Mathematical Programming
https://doi.org/10.1007/s10107-024-02181-1

FULL LENGTH PAPER

Series B

McCormick envelopes in mixed-integer PDE-constrained optimization

Sven Leyffer[1] · **Paul Manns**[2]

Received: 29 March 2024 / Accepted: 6 December 2024

Abstract

McCormick envelopes are a standard tool for deriving convex relaxations of optimization problems that involve polynomial terms. Such McCormick relaxations provide lower bounds, for example, in branch-and-bound procedures for mixed-integer nonlinear programs but have not gained much attention in PDE-constrained optimization so far. This lack of attention may be due to the distributed nature of such problems, which on the one hand leads to infinitely many linear constraints (generally state constraints that may be difficult to handle) in addition to the state equation for a pointwise formulation of the McCormick envelopes and renders bound-tightening procedures that successively improve the resulting convex relaxations computationally intractable. We analyze McCormick envelopes for a model problem class that is governed by a semilinear PDE involving a bilinearity and integrality constraints. We approximate the nonlinearity and in turn the McCormick envelopes by averaging the involved terms over the cells of a partition of the computational domain on which the PDE is defined. This yields convex relaxations that underestimate the original problem up to an a priori error estimate that depends on the mesh size of the discretization. These approximate McCormick relaxations can be improved by means of an optimization-based bound-tightening procedure. We show that their minimizers converge to minimizers to a limit problem with a pointwise formulation of the McCormick envelopes when driving the mesh size to zero. We provide a computational example, for which we certify all of our imposed assumptions. The results point to both the potential of the methodology and the gaps in the research that need to be closed. Our methodology provides a framework first for obtaining pointwise underestimators for nonconvexities and second for approximating them with finitely many linear inequalities in an infinite-dimensional setting.

Mathematics Subject Classification 49M20 · 90C26

✉ Paul Manns
paul.manns@tu-dortmund.de

Sven Leyffer
leyffer@anl.gov

[1] Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL 60439, USA

[2] Faculty of Mathematics, TU Dortmund University, 44227 Dortmund, Germany

Published online: 06 January 2025

1 Introduction

We are interested in the global optimization of mixed-integer PDE-constrained optimization problems (MIPDECOs) that feature nonconvex terms. MIPDECOs arise in many real-world applications such as topology optimization [4, 25, 30] and supply network optimization [20, 41]. A prototypical problem class is

$$\begin{aligned} \min_{u,w}\ & j(u,w) + \alpha R(w) \\ \text{s.t. } & Au + N(u,w) = 0 \\ & w \in C \\ & w(x) \in \mathbb{Z} \text{ for almost every (a.e.) } x \in \Omega, \end{aligned} \tag{MIPDECO}$$

where $w : \Omega \to \mathbb{R}$ is a measurable input function (generally also called *control*) on a computational domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$. The function w needs to lie in a convex set C, for example, satisfy bound constraints or volume restrictions in topology optimization, and additionally satisfies the nonconvex pointwise integrality restriction $w(x) \in \mathbb{Z}$ in Ω. The control function w enters the problem as an input of the state equation (PDE) $Au + N(u, w) = 0$, which we assume to have a unique solution u for a given w. The operator A could be an elliptic differential operator that includes appropriate boundary conditions and $N(u, w)$ a term with a lower differentiability index on u than Au that may involve a nonlinearity. The pair of w and the implied solution u to the state equation will minimize an objective term j, which is often a quadratic fidelity term on u. Moreover, desired structures on w can be promoted by means of the regularization term $\alpha R(w)$ for some $\alpha > 0$. A formal setting is provided in Sect. 2.1.

A beneficial choice for R in the context of mixed-integer PDE-constrained optimization has been the regularization $R = \mathrm{TV}$, where TV denotes the total variation seminorm on the space of integrable functions [1]. Such a regularization may be necessary to guarantee the existence of solutions [29, 34]. The total variation of an integer-valued function is the $(d-1)$-dimensional volume of the interfaces between its level sets weighted by the jump heights over the respective interfaces. This choice implies that instances of (MIPDECO) admit solutions under mild assumptions; see, for example, [11, 36]. Recent work on solving such problems to the satisfaction of the stationarity concept has been studied in [29, 33, 36, 55]. The computation of lower bounds for a linear PDE, thereby yielding a convex set after omitting the integrality constraints, and their integration into a branch-and-bound algorithm have recently been studied by Buchheim et al. in [6–10] for $d = 1$.

In this work we strive to obtain convex relaxations and thus lower bounds on (MIPDECO) in the presence of a structured nonlinearity of the feasible set induced by a specific choice of N. To be precise, we restrict our work to the case that A is an elliptic operator, $R = \mathrm{TV}$, and the nonlinearity N in the state equation is in fact a bilinearity, specifically

$$N(u, w)(x) := u(x)w(x) \text{ for a.e. } x \in \Omega.$$

We replace the nonlinear state equation by a linear state equation and linear inequalities based on McCormick envelopes that have been introduced for finite-dimensional nonconvex optimization problems in [38]. Our aim is to obtain valid and (ideally) tight lower bounds for instances of (MIPDECO). McCormick envelopes and relaxations have already been studied in the context of global optimization with ODEs finite-dimensional control inputs in [39, 47, 48, 52–54, 56, 61, 62]. These articles develop and analyze arguments to show the existence and computation of solutions to ODEs whose solution trajectories are convex and concave lower and upper bounds on the solution trajectories of an original ODE that is influenced by means of a finite-dimensional parameter. In the same spirit such pointwise bounds on the resulting trajectory are transferred to a class of parabolic PDEs in [2]. In the context of infinite-dimensional control inputs for ODE-constrained optimization, the use of McCormick envelopes was proposed in [27] and [51]. The latter generalizes ideas from the aforementioned literature to functions varying pointwise in time. The authors obtain the existence of convex and concave trajectories that over- and underestimate the trajectories for the parameterized ODE when over- and underestimations exist pointwise a.e.

