DFG Deutsche Forschungsgemeinschaft

Priority Programme 1962

POD-Based Bicriterial Optimal Control by the Reference Point Method

Stefan Banholzer, Dennis Beermann, Stefan Volkwein

SPP 1962

Non-smooth and Complementarity-based
Distributed Parameter Systems:
Simulation and Hierarchical Optimization

Preprint Number SPP1962-014

received on April 11, 2017

Edited by
SPP1962 at Weierstrass Institute for Applied Analysis and Stochastics (WIAS)
Leibniz Institute in the Forschungsverbund Berlin e.V.
Mohrenstraße 39, 10117 Berlin, Germany
E-Mail: spp1962@wias-berlin.de

World Wide Web: http://spp1962.wias-berlin.de/

POD-Based Bicriterial Optimal Control by the Reference Point Method

Stefan Banholzer [*] **Dennis Beermann** [*] **Stefan Volkwein** [**]

[*] *Department of Mathematics and Statistics, University of Konstanz, 78457 Konstanz, Germany*
[**] *Department of Mathematics and Statistics, University of Konstanz, 78457 Konstanz, Germany (e-mail: Stefan.Volkwein@uni-konstanz.de)*

Abstract: In the present paper a bicriterial optimal control problem governed by a parabolic partial differential equation (PDE) and bilateral control constraints is considered. For the numerical optimization the reference point method is utilized. The PDE is discretized by a Galerkin approximation utilizing the method of proper orthogonal decomposition (POD). POD is a powerful approach to derive reduced-order approximations for evolution problems. Numerical examples illustrate the efficiency of the proposed strategy.

Keywords: Optimal control, multiobjective optimization, reference point method, proper orthogonal decomposition, a-posteriori error analysis.

1. INTRODUCTION

In real applications, optimization problems are often described by introducing several objective functions conflicting with each other. This leads to *multiobjective* or *multicriterial* optimization problems; Ehrgott (2005), Miettinen (1998) or Stadler (1988). One prominent example is given by an energy efficient heating, ventilation and air-conditioning (HVAC) operation of a building with conflicting objectives such as minimal energy consumption and maximal comfort; Fong et al. (2006) and Kusiak et al. (2011). Finding the optimal control that represents a good compromise is the main issue in these problems. For that reason the concept of Pareto optimal or efficient points is developed. In contrast to scalar-valued optimization problems, the computation of a set of Pareto optimal points is required. Consequently, many scalar-valued constrained optimization problems have to be solved.

In this paper we apply the reference point method Romaus et al. (2009) in order to transform a bicriterial optimal control problem into a sequence of scalar-valued optimal control problems and to solve them using well-known optimal control techniques; see Tröltzsch (2010). Preliminary results combining reduced-order modeling and multiobjective PDE-constrained optimization are recently derived Iapichino et al. (2013), Iapichino et al. (2015) and Peitz et al. (2015).

The paper is organized as follows: In Section 2 we introduce the bicriterial optimization problem under consideration. The reference point method is explained in Section 3. Here we also derive the a-posteriori error estimator which is essential in our reduced-order approach. Section 4 is devoted to recall the POD method for optimal control problems. Numerical examples are presented in Section 5. Finally, a conclusion is drawn in Section 6.

Notation: Throughout this paper, if $x^1, x^2 \in \mathbb{R}^n$ are two vectors, we write $x^1 \leq x^2$ if $x_i^1 \leq x_i^2$ for $i = 1, ..., n$, and similarly for $x^1 < x^2$.

2. PROBLEM FORMULATION

2.1 The state equation

For time $T > 0$ the state equation is given by

$$
\begin{aligned}
y_t(t,\boldsymbol{x}) - \Delta y(t,\boldsymbol{x}) &= \sum_{i=1}^m u_i(t)\chi_i(\boldsymbol{x}) && \text{for } (t,\boldsymbol{x}) \in Q,\\
\frac{\partial y}{\partial \boldsymbol{n}}(t,\boldsymbol{x}) &= 0 && \text{for } (t,\boldsymbol{x}) \in \Sigma,\\
y(0,\boldsymbol{x}) &= y_\circ(\boldsymbol{x}) && \text{for } \boldsymbol{x} \in \Omega,
\end{aligned} \tag{1}
$$

where $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, is a bounded domain with Lipschitz-continuous boundary $\Gamma = \partial\Omega$ and $\boldsymbol{n}$ stands for the outward normal vector. We set $Q = (0,T) \times \Omega$ and $\Sigma = (0,T) \times \Gamma$. Let $H = L^2(\Omega)$ and $V = H^1(\Omega)$ be endowed by the canonical inner products; see Dautray and Lions (2000). The variable $u = (u_1, \ldots, u_m) \in \mathcal{U} = L^2(0,T)^m$ denotes the *control* and $\chi_i \in L^\infty(\Omega)$, $1 = 1, \ldots, m$, are given control shape functions. Furthermore, $y_\circ \in L^\infty(\Omega)$ denotes a given initial heat distribution. We write $y(t)$ when y is considered as a function in $\boldsymbol{x}$ only for fixed $t \in [0,T]$. Recall that

$$W(0,T) = \big\{\varphi \in L^2(0,T;V) \,\big|\, \varphi_t \in L^2(0,T;V')\big\}$$

is a Hilbert space endowed with the common inner product; see, e.g., Dautray and Lions (2000). A weak solution $y \in \mathcal{Y} = W(0,T)$ to (1) is called a *state* and has to satisfy for all test functions $\varphi \in V$:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle y(t),\varphi\rangle_H + \int_\Omega \nabla y(t)\cdot\nabla\varphi\,\mathrm{d}\boldsymbol{x} &= \sum_{i=1}^m u_i(t)\langle\chi_i,\varphi\rangle_H,\\
\langle y(0),\varphi\rangle_H &= \langle y_\circ,\varphi\rangle_H.
\end{aligned} \tag{2}
$$

⋆ The second author gratefully acknowledge financial support by the project *Hybrides Planungsverfahren zur energieeffizienten Wärme- und Stromversorgung von städtischen Verteilnetzen* funded by the German Ministry for Economic Affairs and Energy.

