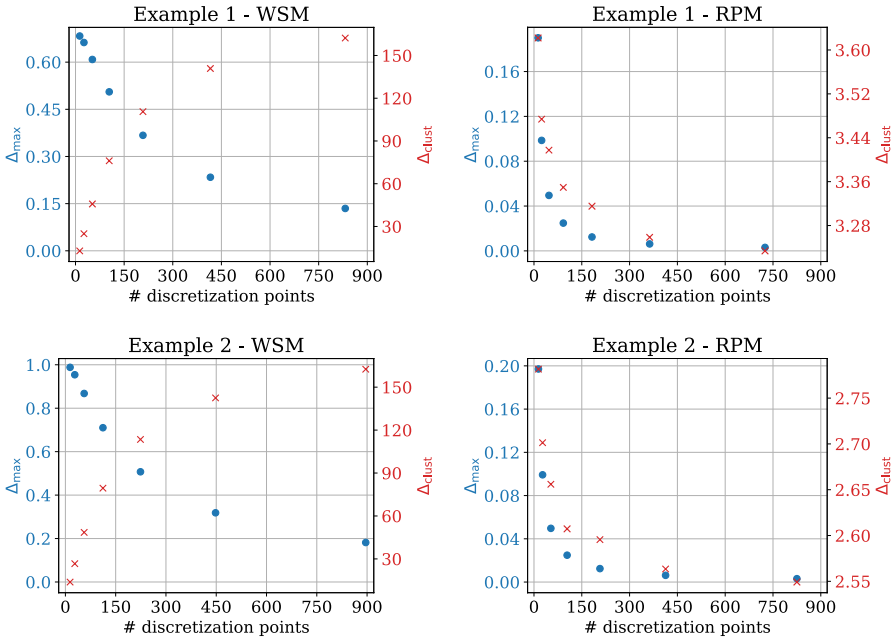

Fig. 2 Evolution of measures of approximation quality ($\Delta_{\max}$ and Δ_{clust}) for the WSM with respect to doubling the number of discretization points on the Pareto front and for the RPM with respect to halving $h^{\|}$ with $\sigma_2 = 1$ and step size $1/h = 100$

even decreases whilst $h^{\|}$ is halved. This behavior can be seen in the right figures and is even better than expected. Note that the size of the Pareto front is approximately doubled when $h^{\|}$ is halved.

The question remains, which method performs better in terms of computational cost. A comparison of the results from the RPM and the first and last result from the WSM in the procedure of doubling the number of discretization points is shown in Table 6.

The observation for both examples are similar with respect to sizes of the Pareto fronts and the measures of approximation quality. Also whilst the RPM is slightly slower than the WSM if the same size for the Pareto front is used, it is about 28 times faster than the WSM when an at least equally good approximation quality is desired.

6 Conclusion

If the controls on the right-hand-side of the constraining PDE to (**P**) are finite dimensional, then conditions that imply primal stationarity can be found. Those conditions can be interpreted as strong stationarity systems of scalarization methods. They are, however, not usable numerically because they contain unknown nonlinearities in the spaces that the conditions are formulated in. Modifying the conditions, we ended

Table 6 Comparison of computation time and approximation quality for RPM and WSM with $\sigma_2 = 1$. WSM (first) indicates WSM with the size of RPM. WSM (last) indicates WSM after doubling the size until a maximal distance on the Pareto front below that of RPM is reached. The step size is chosen as $h = 1/100$

		Time [s]	$\Delta_{\max}$	Δ_{clust}	$\lvert\tilde{\mathscr{P}}_f\rvert$
Ex 1	RPM	$2.42 \cdot 10^{0}$	$1.90 \cdot 10^{-1}$	$3.62 \cdot 10^{0}$	13
	WSM (first)	$1.41 \cdot 10^{0}$	$6.83 \cdot 10^{-1}$	$1.28 \cdot 10^{1}$	13
	WSM (last)	$6.99 \cdot 10^{1}$	$1.34 \cdot 10^{-1}$	$1.62 \cdot 10^{2}$	832
Ex 2	RPM	$1.61 \cdot 10^{0}$	$1.97 \cdot 10^{-1}$	$2.78 \cdot 10^{0}$	14
	WSM (first)	$1.04 \cdot 10^{0}$	$9.88 \cdot 10^{-1}$	$1.38 \cdot 10^{1}$	14
	WSM (last)	$4.49 \cdot 10^{1}$	$1.81 \cdot 10^{-1}$	$1.62 \cdot 10^{2}$	896

up with linear systems as in the case of ample controls. We have shown that both WSM and RPM can be applied to characterize the front of Pareto stationary points for this nonsmooth problem. The reference point method performs significantly better when both approximation quality and computation time are considered, as long as preconditioning is used intelligently in GMRES. In our tests, the preconditioning strategy **aBJ** without updates performs the best. We also saw that the line-search globalized version of the PSN method leads to better performance and convergence of the method over the basic version with uncontrolled step lengths. A simple reduction of the step size numerically does not appear to guarantee this behavior.