We consider a setting, where the existence of such envelopes is straightforward and provide a possibility to on the one hand only have an approximate relaxation on the original problem but which on the other hand allows for a way to accelerate an optimization-based bound-tightening procedure in order to improve the approximate relaxations. Generally, such a bound-tightening procedure can become very compute-intensive, it is called *one of the most expensive bound tightening procedures* in [21]. Moreover, state-constrained optimal control problems, in particular with infinitely many state constraints, are notoriously hard to analyze and solve.

Therefore, we introduce an approximation scheme that allows us to work with a finite number of linear inequalities and variables that is substantially smaller than the number of variables used for the discretization of the PDE. To this end, we partition the computational domain into finitely many grid cells and replace the product uw by a product of local averages $(P_h u)(P_h w)$, where P_h denotes the local averaging defined in (1) below. This enables us to derive approximate McCormick relaxations that can be described by finitely many linear inequalities. Using beneficial regularity or continuity properties of the solution to the underlying PDE, specifically that its solution attains relatively close values in grid cells that are spatially close to each other, we expect that a moderate mesh size may already yield a good approximation of the lower bound. In this way we strive for a much faster computation in order to obtain approximate lower bounds on (MIPDECO), which are required for our goal of solving instances of (MIPDECO) to global optimality. We highlight that the computation of such lower bounds is also interesting for global optimization in its own right even without the context of integer-valued input functions.

We believe that our restrictions on the setting are sensible. First, $R = \mathrm{TV}$ is a useful regularization for MIPDECOs because it results in crisp designs (level sets of the input function). Moreover, the analytical properties of the TV-regularization term also help streamline our analysis although the general strategy can also be carried out for many other choices of R. Similarly, higher-order multilinear terms for $N(u, w)$ involving u, for example, $u^k w$, $k \geq 2$, would make the PDE analysis, namely, the existence of solutions and their regularity, considerably more complicated. In this regard many of

our arguments are streamlined and do not distract from the analysis and approximation of the McCormick relaxation. Consequently, this article will be read as a recipe to first obtain a pointwise generalization of underestimators for nonconvexities as are known from finite-dimensional nonconvex optimization and then approximate this pointwise generalization with finitely many linear inequalities using approximation properties that are due to the PDE setting.

Contribution

We provide a formal definition of an optimal control problem with the aforementioned features and derive McCormick envelopes using pointwise constraints. We then introduce a grid that discretizes the computational domain Ω, and we approximate the inequalities defining the McCormick envelope by a local averaging over the grid cells. In a second step, we discretize the control input function on the same grid, thereby reducing the complexity of the problem further. We prove the existence of solutions to the approximate convex relaxations and an estimate on the lower bound for the optimal control problem depending on the mesh size of the grid. Moreover, we prove that minimizers of the approximate convex relaxations converge to minimizers of the convex relaxation obtained by imposing the McCormick envelopes by means of pointwise constraints when driving the mesh size to zero.

We introduce and analyze the aforementioned optimization-based bound-tightening procedure that allows us to tighten given bounds on the state variable u and in turn to increase the objective value of the (approximate) McCormick relaxation and thus to tighten the induced lower bounds on the optimal objective value of the optimal control problem.

We then verify all assumptions imposed on our analysis for an exemplary optimal control problem that is governed by an elliptic PDE that is defined on $\Omega \subset \mathbb{R}$. In particular, this example allows us to provide tight values for all constants in the estimates we impose. We then give details on how we discretize the PDE and set up a computational experiment where we insert the aforementioned constants. We compute and compare the approximate lower bounds with and without the bound-tightening procedure applied with each other to upper bounds with and without an integrality restriction on the controls in order to assess the quality of the analyzed methodology. We also provide computational results for a second example that is defined on a two-dimensional domain. The observations confirm those of the one-dimensional example.

Structure of the remainder

We continue with a brief introduction to our notation. In Sect. 2 we introduce McCormick envelopes for optimal control problems with semilinear state equations that feature a bilinear term. The McCormick envelopes are first introduced and analyzed in a pointwise fashion. Then a local averaging is introduced in to approximate the pointwise inequalities by finitely many linear inequalities. Then employ a piecewise constant control function ansatz, as is desired in our motivating application of integer

optimal control, to obtain a further approximation with the state variable u still remaining in function space. In Sect. 3 we introduce an elliptic PDE that we use as a showcase to verify all of the assumptions that are imposed for the well-definedness and approximation properties of the different McCormick settings from Sect. 2. We describe the finite-element discretization of the state equation for our guiding example, implement the bound-tightening procedure, and perform our computational experiments in Sects. 4 (1D) and 5 (2D). We draw a conclusion in Sect. 6.

Notation

Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a bounded domain. The projection in the space of integrable functions $L^1(\Omega)$ to piecewise constant functions on a partition $\mathcal{Q}_h = \{Q_h^1, \ldots, Q_h^{N_h}\}$ of Ω is denoted by P_h; specifically

$$P_h : L^1(\Omega) \ni f \mapsto \sum_{i=1}^{N_h} \frac{1}{|Q_h^i|} \int_{Q_h^i} f(x)\,\mathrm{d}\,x\,\chi_{Q_h^i} \in L^1(\Omega). \tag{1}$$

For a measurable set $A \subset \mathbb{R}^d$, $d \in \mathbb{N}$, we denote its d-dimensional Lebesgue measure by $|A|$. We say that a function $w \in L^1(\Omega)$ is of bounded variation if its total variation seminorm $\mathrm{TV}(w)$ is finite. The Banach space of functions in $L^1(\Omega)$ of bounded variation with norm $\|\cdot\|_{\mathrm{BV}} = \|\cdot\|_{L^1} + \mathrm{TV}(\cdot)$ is then denoted by $\mathrm{BV}(\Omega)$; see [1] for details on functions of bounded variation. We recall that the total variation seminorm for $w \in L^1(\Omega)$ is defined as

$$\mathrm{TV}(w) \coloneqq \sup \left\{ \int_\Omega w(x)\,\mathrm{div}\,\phi(x)\,\mathrm{d}\,x \;\middle|\; \phi \in C_c^1(\Omega, \mathbb{R}^d),\ \max_{x\in\Omega} \|\phi(x)\| \le 1 \right\},$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. On one-dimensional domains $\Omega = (a, b)$ for $a, b \in \mathbb{R}$, every element of $\mathrm{BV}(\Omega)$ has a left-continuous representative, and the total variation is the sum of the jump heights of the left-continuous representative (this interpretation also works for the right-continuous representative or other so-called good representatives; see [1]).