It is shown in Dautray and Lions (2000) that (2) admits a unique solution y and

$$\|y\|_{\mathcal{Y}} \le C\big(\|y_\circ\|_H + \|u\|_{\mathcal{U}}\big) \tag{3}$$

for a contant $C \ge 0$. We introduce the linear operator $\mathcal{S}: \mathcal{U} \to \mathcal{Y}$, where $y = \mathcal{S}u$ is the solution to (2) for given $u \in \mathcal{U}$ with $y_\circ = 0$. From (3) it follows that $\mathcal{S}$ is bounded. Moreover, let $\hat{y} \in \mathcal{Y}$ be the solution to (2) for $u = 0$. Then, the affine linear mapping $\mathcal{U} \ni u \mapsto y(u) = \hat{y} + \mathcal{S}u \in \mathcal{Y}$ is affine linear, and $y(u)$ is the weak solution to (1).

2.2 The multiobjective optimal control problem

For given $u_a, u_b \in \mathcal{U}$ with $u_a \le u_b$ in $\mathcal{U}$, the set of admissible controls is given as

$$\mathcal{U}_{\mathsf{ad}} = \big\{u \in \mathcal{U} \,\big|\, u_a(t) \le u(t) \le u_b(t) \text{ in } [0,T]\big\}.$$

Introducing the bicriterial cost functional

$$J : \mathcal{Y} \times \mathcal{U} \to \mathbb{R}^2, \quad J(y,u) = \frac{1}{2}\begin{pmatrix} \|y(T) - y_\Omega\|_H^2 \\ \|u\|_{\mathcal{U}}^2 \end{pmatrix}$$

the multiobjective optimal control problem (MOCP) reads

$$(\mathbf{P}) \qquad \min J(y,u) \quad \text{subject to (s.t.)} \quad (y,u) \in \mathcal{F}(\mathbf{P})$$

with the feasible set

$$\mathcal{F}(\mathbf{P}) = \big\{(y,u) \in \mathcal{Y} \times \mathcal{U}_{\mathsf{ad}} \,\big|\, y \text{ solves } (2)\big\}.$$

Next we define the reduced cost function $\hat{J} = (\hat{J}_1, \hat{J}_2) : \mathcal{U} \to \mathbb{R}^2$ by $\hat{J}(u) = J(\hat{y} + \mathcal{S}u, u)$ for $u \in \mathcal{U}$. Then, $(\mathbf{P})$ can be equivalently formulated as

$$(\hat{\mathbf{P}}) \qquad \min \hat{J}(u) \quad \text{s.t.} \quad u \in \mathcal{U}_{\mathsf{ad}}.$$

Problem $(\hat{\mathbf{P}})$ involves the minimization of a vector-valued objective. This is done by using the concepts of *order relation* and *Pareto optimality*; see, e.g., Ehrgott (2005). In $\mathbb{R}^2$ we make use of the following order relation: For all $z^1, z^2 \in \mathbb{R}^2$ we have

$$z^1 \le z^2 \Leftrightarrow z^2 - z^1 \in \mathbb{R}^2_+ = \big\{z \in \mathbb{R}^2 \,\big|\, z_i \ge 0 \text{ for } i = 1,2\big\}.$$

Definition 1. The point $\bar{u} \in \mathcal{U}_{\mathsf{ad}}$ is called *Pareto optimal* for $(\hat{\mathbf{P}})$ if there is no other control $u \in \mathcal{U}_{\mathsf{ad}} \setminus \{\bar{u}\}$ with $\hat{J}_i(u) \le \hat{J}_i(\bar{u})$, $i = 1,2$, and $\hat{J}_j(u) < \hat{J}_j(\bar{u})$ for at least one $j \in \{1,2\}$.

3. THE REFERENCE POINT METHOD

3.1 The reference point problem

The theoretical and numerical challenge is to present the decision maker with an approximation of the *Pareto front*

$$\mathcal{P} = \big\{\hat{J}(u) \,\big|\, u \in \mathcal{U}_{\mathsf{ad}} \text{ is Pareto optimal}\big\} \subset \mathbb{R}^2$$

In order to do so, we follow the ideas laid out in Peitz et al. (2015) and make use of the *reference point method*: Given a reference point $z = (z_1, z_2) \in \mathbb{R}^2$ that satisfies

$$z < \hat{J}(u) \quad \text{for all } u \in \mathcal{U}_{\mathsf{ad}} \tag{4}$$

we introduce the *distance function* $F_z : \mathcal{U} \to \mathbb{R}$ by

$$F_z(u) = \tfrac{1}{2}\,|\hat{J}(u) - z|^2 = \tfrac{1}{2}\big(\hat{J}_1(u) - z_1\big)^2 + \tfrac{1}{2}\big(\hat{J}_2(u) - z_2\big)^2.$$

The mapping F_z measures the geometrical distance between $\hat{J}(u)$ and z.

Lemma 2. The mapping F_z is strictly convex.

Proof. The mapping F_z is of the form $F_z = \sum_{i=1}^2 g_i \circ \hat{J}_i$ where, because of (4), we have $g_i : (z_i, \infty) \to \mathbb{R}_0^+$ with $g_i(\xi) = (\xi - z_i)^2/2$. Because of the affine linearity of $u \mapsto y(u)$, $\hat{J}_1$ is convex and $\hat{J}_2$ strictly convex. Further, g_i is strictly convex and monotone increasing for $i = 1,2$. Altogether, F_z itself is strictly convex. □

Suppose that z is componentwise strictly smaller than every objective value which we can achieve within $\mathcal{U}_{\mathsf{ad}}$. The goal is that – by approximating z as best as possible – we get a Pareto optimal point for $(\hat{\mathbf{P}})$. Therefore, we have to solve the *reference point problem*

$$(\hat{\mathbf{P}}_z) \qquad \min F_z(u) \quad \text{s.t.} \quad u \in \mathcal{U}_{\mathsf{ad}}$$

which is a scalar-valued minimization problem.

Theorem 3. For any $z \in \mathbb{R}^2$ the reference point problem admits a unique solution $\bar{u}_z \in \mathcal{U}_{\mathsf{ad}}$.