Acknowledgements The authors would like to express great gratitude to the anonymous reviewers, as they caught a major oversight in the analysis of this article's first version.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was supported by the German Research Foundation (DFG) under grant number VO 1658/5-2 within the priority program *Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization* (SPP 1962).

Data availability All data analyzed in this work was generated by the corresponding code and is available from the authors upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Alnaes, M., Blechta, J., Hake, J., Johansson, A., Kehlet, B., Logg, A., Richardson, C., Ring, J., Rognes, M.E., Wells, G.N.: The fenics project version 1.5. Arch. Numer. Softw. **3**(100), 9–23. https://doi.org/10.11588/ans.2015.100.20553
2. Banholzer, S.: POD-based bicriterial optimal control of convection-diffusion equations. Master's thesis, Universität, Konstanz (2017)
3. Banholzer, S.: ROM-based multiobjective optimization with PDE constraints. Ph.D. thesis, Universität Konstanz, Konstanz (2021)
4. Banholzer, S., Volkwein, S.: Hierarchical convex multiobjective optimization by the Euclidean reference point method (2019)
5. Bernreuther, M., Müller, G., Volkwein, S.: 1 reduced basis model order reduction in optimal control of a nonsmooth semilinear elliptic PDE. In: Optimization and Control for Partial Differential Equations, pp. 1–32. De Gruyter (2022)
6. Christof, C., Meyer, C., Walther, S., Clason, C.: Optimal control of a non-smooth semilinear elliptic equation. Math. Control Relat. Fields **8**(1), 247–276 (2018). https://doi.org/10.3934/mcrf.2018011
7. Christof, C., Müller, G.: Multiobjective optimal control of a non-smooth semilinear elliptic partial differential equation **27**, S13. https://doi.org/10.1051/cocv/2020060
8. Ehrgott, M.: Multicriteria Optimization, second edn. Springer, Berlin
9. Gerdts, M., Horn, S., Kimmerle, S.J.: Line search globalization of a semismooth Newton method for operator equations in Hilbert spaces with applications in optimal control. J. Ind. Manag. Optim. **13**(1), 47–62 (2017). https://doi.org/10.3934/jimo.2016003
10. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order. Springer, Berlin (2001)
11. Ito, K., Kunisch, K.: On a semi-smooth Newton method and its globalization. Math. Program. **118**(2), 347–370 (2007). https://doi.org/10.1007/s10107-007-0196-3
12. Khoromskij, B.N., Wittum, G.: Numerical Solution of Elliptic Differential Equations by Reduction to the Interface. Springer, Berlin (2004)
13. Kikuchi, F., Nakazato, K., Ushijima, T.: Finite element approximation of a nonlinear eigenvalue problem related to MHD equilibria. Jpn. J. Appl. Math. **1**(2), 369–403 (1984). https://doi.org/10.1007/bf03167065
14. Miettinen, K.: Nonlinear multiobjective optimization. Springer, US (1998)
15. Mueller-Gritschneder, D., Graeb, H., Schlichtmann, U.: A successive approach to compute the bounded pareto front of practical multiobjective optimization problems. SIAM J. Optim. **20**(2), 915–934 (2009). https://doi.org/10.1137/080729013
16. Rappaz, J.: Approximation of a nondifferentiable nonlinear problem related to MHD equilibria. Numer. Math. **45**(1), 117–133 (1984). https://doi.org/10.1007/bf01379665
17. Romaus, C., Bocker, J., Witting, K., Seifried, A., Znamenshchykov, O.: Optimal energy management for a hybrid energy storage system combining batteries and double layer capacitors. In: 2009 IEEE Energy Conversion Congress and Exposition. IEEE (2009)
18. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2 edn. SIAM
19. Temam, R.: A non-linear eigenvalue problem: the shape at equilibrium of a confined plasma. Arch. Ration. Mech. Anal. **60**(1), 51–73 (1975). https://doi.org/10.1007/bf00281469
20. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A.P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D.V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavi c, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T.,

Jones, T.R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., and, Y.V.B.: SciPy 1.0: fundamental algorithms for scientific computing in python. Nature Methods **17**(3), 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2

21. Xin, J.: An Introduction to Fronts in Random Media. Springer, New York (2009)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.