2 McCormick envelopes for optimal control problems with state equations that have bilinear terms

In this section we derive and analyze McCormick envelopes and relaxations for a class of nonconvex optimal control problems where the nonconvexity stems from bilinear terms in the state equation. We first introduce a prototypical optimal control problem in Sect. 2.1. We then derive pointwise a.e. McCormick envelopes in Sect. 2.2 and, subsequently, provide a local averaging in two stages that yields approximate lower bounds but implies problems, still in function space, which feature only finitely many additional linear inequalities in Sect. 2.3. In Sect. 2.4 we prove the validity of

a bound-tightening procedure that may improve (increase) the (approximate) lower bounds.

2.1 Optimal control problem with nonconvex bilinearity

We consider the following prototypical setting with a nonconvexity induced by a bilinear term in the state equation, but note that the strategy we develop can be transferred to many other settings with lower or slightly different regularity assumptions with a small amount of modifications.

Let Ω be a bounded domain. Let U be a reflexive Banach space that is compactly embeded into $H := L^2(\Omega)$, that is, $U \hookrightarrow^c H$, so that we may work with the triple $U \hookrightarrow^c H \cong H^* \hookrightarrow^c U^*$ of compact embeddings and the isometric isomorphic identification $H \cong H^*$. In optimal control terminology, U is the *state space*. For the *control space*, we consider the space $W := \mathrm{BV}(\Omega)$ of functions of bounded variation, which is motivated by the compactness and approximation properties it provides in the control space and our intended application in the context of integer optimal control. Specifically, $\mathrm{BV}(\Omega)$-regularity allows us to approximate a function by its average on a partition of the domain with an error that is bounded by a scalar multiple of the mesh size. Note that other spaces like the Sobolev space $H^1(\Omega)$ also provide this property.

The optimal control problem is

$$\begin{aligned} \min_{u,w}\ & j(u,w) + \alpha\,\mathrm{TV}(w) \\ \text{s.t. } & Au + uw = f \quad \text{in } U^*, \\ & w \in C,\, u \in U \end{aligned} \tag{OCP}$$

for a nonempty, bounded, closed, and convex set $C \subset H$, an objective $j : U \times H \to \mathbb{R}$ that is Lipschitz continuous on bounded sets and a linear and bounded differential operator $A : U \to U^*$ with bounded inverse $A^{-1} : U^* \to U$. The nonconvexity of the problem stems from the bilinear term uw in the state equation.

Note that the assumptions on C imply that it is a weakly sequentially compact subset of H so that we obtain the regularity $w \in H$ for all feasible w. Together with the TV-seminorm in the objective, we obtain the regularity $w \in W$ for all feasible w with finite objective value and the assumed structure $j : U \times H \to \mathbb{R}$ is well-defined. Moreover, we note that the continuous embedding $W \hookrightarrow H$ holds for $d \in \{1, 2\}$ so that we can equivalently consider W or $W \cap H$ as control space. For $d \geq 3$, defining $W \cap H$ as control space here is slightly more specific but not necessary for our analysis and would complicate our notation later. The vector $f \in U^*$ is a fixed datum that parameterizes the optimization problem, and we assume that the product uw is a measurable function that is (also) an element of U^* for all tuples $(u, w) \in U \times W$. This follows, for example, if C is bounded in $L^\infty(\Omega)$, too. The term $\mathrm{TV}(w)$ denotes the total variation seminorm of w that enforces its required $\mathrm{BV}(\Omega)$-regularity, and $\alpha > 0$ is a positive scalar. Assuming that the state equation and the original optimal control problem admit unique solutions, the presence of the term uw

implies a nonlinear dependence of the solution to the state equation on w, which in turn implies that (OCP) is generally nonconvex even if j is convex.

2.2 Pointwise McCormick envelopes

McCormick envelopes have been introduced for finite-dimensional nonconvex optimization problems in [38]. The key idea for the case of (OCP) is to derive a new optimal control problem, the so-called McCormick envelope, by relaxing the state constraints. Specifically, the bilinear term is replaced by a new variable z, which is a measurable function that is also an element of U^*, and a set of linear inequalities on z, u, and w that preserve the feasibility of the state equation; that is, if u and w satisfy $Au = uw + f$, then the choice $z = uw$ satisfies all of the newly introduced linear inequalities. Here, we make the following assumption.

Assumption 2.1 Let there be bounds u_ℓ, $u_u \in H$ and w_ℓ, $w_u \in L^\infty(\Omega)$ so that $u_\ell \le u \le u_u$ and $w_\ell \le w \le w_u$ hold pointwise a.e. for all $(u, w) \in U \times W$ that solve $Au + uw = f$ and satisfy $w \in C$.

Clearly, the multiplication of these bounds, $u_\ell w_\ell$, $u_\ell w_u$, $u_u w_\ell$, $u_u w_u \in H$ by Hölder's inequality. We note that the assumption of bounds on the state is quite strong, but we emphasize that such bounds are generally required only in the subset of the computational domain Ω on which N actually acts. For example, terms like $N(u, w) = \chi_A uw$ for compact subsets $A \subset\subset \Omega$ may occur in topology optimization problems like in [25] because the control design may be restricted to A and is extended by zero to the complement of A. Consequently, higher interior regularity of the PDE solutions can help to establish implementable (i.e., generally uniform) bounds. For rich classes of PDEs that are governed by an elliptic operator like in our setting, $L^\infty(\Omega)$-bounds on u can be established by following of the strategy and results in Appendix B in [28] under assumptions on f and the coefficients of the elliptic operator, see also Theorem 4.5 and its proof in [58].