Proof. By Lemma 2 the mapping F_z is convex. Now, the proof is identical to the proof of Theorem 2.14 in Tröltzsch (2010) and uses the strict convexity of F_z along with the fact that $\mathcal{U}_{\mathsf{ad}}$ is bounded and closed in $\mathcal{U}$. □

Theorem 4. Let (4) hold and $\bar{u}_z \in \mathcal{U}_{\mathsf{ad}}$ be an optimal solution to $(\hat{\mathbf{P}}_z)$ for a given $z \in \mathbb{R}^2$. Then $\bar{u}_z$ is Pareto optimal for $(\hat{\mathbf{P}})$.

Proof. We follow along the lines of Theorem 4.20 in Ehrgott (2005): Assume that $\bar{u}_z \in \mathcal{U}_{\mathsf{ad}}$ is not Pareto optimal, then there exists a point $u \in \mathcal{U}_{\mathsf{ad}}$ with $\hat{J}(u) \le \hat{J}(\bar{u}_z)$ and $\hat{J}_j(u) < \hat{J}_j(\bar{u}_z)$ for $j \in \{1,2\}$. Using (4) we get

$$0 < \hat{J}_i(u) - z_i \le \hat{J}_i(\bar{u}_z) - z_i \quad \text{for } i = 1,2$$

and strictly smaller for $i = j$. Together, this yields $F_z(u) < F_z(\bar{u}_z)$ which is a contradiction to the assumption that $\bar{u}_z$ is optimal for $(\hat{\mathbf{P}}_z)$. □

By solving $(\hat{\mathbf{P}}_z)$ consecutively with an adaptive variation of z, we are able to move along the Pareto front in a uniform manner. This way, we get a sequence $\{z^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^2$ of reference points along with optimal controls $\{u^k\}_{k\in\mathbb{N}} \subset \mathcal{U}_{\mathsf{ad}}$ that solve $(\hat{\mathbf{P}}_z)$ with $z = z^k$ as well as $\{\hat{J}^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^2$ with $\hat{J}^k = \hat{J}(u^k)$. To be more precise, the next reference point z^{k+1} is chosen as

$$z^{k+1} = \hat{J}^k + h_J \frac{\hat{J}^k - \hat{J}^{k-1}}{|\hat{J}^k - \hat{J}^{k-1}|} + h_z \frac{\hat{J}^k - z^k}{|\hat{J}^k - z^k|} \text{ for } k \ge 2, \tag{5}$$

where $h_J, h_z \ge 0$ are chosen to control the coarseness of the approximation to the Pareto front. The algorithm is initialized by applying the weighted sum method to $(\hat{\mathbf{P}})$; Zadeh (1963). This yields the first iterates $\hat{J}^1$, $\hat{J}^2 \in \mathcal{P}$. We therefore do not require z^1, z^2 and compute z^3 by setting $h_z = 0$ in (5). Note that the algorithm only moves in one direction: If $\hat{J}_1^1 > \hat{J}_1^2$, then it turns to the upper left in the $\mathbb{R}^2$-plane. Therefore, we perform the algorithm twice, the second time with switched roles of $\hat{J}^1$, $\hat{J}^2$ to cover the other direction as well.

3.2 Optimality conditions

Applying the chain rule, we get for any $u \in \mathcal{U}$

$$\nabla F_z(u) = (\hat{J}_1(u) - z_1)\nabla\hat{J}_1(u) + (\hat{J}_2(u) - z_2)\nabla\hat{J}_2(u). \tag{6}$$

It is well known (Hinze et al. (2009)) that the gradients of $\hat{J}_1$ and $\hat{J}_2$ take the form

$$\big[\nabla \hat{J}_1(u)\big]_i = \int_\Omega \chi_i p(\cdot)\,\mathrm{d}\boldsymbol{x},\ 1 \le i \le m,\ \nabla \hat{J}_2(u) = u, \tag{7}$$

where $p \in \mathfrak{Y}$ is the weak solution to the *adjoint equation*

$$\begin{aligned} -p_t(t,\boldsymbol{x}) &= \Delta p(t,\boldsymbol{x}) && \text{for } (t,\boldsymbol{x}) \in Q,\\ \frac{\partial p}{\partial \boldsymbol{n}}(t,\boldsymbol{x}) &= 0 && \text{for } (t,\boldsymbol{x}) \in \Sigma,\\ p(T,\boldsymbol{x}) &= y(T,\boldsymbol{x}) - y_\Omega(\boldsymbol{x}) && \text{for } \boldsymbol{x} \in \Omega. \end{aligned} \tag{8}$$

In particular, we have

$$\|p\|_{\mathfrak{Y}} \le C\|y(T) - y_\Omega\|_H \tag{9}$$

for a constant $C \ge 0$ which does not depend on u.

We infer from (6)

$$\begin{aligned} \big[\nabla F_z(u)\big]_i = (\hat{J}_1(u) - z_1) \int_\Omega \chi_i p(\cdot)\,\mathrm{d}\boldsymbol{x}&\\ + (\hat{J}_2(u) - z_2)u_i \in \mathfrak{U} \quad \text{for } i = 1,\ldots,m.& \end{aligned}$$

The *first-order necessary optimality condition* for an optimal $\bar{u}_z \in \mathfrak{U}_{\mathsf{ad}}$ now reads as the variational inequality

$$0 \le \langle \nabla F_z(\bar{u}_z), u - \bar{u}_z\rangle_{\mathfrak{U}} \quad \text{for all } u \in \mathfrak{U}_{\mathsf{ad}}. \tag{10}$$