We note that the different parts of the proofs do not require the full assumed regularity here, but again we prefer this slightly more restrictive setting in the interest of a cleaner presentation. Under Assumption 2.1, the McCormick relaxation of (OCP) can be derived pointwise a.e. in a similar way as for finite-dimensional problems; see [38]:

$$
\begin{aligned}
\min_{u,w,z}\ & j(u, w) + \alpha\,\mathrm{TV}(w) \\
\text{s.t. }\ & Au + z = f \ \text{ in } U^*, \\
& z \ge u_\ell w + u w_\ell - u_\ell w_\ell \ \text{ a.e.}, \\
& z \ge u_u w + u w_u - u_u w_u \ \text{ a.e.}, \\
& z \le u_u w + u w_\ell - u_u w_\ell \ \text{ a.e.}, \\
& z \le u_\ell w + u w_u - u_\ell w_u \ \text{ a.e.}, \\
& u \in U,\ u_\ell \le u \le u_u \ \text{ a.e.}, \\
& w \in C,\ w_\ell \le w \le w_u \ \text{ a.e.}
\end{aligned}
\tag{McC}
$$

We obtain that (McC) is a relaxation of (OCP) with a convex feasible set in the proposition below. As mentioned in the introduction, (generalized [54]) McCormick relaxations have been studied for ODEs with finite-dimensional inputs to obtain concave and convex upper and lower bounds for their solution trajectories [39, 47, 48, 52–54, 56, 61, 62]. This implies analogous relaxation results for their settings. In the context of ODE-constrained optimal control, such a result was shown for a pointwise (that is infinite-dimensional) control for nonlinearities given by factorable functions and appropriate Lipschitz assumptions, see Lemma 1, Assumption 3, and Theorem 1 in [51]. Their observations can likely be used to transfer some of our ideas to more general nonlinearities than present in (OCP). We stress that the analysis of the PDE becomes increasingly difficult when higher-degree monomials are present and generally requires a study of the specific case to determine if the PDE has a solution and if yes if it lies in a function space that provides enough regularity for the analysis below. Using McCormick relaxations for such infinite-dimensional control settings and envelopes around solution trajectories was also proposed in [27].

Proposition 2.2 *Let Assumption 2.1 hold.*

1. *If* (u, w) *is feasible for* (OCP)*, then* (u, w, z) *with* $z = uw$ *is feasible for* (McC)*. In particular, the infimum* (McC) *is a lower bound on the infimum of* (OCP)*.*
2. *The feasible set of* (McC) *is convex and bounded in* $U \times H \times H$*.*

Proof Regarding feasibility, the state equation, the inclusion in C, and the two last pointwise inequalities are immediate. The satisfaction of the four additional pointwise inequalities for the choice $z = uw$ follows as in the finite-dimensional case by rearranging them. For example, we obtain $(u - u_\ell)w \geq (u - u_\ell)w_\ell$ a.e. for the first inequality. The state equation in (McC) and the additional inequalities are affine in (u, w, z). Moreover, C is assumed to be convex so that (McC) has a convex feasible set. The feasible w are bounded in H. In combination with the pointwise bounds on u, we obtain boundedness of the feasible z in H from the McCormick inequalities. In turn, the continuous invertibility of A and the embedding $H \cong H^* \hookrightarrow^c U^*$ imply the boundedness of the feasible u in U. □

Remark 2.3 Because PDE-constrained optimization problems generally feature a large number of variables after discretization due to their distributed nature, the pointwise McCormick inequalities add many linear constraints—on the order of six times the number of discretization cells—to a later fully discretized problem if one enforces them on every degree of freedom of a finite-element discretization of the state variable.

Without any additional constraints, we do not know whether we can actually bound z uniformly in H, however, so we add these bound constraints on u, which are again satisfied by all feasible points for (OCP).

Remark 2.4 For the product xy of a binary-valued variable x and a fractional-valued variable y with $y_\ell \leq y \leq y_u$, as is present in our guiding MIPDECO example, a envelopes arises by replacing the product with a new variable z and use the so-called big-M linearization, where y_ℓ and y_u assume the roles of the big-Ms. This yields the linear inequality system

$$y_\ell x \leq z \leq y_u z \quad \text{and} \quad (x - 1)y_u \leq z - y \leq (x - 1)y_\ell.$$

Therein, the two bounds on z correspond to the first and third inequality in the McCormick envelopes as introduced in (McC) and the bounds on $z - y$ correspond to the second and fourth inequality in the McCormick envelopes so that these two approaches are equivalent for our guiding example.

Proposition 2.5 *In the setting of Sect. 2.1, the problems* (OCP) *and* (McC) *admit solutions.*

Proof The existence of solutions for (OCP) follow with standard arguments from the direct method of calculus of variations.

By construction of (McC) as a relaxation of (OCP), its feasible set is nonempty, and there exists a minimizing sequence $(u^n, w^n, z^n)_n \subset U \times H \times H$ of feasible points, which has at least one accumulation point $(u, w, z) \in H \times H \times H$ with respect to weak convergence in $H \times H \times H$, which follows directly from Proposition 2.2. Because A^{-1} is a bounded, linear operator, u satisfies the state equation and is an element of U, too. Moreover, because H is compactly embedded in U^* and thus $z^n \to z$ holds in U^*, we even obtain

$$u^n = A^{-1}(f - z^n) \to A^{-1}(f - z) = u \text{ in } U$$

and in turn $u^n \to u$ in H and pointwise a.e. after restricting to a suitable subsubsequence. The pointwise a.e. convergence yields $u_\ell \le u \le u_u$.