Next, we investigate second-order derivatives: we find

$$\begin{aligned} \nabla^2 F_z(u)v = \sum_{i=1}^{2} \Big((\hat{J}_i(u) - z_i)\nabla^2 \hat{J}_i(u)v&\\ + \langle \nabla \hat{J}_i(u), v\rangle_{\mathfrak{U}} \nabla \hat{J}_i(u)\Big) \text{ for } v \in \mathfrak{U},& \end{aligned}$$

where

$$\big[\nabla^2 \hat{J}_1(u)v\big]_i = \int_\Omega \chi_i q(\cdot)\,\mathrm{d}\boldsymbol{x}, \quad \nabla^2 \hat{J}_2(u)v = v$$

and $q \in \mathfrak{Y}$ solves the *second adjoint equation*

$$\begin{aligned} -q_t(t,\boldsymbol{x}) &= \Delta q(t,\boldsymbol{x}) && \text{for } (t,\boldsymbol{x}) \in Q,\\ \frac{\partial q}{\partial \boldsymbol{n}}(t,\boldsymbol{x}) &= 0 && \text{for } (t,\boldsymbol{x}) \in \Sigma,\\ q(T,\boldsymbol{x}) &= \tilde{y}(T,\boldsymbol{x}) && \text{for } (t,\boldsymbol{x}) \in \Omega \end{aligned}$$

and $\tilde{y} = \mathcal{S}v \in \mathfrak{Y}$ is the solution to (2) for $u = v$ and $y_\circ = 0$. We are interested in whether the second derivative of F_z is coercive. Let $u \in \mathfrak{U}_{\mathsf{ad}}$ and $v \in \mathfrak{U}$:

$$\begin{aligned} &\langle \nabla^2 F_z(u)v, v\rangle_{\mathfrak{U}}\\ &= \sum_{i=1}^{2} (\hat{J}_i(u) - z_i)\langle \nabla^2 \hat{J}_i(u)v, v\rangle_{\mathfrak{U}} + \big|\langle \nabla \hat{J}_i(u), v\rangle_{\mathfrak{U}}\big|^2\\ &\ge (\hat{J}_1(u) - z_1) \int_0^T \int_\Omega \Big(\sum_{k=1}^{m} \chi_k v_k\Big) q \,\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\quad + (\hat{J}_2(u) - z_2)\|v\|_{\mathfrak{U}}^2\\ &= (\hat{J}_1(u) - z_1)\|\tilde{y}(T)\|_H^2 + (\hat{J}_2(u) - z_2)\|v\|_{\mathfrak{U}}^2 \end{aligned}$$

with $\tilde{y} = \mathcal{S}v$. This yields the following result.

Theorem 5. Let (4) hold. For any $u \in \mathfrak{U}_{\mathsf{ad}}$ the hessian $\nabla^2 F_z(u)$ satisfies

$$\langle \nabla^2 F_z(u)v, v\rangle_{\mathfrak{U}} \ge \kappa_z(u)\,\|v\|_{\mathfrak{U}}^2$$

for $\kappa_z(u) = \hat{J}_2(u) - z_2 > 0$. Moreover, if

$$\bar{\kappa}_z = \min\big\{\hat{J}_2(u) - z_2 \,\big|\, u \in \mathfrak{U}_{\mathsf{ad}}\big\} > 0 \tag{11}$$

then $\nabla^2 F_z$ is uniformly positive definite with coercivity constant $\bar{\kappa}_z$.

Remark 6. It is a major advantage for the upcoming error estimation that a coercivity constant for $\nabla^2 F_z$ can be chosen as $\hat{J}_2(u) - z_2$. Usually, the smallest eigenvalue of $\nabla^2 F_z(u)$ has to be computed to gain information about the coercivity constant; Trenz (2016). From a numerical point of view, this is both very expensive and unstable, whereas the value $\kappa_z(u)$ is easily available. ◇

3.3 A-posteriori error estimation

We want to estimate the error $\|\bar{u}_z - \tilde{u}\|_{\mathfrak{U}}$, where $\bar{u}_z \in \mathfrak{U}_{\mathsf{ad}}$ is the (unknown) optimal control of ($\hat{\mathbf{P}}_z$) and $\tilde{u} \in \mathfrak{U}_{\mathsf{ad}}$ is an given admisible control. We follow along the lines of Kammann et al. (2005) and consider the perturbed problem

$$\min F_z(u) + \langle \zeta, u\rangle_{\mathfrak{U}} \quad \text{s.t.} \quad u \in \mathfrak{U}_{\mathsf{ad}} \tag{12}$$

for a fixed function $\zeta \in \mathfrak{U}$. Notice that the functional $F_z + \langle \zeta, \cdot\rangle_{\mathfrak{U}}$ is also strictly convex. Hence, (12) has a unique solution $\bar{u}_\zeta$ for any $\zeta \in \mathfrak{U}$. The first-order sufficient optimality condition for (12) reads

$$\langle \nabla F_z(\bar{u}_\zeta) + \zeta, u - \bar{u}_\zeta\rangle_{\mathfrak{U}} \ge 0 \quad \text{for all } u \in \mathfrak{U}_{\mathsf{ad}}. \tag{13}$$

We will later show that a perturbation $\zeta = \zeta(\tilde{u})$ can be computed numerically such that for a known admissible control $\tilde{u} \in \mathfrak{U}_{\mathsf{ad}}$ the variational inequality (13) holds. This implies that $\tilde{u}$ solves (12).

Suppose that (11) holds. Inserting $\tilde{u}$ in (10) as u and $\bar{u}_z$ in (13) as u yields

$$\begin{aligned} 0 &\le \langle \nabla F_z(\bar{u}_z), \tilde{u} - \bar{u}_z\rangle_{\mathfrak{U}} + \langle \nabla F_z(\tilde{u}) + \zeta, \bar{u}_z - \tilde{u}\rangle_{\mathfrak{U}}\\ &\le -\langle \nabla F_z(\bar{u}_z) - \nabla F_z(\tilde{u}), \bar{u}_z - \tilde{u}\rangle_{\mathfrak{U}} + \|\zeta\|_{\mathfrak{U}}\|\bar{u}_z - \tilde{u}\|_{\mathfrak{U}}. \end{aligned}$$

Using the mean value theorem for the Fréchet-differentiable function ∇F_z yields the existence of a $\hat{u} \in \mathfrak{U}$ with

$$\begin{aligned} 0 &\le -\langle \nabla^2 F_z(\hat{u})(\bar{u}_z - \tilde{u}), \bar{u}_z - \tilde{u}\rangle_{\mathfrak{U}} + \|\zeta\|_{\mathfrak{U}}\|\bar{u}_z - \tilde{u}\|_{\mathfrak{U}}\\ &\le -\bar{\kappa}_z\|\bar{u}_z - \tilde{u}\|_{\mathfrak{U}}^2 + \|\zeta\|_{\mathfrak{U}}\|\bar{u}_z - \tilde{u}\|_{\mathfrak{U}}. \end{aligned}$$

Summarizing, we deduce the rigorous estimate

$$\|\bar{u}_z - \tilde{u}\|_{\mathfrak{U}} \le \frac{1}{\bar{\kappa}_z}\|\zeta\|_{\mathfrak{U}}.$$

We still have to identify the function $\zeta \in \mathfrak{U}$. Exactly as in Kammann et al. (2005), we use

$$\zeta_i(t) = \begin{cases} [(\nabla F_z(\tilde{u}))_i(t)]_- & \text{if } \tilde{u}_i(t) = (u_a)_i(t),\\ -[(\nabla F_z(\tilde{u}))_i(t)]_+ & \text{if } \tilde{u}_i(t) = (u_b)_i(t),\\ -(\nabla F_z(\tilde{u}))_i(t) & \text{otherwise}, \end{cases} \tag{14}$$

where we have used the decomposition for a real number $\xi \in \mathbb{R}$ as $\xi = [\xi]_+ - [\xi]_-$ with $[\xi]_+ = \max(0,\xi)$ and $[\xi]_- = -\min(0,\xi)$. Thus, we have proved the next theorem.

Theorem 7. Assume that (11) holds and $\bar{u}_z \in \mathfrak{U}_{\mathsf{ad}}$ is the unique solution to ($\hat{\mathbf{P}}_z$). For an arbitrary control $\tilde{u} \in \mathfrak{U}_{\mathsf{ad}}$ let the function $\zeta \in \mathfrak{U}$ be defined as in (14). Then we can estimate the error by

$$\|\bar{u}_z - \tilde{u}\|_{\mathfrak{U}} \le \Delta(\tilde{u}) \quad \text{with } \Delta(\tilde{u}) = \frac{1}{\bar{\kappa}_z}\,\|\zeta\|_{\mathfrak{U}}. \tag{15}$$

3.4 Continuity of the mapping $z \mapsto \bar{u}_z$

An important aspect is how the optimal control $\bar{u}_z$ adapts to the choice of the reference point.

Theorem 8. Suppose that (11) holds. Let $\bar{u}_z$ be the optimal solution to $(\hat{\mathbf{P}}_z)$ for a given feasible reference point $z \in \mathcal{Z}$ with

$\mathcal{Z} = \big\{ \tilde{z} \in \mathbb{R}^2 \big| \, \tilde{z}_1 < \hat{J}_1(u), \, \tilde{z}_2 < \hat{J}_2(u) - \bar{\kappa} \text{ for all } u \in \mathcal{U}_{\mathsf{ad}} \big\}.$

Then the mapping $z \mapsto \bar{u}_z$ is continuous from $\mathcal{Z}$ to $\mathcal{U}_{\mathsf{ad}}$.

Proof. Let $z_1 = (z_{11}, z_{12})$, $z_2 = (z_{21}, z_{22}) \in \mathcal{Z}$ be chosen arbitrarily with according optimal controls $\bar{u}_1 = \bar{u}_{z_1}$, $\bar{u}_2 = \bar{u}_{z_2}$. By using (10) we get

$$\begin{aligned} 0 &\le \langle \nabla F_{z_1}(\bar{u}_1), \bar{u}_2 - \bar{u}_1 \rangle_{\mathcal{U}} + \langle \nabla F_{z_2}(\bar{u}_2), \bar{u}_1 - \bar{u}_2 \rangle_{\mathcal{U}} \\ &= -\langle \nabla F_{z_1}(\bar{u}_1) - \nabla F_{z_1}(\bar{u}_2), \bar{u}_1 - \bar{u}_2 \rangle_{\mathcal{U}} \\ &\quad + \langle \nabla F_{z_2}(\bar{u}_2) - \nabla F_{z_1}(\bar{u}_2), \bar{u}_1 - \bar{u}_2 \rangle_{\mathcal{U}}. \end{aligned} \tag{16}$$

Arguing as in the proof of Theorem 7 the first term is bounded by $-\bar{\kappa}\|\bar{u}_1 - \bar{u}_2\|^2_{\mathcal{U}}$. Hence, (16) implies that

$$\bar{\kappa}\, \|\bar{u}_1 - \bar{u}_2\|^2_{\mathcal{U}} \le \langle \nabla F_{z_2}(\bar{u}_2) - \nabla F_{z_1}(\bar{u}_2), \bar{u}_1 - \bar{u}_2 \rangle_{\mathcal{U}}. \tag{17}$$

From (6) we find that $\nabla F_z(u)$ depends linearly on z and

$$\begin{aligned} &\langle \nabla F_{z_2}(\bar{u}_2) - \nabla F_{z_1}(\bar{u}_2), \bar{u}_1 - \bar{u}_2 \rangle_{\mathcal{U}} \\ &= \sum_{i=1}^{2} (z_{1i} - z_{2i}) \, \langle \nabla \hat{J}_i(\bar{u}_2), \bar{u}_1 - \bar{u}_2 \rangle_{\mathcal{U}} \\ &\le \big| z_1 - z_2 \big| \Big(\sum_{i=1}^{2} \langle \nabla \hat{J}_i(\bar{u}_2), \bar{u}_1 - \bar{u}_2 \rangle^2_{\mathcal{U}} \Big)^{1/2} \\ &\le C(z_2) \big| z_1 - z_2 \big| \|\bar{u}_1 - \bar{u}_2\|_{\mathcal{U}}, \end{aligned}$$

where $C(z_2) = (\sum_{i=1}^{2} \|\nabla \hat{J}_i(\bar{u}_2)\|^2_{\mathcal{U}})^{1/2}$ is independent of $\bar{u}_2$ and therefore independent of z_2. We derive from (17):

$$\|\bar{u}_1 - \bar{u}_2\|_{\mathcal{U}} \le \frac{C(z_2)}{\bar{\kappa}} \big| z_1 - z_2 \big|.$$

Fixing z_2 and letting $z_1 \to z_2$ in $\mathbb{R}^2$ now yields $\bar{u}_1 \to \bar{u}_2$ in $\mathcal{U}$ which proves the claim. □