The assumptions on C imply $w \in C$. Because j is bounded below, $(\mathrm{TV}(w^n))_n$ is bounded; and again after restricting to a suitable subsubsequence, we obtain $w^n \to w$ in $L^1(\Omega)$ and pointwise a.e. from the weak-* sequential compactness properties of the TV-seminorm; see Theorem 3.23 in [1]. We obtain $w_\ell \le w \le w_u$.

It remains to assert that the weak limit z satisfies the McCormick inequalities before proving the optimality. This follows from Lemma A.1.

The continuity properties of j and the lower semi-continuity of the TV-seminorm with respect to convergence in $L^1(\Omega)$ and in turn also in H yield that (u, w, z) is a minimizer. □

Remark 2.6 In our examples, the space W will be (a subspace of) the space of functions of bounded variation $\mathrm{BV}(0, 1)$, which is not reflexive. There are important sets C that are subsets of such spaces and also convex, bounded, and weakly sequentially closed in H. An example is $\{w \in \mathrm{BV}(0, 1) \mid \mathrm{TV}(w) \le \kappa\}$ for some $\kappa > 0$; see also [8] for such a constraint set in the context of integer optimal control.

The challenges of the pointwise McCormick envelopes for (McC) have already been sketched in Remark 2.3. The situation is further complicated if one wants to improve the lower bound induced by (McC). A straightforward application of a bound-tightening procedure implies that one alternatingly picks $\tilde{x} \in \Omega$ and solves (one of) the optimization problems

$$\inf_{u,w,z} u(\tilde{x}) \text{ s.t. } (u, w, z) \text{ is feasible for (McC)}$$

and

$$\sup_{u,w,z} u(\tilde{x}) \text{ s.t. } (u, w, z) \text{ is feasible for (McC)}$$

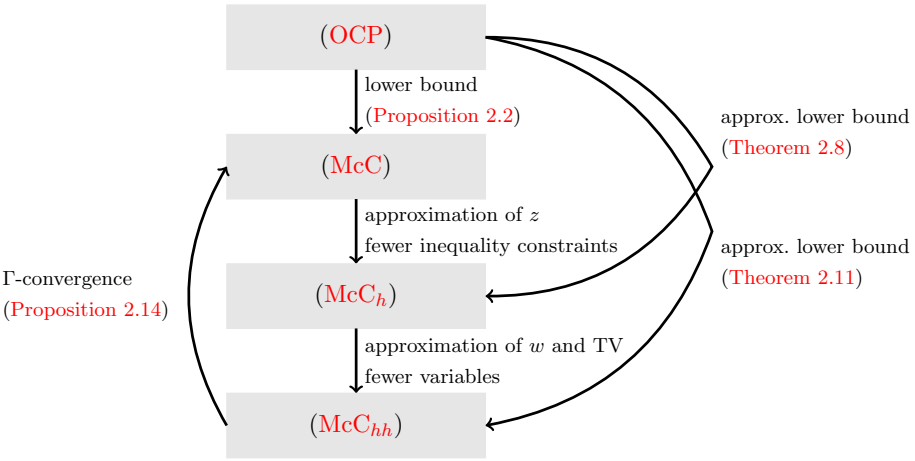


Fig. 1 Two-stage approximation of (McC) by reduced problems (McC_h) and (McC_{hh}) in order to obtain approximate lower bounds on (OCP)

in order to then update $u_\ell(\tilde{x})$ and $u_u(\tilde{x})$ to the computed values. Of course, this is sensible only if u in $C(\bar{\Omega})$ or has a meaningful pointwise interpretation. In a practical implementation, such optimization problems may, for example, be solved for nodes $\tilde{x}$ of a finite-element discretization with a nodal basis. Since improving the bounds yields further possible improvements, it makes sense to repeatedly solve such problems until no further progress is made. In total, the bound-tightening procedure can become computationally very expensive. Therefore, in the next section we analyze a local averaging of the nonlinearity that leads to approximate lower bounds by means of a grid whose mesh size can then be chosen coarser than the mesh size of the grid that is used to the discretize the state variable.

2.3 Approximate McCormick relaxations by local averaging

We propose, analyze, and assess a local averaging that modifies the problem before introducing the variable z. Thus, the variable z and in a second approximation step w, which only enters the relaxed PDE implicitly through z, and the bounds on the product uw are approximated by means of a coarse grid that allows us to reduce the size of the resulting optimization problem. We give an overview on the approximation arguments of this subsection in Fig. 1.

Let $\mathcal{Q}_h = \{Q_h^1, \ldots, Q_h^N\}$ be a partition of the domain Ω into intervals / squares / cubes /... with mesh size h. Locally averaging the nonlinearity of the state equation from (McC) means that $Au + uw = f$ is replaced by the variant

$$Au + (P_h u)(P_h w) = f \tag{2}$$

with P_h being defined in (1). We define the locally averaged McCormick relaxation with respect to the partition $\mathcal{Q}_h$ as

$$
\begin{aligned}
\min_{\substack{u,w,\\ z_1,\dots,z_{N_h}}} \quad & j(u,w) + \alpha\,\mathrm{TV}(w) \\
\text{s.t.} \quad & Au + z = f \quad \text{in } U^*, \\
& z = \sum_{i=1}^{N_h} z_i \chi_{Q_h^i}, \\
& z_i \geq u_\ell^i (P_h w) + (P_h u) w_\ell^i - u_\ell^i w_\ell^i && \text{on } Q_h^i \text{ for all } i \in \{1,\dots,N_h\}, \\
& z_i \geq u_u^i (P_h w) + (P_h u) w_u^i - u_u^i w_u^i && \text{on } Q_h^i \text{ for all } i \in \{1,\dots,N_h\}, \\
& z_i \leq u_u^i (P_h w) + (P_h u) w_\ell^i - u_u^i w_\ell^i && \text{on } Q_h^i \text{ for all } i \in \{1,\dots,N_h\}, \\
& z_i \leq u_\ell^i (P_h w) + (P_h u) w_u^i - u_\ell^i w_u^i && \text{on } Q_h^i \text{ for all } i \in \{1,\dots,N_h\}, \\
& u \in U, u_\ell^i \leq P_h u \leq u_u^i && \text{on } Q_h^i \text{ for all } i \in \{1,\dots,N_h\}, \\
& w \in C,\ w_\ell \leq w \leq w_u \text{ a.e.}, \\
& z_1,\dots,z_{N_h} \in \mathbb{R}.
\end{aligned}
\tag{McC$_h$}
$$