Remark 9. The set $\mathcal{U}_{\mathsf{ad}}$ is bounded. Combining (3) and (9) it follows that the weak solution p to (8) is bounded in $\mathcal{Y}$ by a constant independent of $u \in \mathcal{U}_{\mathsf{ad}}$. Consequently, the norms $\|\nabla \hat{J}_1(u)\|_{\mathcal{U}}$ and $\|\nabla \hat{J}_2(u)\|_{\mathcal{U}}$ are bounded by a constant independent of $u \in \mathcal{U}_{\mathsf{ad}}$. Hence the constant $C(z_2)$ in the proof of Theorem 8 is bounded from above by a constant independent of $u \in \mathcal{U}_{\mathsf{ad}}$. This implies that the mapping $z \mapsto \bar{u}_z$ is even Lipschitz-continuous. ◇

4. REDUCED-ORDER MODELING (ROM) BY POD

The previously described reference point algorithm makes it necessary to repeatedly solve $(\hat{\mathbf{P}}_z)$ which ultimately goes back to computing the state equation (1) and its adjoint equation (8) for many different instances. This multi-query context makes model-order reduction techniques conceivable; we focus in particular on proper orthogonal decomposition (POD); see, e.g., Holmes et al. (2012): After having computed the first control u^1 by means of a weighted sum problem, a finite-dimensional subspace

$$V^\ell = \operatorname{span}\{\psi_1, \ldots, \psi_\ell\} \subset V$$

is created such that the projected trajectory of $y(u^1)$ has a least-squares deviation from its full-dimensional version at given time instances $0 = t_1 < t_2 < \ldots < t_n = T$. Let X denote either the space H or V. Then we consider the POD cost function $\mathcal{I} : V^\ell \to \mathbb{R}$ which is given by

$$\mathcal{I}_n(\psi_1, ..., \psi_\ell) = \sum_{j=1}^{n} \alpha_j \, \Big\| y(t_j) - \sum_{i=1}^{\ell} \langle y(t_j), \psi_i \rangle_X \, \psi_i \Big\|^2_X$$

with positive (trapezoidal) weights α_j. The POD optimization problem then reads

$$\min \mathcal{I}_n(\psi_1, ...\psi_\ell) \text{ s.t. } \langle \psi_i, \psi_j \rangle_X = \delta_{ij} \; (i,j = 1, ..., \ell). \tag{18}$$

The solution to (18) is well-known by the singular value decomposition of a certain operator, compare e.g. Kammann et al. (2005).

Now suppose that we have computed a POD basis $\{\psi_i\}_{i=1}^{\ell} \subset X$ of rank $\ell \le n$. The POD Galerkin solution

$$y^\ell(t) = \sum_{i=1}^{\ell} a_i^\ell(t) \psi_i \text{ for } t \in [0,T], \; a^\ell : [0,T] \to \mathbb{R}^\ell \tag{19}$$

satisfies the following POD Galerkin scheme

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} \, \langle y^\ell(t), \psi_j \rangle_H + \int_\Omega \nabla y^\ell(t) \cdot \nabla \psi_j \, \mathrm{d}\boldsymbol{x} \\ &\qquad = \sum_{i=1}^{m} u_i(t) \, \langle \chi_i, \psi_j \rangle_H, \quad t \in (0,T), \\ &\langle y^\ell(0), \psi_j \rangle_H = \langle y_\circ, \psi_j \rangle_H \end{aligned} \tag{20}$$

for $j = 1, ..., \ell$ and $t \in (0,T)$. It is known that (20) admits a unique solution $y^\ell = y^\ell(u) \in H^1(0,T;V^\ell) \subset \mathcal{Y}$; see, e.g., Gubisch and Volkwein (2016).

Inserting (19) into (20) we get the following system in $\mathbb{R}^\ell$:

$$M\dot{a}^\ell(t) + Sa^\ell(t) = Bu(t), \quad a^\ell(0) = a_\circ$$

with the mass matrix $M \in \mathbb{R}^{\ell \times \ell}$, the stiffness matrix $S \in \mathbb{R}^{\ell \times \ell}$ and the control operator $B \in \mathbb{R}^{\ell \times m}$ which are given by

$$\begin{aligned} M_{ij} &= \langle \psi_i, \psi_j \rangle_H && \text{for } i,j = 1, ..., \ell, \\ S_{ij} &= \int_\Omega \nabla \psi_i \cdot \nabla \psi_j \, \mathrm{d}\boldsymbol{x} && \text{for } i,j = 1, ..., \ell, \\ B_{ij} &= \langle \psi_j, \chi_i \rangle_H && \text{for } i = 1, ..., m, \; j = 1, ..., \ell. \end{aligned}$$

The POD Galerkin approximation to $(\mathbf{P})$ is given by the minimization problem

$$(\hat{\mathbf{P}}^\ell) \qquad \min \hat{J}^\ell(u) \quad \text{s.t.} \quad u \in \mathcal{U}_{\mathsf{ad}},$$

where we set $\hat{J}^\ell(u) = J(y^\ell(u), u)$ and $y^\ell(u)$ denotes the solution to (20) for $u \in \mathcal{U}_{\mathsf{ad}}$.

To solve $(\hat{\mathbf{P}}^\ell)$ we apply the reference point method utilizing the corresponding distance function

$$F_z^\ell(u) = \frac{1}{2} \, |\hat{J}^\ell(u) - z|^2.$$

with a reference point $z \in \mathbb{R}^2$. Thereby, we obtain a POD suboptimal control $\bar{u}_z^\ell \in \mathcal{U}_{\mathsf{ad}}$ for any z. The resulting error is then estimated using (15) with $\tilde{u} = \bar{u}_z^\ell$. This indicates the quality of the current POD basis, which can then be recomputed if necessary.

5. NUMERICAL EXPERIMENTS

5.1 Setting

All computations were carried out on a standard PC with Ubuntu 14.04 LTS, Intel(R) Core i7-4600U CPU @ 2.10GHz x4, 11.4 GiB RAM.

For the state equation, we choose a two-dimensional domain $\Omega = (0,1)^2$ and $m = 9$ shape functions which are indicator functions representing a uniform partition of the domain. The initial condition is $y_\circ \equiv 0$ and we chose

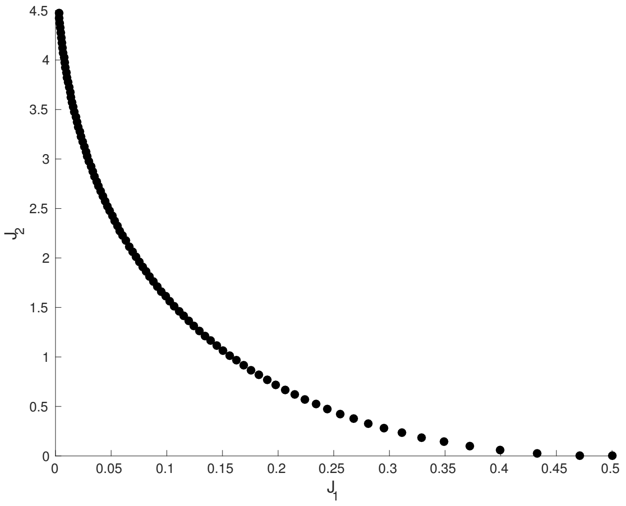

Fig. 1. Discrete Pareto front consisting of 64 points computed using POD-ROM with $\ell = 2$.

$T = 1$. The bilateral constraints were set to $u_a \equiv 0$ and $u_b \equiv 1$. For the cost functional J, the terminal condition is $y_\Omega \equiv 1$.

Problem $(\hat{\mathbf{P}}_z)$ was solved using a projected Newton-CG method; see Kelley (1999). For the coefficients in (5), two different settings were used as will be explained later in detail. The discretization was done using linear finite elements (FE) on a grid with $N_x = 729$ nodes. The time interval was discretized uniformly using $N_t = 100$ time instances. The time discretization of the state and adjoint equation were done by an implicit Euler scheme. For the POD basis computation we utilize $X = V$.

5.2 Results

First results can be seen in Figure 1, where the Pareto front was computed using the POD-ROM ansatz described in Section 4 with $\ell = 2$ basis functions.

One can observe that the front becomes coarser towards the lower right. This is due to the fact that in this area, $\hat{J}_2^\ell$ is relatively small which corresponds to little to no control effort. For these objective values, a very small increment in the control variable can yield a large improvement towards the desired state. However, this is due to the nature of the problem and is also observed using different multiobjective optimization techniques, like, e.g., the weighted sum method. Apart from this, the results are quite satisfactory: We have achieved a rather uniform discretization of the Pareto front which can be presented to the decision maker.

In Figures 2 and 3, the corresponding reference points can be seen. Note that the axis are not scaled equally so the pictures look distorted. We have utilized two different strategies for choosing the parameters in (5), one resulting in a very tight neighbourhood of the front, the other one in a curve further apart. This has consequences which will become apparent when considering the error estimation.

Recall that in the estimation for the real error (15), the term $\bar{\kappa}_z$ appears which is given by the condition (11). As

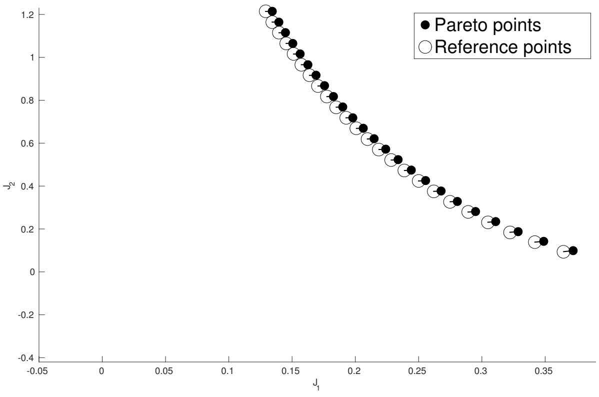

Fig. 2. Part of the Pareto front with according reference points, chosen close to the front itself.

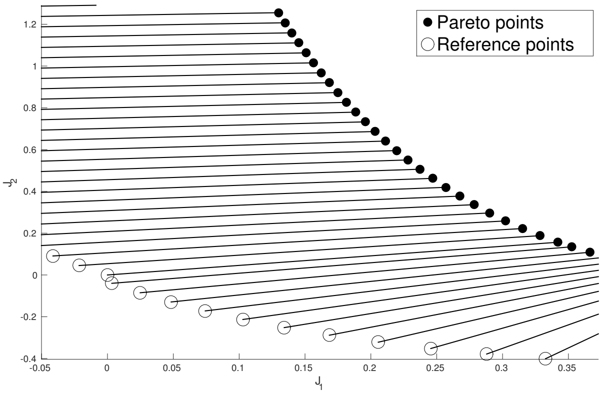

Fig. 3. Part of the Pareto front with according reference points, chosen far from the front itself.

it turns out, this condition is too strong for our numerical purposes. Instead, we only require that the inequality holds for all u in a neighbourhood of the optimal control $\bar{u}$. As an approximation, we therefore compute

$$\kappa_z(\bar{u}_z^\ell) = \hat{J}_2^\ell(\bar{u}_z^\ell) - z_2,$$

where $\bar{u}_z^\ell$ is the suboptimal control solving $(\hat{\mathbf{P}}^\ell)$, and use this value instead of $\bar{\kappa}_z$. Furthermore, an additional termination check is implemented to terminate the multiobjective optimization if $\kappa_z(\bar{u}_z^\ell)$ falls below a certain tolerance $0 < \varepsilon_\kappa \ll 1$.

By (11), if the reference points are further apart from the Pareto front, $\bar{\kappa}_z$ should be larger which would correspond to a tighter estimator $\Delta(\tilde{u})$ in (15) for the choice $\tilde{u} = \bar{u}_z^\ell$. We have analyzed the behavior of the error estimator along the multiobjective optimization iteration: For each reference point $z \in \mathbb{R}^2$, we obtain a POD-based suboptimal control $\bar{u}_z^\ell$ which is an approximation to the (unknown) optimal control $\bar{u}_z$ that is approximated by the FE control denoted by $\bar{u}_z^h$. In Figures 4 and 5, we observe the behavior of the true error $\|\bar{u}_z^h - \bar{u}_z^\ell\|_\mathcal{U}$ as well as the error estimator $\Delta(\bar{u}_z^\ell)$ in the course of the multiobjective optimization, both for close and far reference points.