Note that (McC$_h$) is still in function space regarding u and w. In (McC$_h$) the statement that the four McCormick inequalities on z_i hold on Q_h^i might be misunderstood. We therefore highlight that $P_h u$ and $P_h w$ are constant on the cells Q_h^i so that there are exactly four McCormick inequalities per cell $Q_i^h \in \mathcal{Q}_h$. The same argument applies for the bounds on $P_h u$. The bounds $u_\ell^i \in \mathbb{R}$ and $u_u^i \in \mathbb{R}$ assume the roles of the functions u_ℓ and u_u per grid cell for the locally averaged function $P_h u$. Specifically, the function $P_h u$ is bounded from below by u_ℓ^i and from above by u_u^i on the i-th grid cell Q_h^i. Similarly, the bounds w_ℓ^i and w_u^i assume the roles of w_ℓ and w_u.

By employing only finitely many bounds on $P_h u$, we can circumvent the aforementioned regularity problems if the number of constraints is not too large; that is, its Lagrange multiplier may sensibly be interpreted as a vector in $\mathbb{R}^n$. Moreover, the local averaging opens up possibilities for adaptive McCormick relaxations when using a corresponding adaptive refinement of the partition $\mathcal{Q}_h$ and will also reduce the computational effort for improving the (approximate) lower bound. We can now impose inequalities on the locally averaged state $P_h u$ and tighten them algorithmically in order to reduce the gap between the relaxation and the original problem. We now impose an assumption that implies that (McC$_h$) admits a solution and is an approximate relaxation of (OCP). This means that the optimal objective of (OCP) is bounded from below by the objective of (McC$_h$) minus a bound on the approximation error.

Assumption 2.7 In addition to the setting from Sect. 2.1, we assume the following.

1. Let $\{\mathcal{Q}_h\}_h$ be a sequence of partitions of the domain Ω with mesh sizes h.
2. Let the equation (2) have a unique solution for all $w \in \mathrm{BV}(\Omega)$.

3. For all h and all $Q_i \in \mathcal{Q}_h$, let the bounds u_ℓ^i, u_u^i, w_ℓ^i, $w_u^i \in \mathbb{R}$ be such that $u_\ell^i \le (P_h u)|_{Q_i} \le u_u^i$ and $w_\ell^i \le (P_h w)|_{Q_i} \le w_u^i$ hold for all $w \in C \cap [w_\ell, w_u]$ and $u \in U$ that solve $Au + (P_h u)(P_h w) = f$.

We will verify all parts of Assumption 2.7 for an example in Sect. 3.

Theorem 2.8 *Let Assumption 2.7 hold. Then* (McC_h) *has a convex feasible set that is bounded in* $U \times H \times \mathbb{R}^{N_h}$ *and admits a minimizer. Let* (u, w) *be feasible for* (OCP). *Then the tuple* (u_h, w_h, z_h) *with* $w_h := w$, u_h *being the unique solution to* (2), *and* $z_h := (P_h u_h)(P_h w_h)$ *is feasible for* (McC_h), *and the objective value satisfies*

$$|j(u_h, w_h) + \alpha\,\mathrm{TV}(w_h) - j(u, w) - \alpha\,\mathrm{TV}(w)| \le L_u \|u - u_h\|_U, \tag{3}$$

where L_u *is the Lipschitz constant of* j *with respect to its first argument on the feasible set of* (McC_h). *Moreover,*

$$m_{(\mathrm{OCP})} \ge m_{(\mathrm{McC}_h)} - L_u \|\bar{u} - \bar{u}_h\|_U \tag{4}$$

if $m_{(\mathrm{McC}_h)}$ *denotes the infimum of* (McC_h), $m_{(\mathrm{OCP})}$ *denotes the infimum of* (OCP), $(\bar{u}, \bar{w})$ *denotes the minimizer of* (OCP), *and* $\bar{u}_h$ *the corresponding solution to* (2).

Proof The convexity and boundedness of the feasible set follow as in Proposition 2.2. The feasibility of (u_h, w_h, z_h) for (McC_h) follows by construction. Because the feasible set of (McC) is nonempty (see Proposition 2.5), there exist feasible points of (McC_h). Because the feasible set of (McC_h) is bounded, the desired constant L_u exists by assumption on j.

Consequently, we can consider a minimizing sequence $\{(u_h^k, w_h^k, z_h^k)\}_k$ for (McC_h) and apply similar arguments as in the proof of Proposition 2.5 in order to prove existence of a minimizer. The z_h^k have a finite-dimensional piecewise constant ansatz and are bounded because of the boundedness of w implied by the properties of C so that they admit a (norm-)convergent subsequence in H with limit $\bar{z}_h$. Because of the continuous invertibility of A, there is a corresponding subsequence of $\{u_h^k\}_k$ that converges to $\bar{u}_h = A^{-1}(-\bar{z}_h + f)$. By possibly passing to a further subsequence, the corresponding subsequence of $\{w_h^k\}_k$ converges weakly* to $\bar{w}_h$ in W. This implies $\bar{w}_h \in C$ and $w_\ell \le \bar{w}_h \le w_u$. Passing to a further subsequence, we obtain pointwise a.e. convergence for all three subsequences and in turn feasibility of the limit triple $(\bar{u}_h, \bar{w}_h, \bar{z}_h)$ for (McC_h), where we use that the projection to the piecewise constant functions is continuous with respect to all L^p-norms. For the objective, we obtain that the first term converges because of its assumed continuity properties. Moreover, the TV-term is lower semicontinuous with respect to convergence in H; and, in turn, we obtain that $(\bar{u}_h, \bar{w}_h, \bar{z}_h)$ realizes the infimal objective value and is thus a minimizer.

The approximation bound (3) follows from the Lipschitz continuity of j and the assumed choice (u_h, w_h, z_h).

To prove the approximate lower bound (4), let $(\bar{u}, \bar{w})$ be a minimizer to (OCP), and let $\bar{w}_h = \bar{w}$. Then Assumption 2.7 2 gives a unique solution $\bar{u}_h$ to (2), and we choose $\bar{z}_h = (P_h u_h)(P_h w)$. Assumption 2.7 3 yields that the triple $(\bar{u}_h, \bar{w}_h, \bar{z}_h)$ is feasible

for (McC_h). We deduce by means of $\bar{w}_h = \bar{w}$ and the optimality of $(\bar{u}, \bar{w}, \bar{z})$ that

$$\begin{aligned} m_{(\mathrm{OCP})} = j(\bar{u}, \bar{w}) + \alpha\,\mathrm{TV}(\bar{w}) &\geq j(\bar{u}_h, \bar{w}_h) + \alpha\,\mathrm{TV}(\bar{w}_h) - L_u\|u - u_h\|_U \\ &\geq m_{(\mathrm{McC}_h)} - L_u\|u - u_h\|_U. \end{aligned}$$

□

We note that approximation results on the difference $\|u - u_h\|_U$ that are required to obtain a meaningful bound can generally be obtained for broad classes of elliptic and parabolic PDEs.

The locally averaged McCormick relaxation (McC_h) has the drawback that while only a local average of w is required for the McCormick inequalities, the full information is required in order to represent the total variation term correctly. To reduce the number of optimization variables further, one is tempted to replace $\mathrm{TV}(w)$ by $\mathrm{TV}(P_h w)$. Then one could use $P_h w = \sum_{i=1}^{N_h} w_i \chi_{Q_h^i}$ for real coefficients w_i, $i \in \{1, \ldots, N_h\}$, in order to replace the optimization over w in $\mathrm{BV}(\Omega)$ by an optimization over $\mathbb{R}^N$. The lower semicontinuity inequality $\mathrm{TV}(w) \leq \liminf_{h \searrow 0} \mathrm{TV}(P_h w)$ is strict in general if the dimension of Ω is larger than one. To make things worse, even a Γ-convergence result cannot be obtained directly because the geometries of the functions in the limit $h \searrow 0$ are restricted due to the restrictions of the geometry of the elements of the partitions $\mathcal{Q}_h$. This is noted at the end of section 1 in [14] with references to [26] and [13] that employ discrete (anisotropic) approximations of the total variation. Specifically, the isotropic unit ball that is due to the Euclidean norm in the definition of the total variation is not recovered in the limit process for $h \searrow 0$, and other limit functionals that have nonisotropic unit balls may arise for periodic discretizations; this situation is analyzed and reported in [18].

We introduce a variant of the McCormick relaxation where both the McCormick inequalities and the control discretization employ a local averaging with respect to the same grid based on the following assumption. To be able to obtain the desired approximation bound and a Γ-convergence to the limit problem (McC), we assume that a regularization of the total variation penalty, such as the L^2-regularization of the dual formulation presented in [15], is coupled to the grid used for the local averaging.

Assumption 2.9 In addition to Assumption 2.7, we assume that there is a sequence of approximations $(\mathrm{TV_h})_h$ of TV that is consistent with local averaging. Specifically, we assume for all h that $\mathrm{TV_h} : L^1(\Omega) \to [0, \infty]$ is lower semicontinuous, $\mathrm{TV_h} \circ P_h \leq \mathrm{TV}$, and $\mathrm{TV_h} \circ P_h \leq \mathrm{TV_h}$.

Remark 2.10 In one-dimensional domains, the choice $\mathrm{TV} = \mathrm{TV_h}$ immediately satisfies Assumption 2.9, which follows from Lemma 2.15 below. On multidimensional domains, one can, for example, choose suitable finite-element discretizations of the total variation seminorm like a approximation with lowest-order Raviart–Thomas elements; see §4.1 in [14]. Note that a correct discretization of the TV seminorm is not trivial in our intended context of mixed-integer PDE-constrained and we refer to the recent article [49] for details on the difficulties and a solution based on the aforementioned lowest-order Raviart–Thomas elements.

The variant of the McCormick relaxation that replaces $\mathrm{TV}(w)$ by $\mathrm{TV_h}(w)$ is:

$$
\begin{aligned}
\min_{\substack{u,w_1,\dots,w_{N_h},\\ z_1,\dots,z_{N_h}}} \quad & j(u,w) + \alpha\,\mathrm{TV_h}(w) \\
\text{s.t.} \quad & Au + z = f \quad \text{in } U^*, \\
& w = \sum_{i=1}^{N_h} w_i \chi_{Q_h^i},\ z = \sum_{i=1}^{N_h} z_i \chi_{Q_h^i}, \\
& z_i \geq u_\ell^i w_i + (P_h u) w_\ell^i - u_\ell^i w_\ell^i && \text{on } Q_h^i \text{ for all } i \in \{1,\dots,N_h\}, \\
& z_i \geq u_u^i w_i + (P_h u) w_u^i - u_u^i w_u^i && \text{on } Q_h^i \text{ for all } i \in \{1,\dots,N_h\}, \\
& z_i \leq u_u^i w_i + (P_h u) w_\ell^i - u_u^i w_\ell^i && \text{on } Q_h^i \text{ for all } i \in \{1,\dots,N_h\}, \\
& z_i \leq u_\ell^i w_i + (P_h u) w_u^i - u_\ell^i w_u^i && \text{on } Q_h^i \text{ for all } i \in \{1,\dots,N_h\}, \\
& u \in U,\ u_\ell^i \leq P_h u \leq u_u^i && \text{on } Q_h^i \text{ for all } i \in \{1,\dots,N_h\}, \\
& w \in C,\ w_\ell \leq w \leq w_u \text{ a.e.}, \\
& z_1,\dots,z_{N_h} \in \mathbb{R}
\end{aligned}
\tag{McC$_{hh}$}
$$

Note that (McC$_{hh}$) is still in function space regarding u. With the help of Assumption 2.9, we can show that (McC$_{hh}$) is an approximate lower bound on (McC$_h$) and in turn on (OCP). We note that the inequalities in Assumption 2.9 are implementable in practice and can be relaxed to approximate inequalities with an h-dependent error that tends to zero for $h \searrow 0$.