It can be seen that indeed, we are able to achieve a much tighter estimate for the choice of distant reference points. Nevertheless, the estimator can overshoot the real error

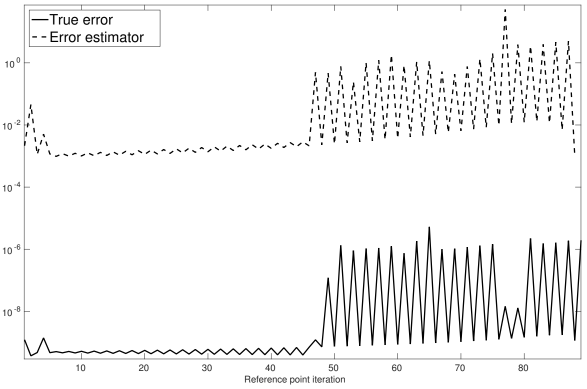

Fig. 4. True error $\|\bar{u}_z^h - \bar{u}_z^\ell\|_{\mathcal{U}}$ and error estimator $\Delta(\bar{u}_z^\ell)$ for $\ell = 2$, using reference points z close to the Pareto front.

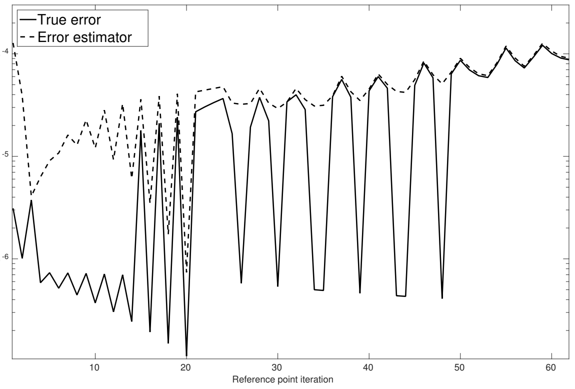

Fig. 5. True error $\|\bar{u}_z^h - \bar{u}_z^\ell\|_{\mathcal{U}}$ and error estimator $\Delta(\bar{u}_z^\ell)$ for $\ell = 2$, using reference points z far from the Pareto front

for factors of up to 10^3. This can be explained by the fact that we have used some rather rough estimates to get to the bound $\bar{\kappa}_z$ in (15). The neglected terms then explain the observed discrepancies.

The zigzag behavior of the error estimates can be easily explained since in the implementation, we do not cover both directions in the objective space consecutively. Instead, we alternate between the two directions. We therefore alternate between different regions of the Pareto front where different properties of the reference point problem prevail.

6. CONCLUSION

In the present paper it is illustrated that POD reduced-order strategies can be efficiently combined with the reference point method in order to solve bicriterial optimization problems. A rigorous a-posteriori error estimator allows us to control the POD error in order to ensure a desired tolerance. As a next step we plan to involve also convective terms in the state equation in order to model a heat convection in a building.

REFERENCES

R. Dautray and J.-L. Lions: *Mathematical Analysis and Numerical Methods for Science and Technology. Volume 5: Evolution Problems I.* Springer-Verlag, Berlin, 2000.

M. Ehrgott: *Multicriteria Optimization.* Springer, Berlin, 2005.

K. F. Fong, V. I. Hanby, and T.-T. Chow: HVAC system optimization for energy management by evolutionary programming. *Energy and Buildings*, 38:220-231, 2006.

M. Gubisch and S. Volkwein: Proper orthogonal decomposition for linear-quadratic optimal control. To appear in P. Benner, A. Cohen, M. Ohlberger, and K. Willcox (eds.), Model Reduction and Approximation: Theory and Algorithms. *SIAM*, Philadelphia, PA, 2016.

M. Hinze, R. Pinnau, M. Ulbrich and S. Ulbrich: *Optimization with PDE Constraints.* Springer, 2009.

P. Holmes, J.L. Lumley, G. Berkooz, and C.W. Rowley: *Turbulence, Coherent Structures, Dynamical Systems and Symmetry.* Cambridge Monographs on Mechanics. Cambridge University Press, Cambridge, 2nd ed. edition, 2012.

L. Iapichino, S. Trenz and S. Volkwein: Reduced-order multiobjective optimal control of semilinear parabolic problems. http://nbn-resolving.de/urn:nbn:de:bsz:352-0-313874, 2015.

L. Iapichino, S. Ulbrich and S. Volkwein: Multiobjective PDE-constrained optimization using the reduced-basis method. http://nbn-resolving.de/urn:nbn:de:bsz:352-250190, 2013.

E. Kammann, F. Tröltzsch and S. Volkwein: A posteriori error estimation for semilinear parabolic optimal control problems with application to model reduction by POD. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47:555-581, 2013.

C.T. Kelley: *Iterative Methods for Optimization*, SIAM Frontiers in Applied Mathematics, no 18, 1999.

A. Kusiak, F. Tang and G. Xu: Multi-objective optimization of HVAC system with an evolutionary computation algorithm. *Energy*, 36:2440-2449, 2011.

K. Miettinen: *Nonlinear Multiobjective Optimization.* International Series in Operations Research & Management Science, Springer, 1998.

S. Peitz, S. Oder-Blöbaum and M. Dellnitz: Multiobjective optimal control methods for fluid flow using reduced order modeling. http://arxiv.org/pdf/1510.05819v2.pdf, 2015.

C. Romaus, J. Böcker, K. Witting, A. Seifried, and O. Znamenshchykov: Optimal energy management for a hybrid energy storage system combining batteries and double layer capacitors. *IEEE*, pages 1640-1647, San Jose, CA, USA, 2009.

W. Stadler: *Multicriteria Optimization in Engineering and in the Sciences.* Plenum Press, New York, 1988.

S. Trenz: *A-posteriori Error Estimates for Reduced Order Models in Nonlinear Optimization Problems.* Ph.D. thesis, University of Konstanz, 2016.

F. Tröltzsch: *Optimal Control of Partial Differential Equations. Theory, Methods and Applications.* American Math. Society, Providence, volume 112, 2010.

L. Zadeh. Optimality and non-scalar-valued performance criteria. *IEEE Transactions on Automatic Control*, 8, 1963.