Theorem 2.11 *Let Assumption 2.9 hold. Then* (McC$_{hh}$) *has a convex feasible set that is bounded in* $U \times \mathbb{R}^{N_h} \times \mathbb{R}^{N_h}$ *and admits a minimizer. We define* $w_h := \sum_{i=1}^{N_h} w_i^h \chi_{Q_h^i}$ *for a feasible point* $(u_{hh}, w_h^1, \dots, w_h^{N_h}, z_h^1, \dots, z_h^{N_h})$ *of* (McC$_{hh}$). *Then the tuple* $(u_{hh}, w_h, z_h^1, \dots, z_h^{N_h})$ *is feasible for* (McC$_h$).

Let $(u_h, w_h, z_h^1, \dots, z_h^{N_h})$ *be feasible for* (McC$_h$). *Then* $(u_h, w_h^1, \dots, w_h^{N_h}, z_h^1, \dots, z_h^{N_h})$ *with* $w_h^i := (P_h w_h)|_{Q_h^i}$, $i \in \{1,\dots,N_h\}$, *is feasible for* (McC$_{hh}$). *Moreover,*

$$
j(\bar{u}_h, w_h) + \alpha\,\mathrm{TV}(w_h) \geq m_{(\mathrm{McC}_{hh})} - L_w \|w_u - w_\ell\|_{L^\infty}^{\frac{1}{2}} \sqrt{d}^{\frac{1}{2}}\, \mathrm{TV}(\bar{w}_h)^{\frac{1}{2}} h^{\frac{1}{2}}, \tag{5}
$$

where $m_{(\mathrm{McC}_{hh})}$ *is the infimum of* (McC$_{hh}$) *and* $\bar{w}_h$ *is a minimizer of* (McC$_h$), $\bar{u}_h$ *the corresponding solution to* (2), *and* L_w *is the Lipschitz constant of* j *with respect to its second argument on* C.

Moreover,

$$
m_{(\mathrm{OCP})} \geq m_{(\mathrm{McC}_{hh})} - L_u \|\bar{u} - \bar{u}_h\|_U - L_w \|w_u - w_\ell\|_{L^\infty}^{\frac{1}{2}} \sqrt{d}^{\frac{1}{2}}\, \mathrm{TV}(\bar{w})^{\frac{1}{2}} h^{\frac{1}{2}}, \tag{6}
$$

where $m_{(\mathrm{OCP})}$ *is the infimum of* (OCP), $(\bar{u}, \bar{w})$ *is a minimizer of* (OCP), *and* L_u *is the Lipschitz constant of* j *with respect to its first argument on the feasible set of* (McC$_{hh}$).

Proof The convexity and boundedness of the feasible set follow as in Proposition 2.2. The respective feasibility relations follow by construction, the definitions of (McC_h) and (McC_{hh}), and the assumed properties of C. The existence of minimizers for (McC_{hh}) follows with the same arguments as for (McC_h) with the only difference being that the existence of a convergent subsequence in H of the control inputs w as part of a minimizing sequence can already be deduced from the finite-dimensional ansatz. Because the feasible set of (McC_h) is bounded, the desired constant L_u exists. To prove the approximate lower bound (5), we deduce that, for all $(u_h, w_h, z_1^h, \dots, z_{N_h}^h)$ that are feasible for (McC_h), the estimates

$$\begin{aligned} m_{(\mathrm{McC}_{hh})} &\leq j(u_h, P_h w_h) + \alpha\, \mathrm{TV_h}(P_h w_h) \\ &\leq j(u_h, w_h) + \alpha\, \mathrm{TV_h}(P_h w_h) + L_w \|w_h - P_h w_h\|_H \\ &\leq j(u_h, w_h) + \alpha\, \mathrm{TV}(w_h) + L_w \|w_u - w_\ell\|_{L^\infty}^{\frac{1}{2}} \|w_h - P_h w_h\|_{L^1(\Omega)}^{\frac{1}{2}} \\ &\leq j(u_h, w_h) + \alpha\, \mathrm{TV}(w_h) + L_w \|w_u - w_\ell\|_{L^\infty}^{\frac{1}{2}} \sqrt{d}^{\frac{1}{2}}\, \mathrm{TV}(w_h)^{\frac{1}{2}} h^{\frac{1}{2}}, \end{aligned}$$

hold, where we have used Assumption 2.9 and the Lipschitz continuity of j with respect to the second argument for the second inequality and (the proof of) (12.24) in Theorem 12.26 in [31] for the third inequality.

To prove the approximate lower bound (6), we chain this estimate with the arguments from Theorem 2.8. Let $(\bar{u}, \bar{w})$ be a minimizer to (OCP), and let $\bar{w}_h = \bar{w}$. Then Assumption 2.7 2 gives a unique solution $\bar{u}_h$ to (2), and we choose $\bar{z}_h = (P_h u_h)(P_h \bar{w})$. Assumption 2.7 3 gives the required feasibility, and using $P_h \bar{w}$ as well as $(\bar{u}_h, P_h \bar{w}, \bar{z}_h^1, \dots, \bar{z}_h^{N_h})$ in the estimate above gives

$$\begin{aligned} m_{(\mathrm{McC}_{hh})} &\leq j(\bar{u}_h, \bar{w}) + \alpha\, \mathrm{TV}(\bar{w}) + L_w \|w_u - w_\ell\|_{L^\infty}^{\frac{1}{2}} \sqrt{d}^{\frac{1}{2}}\, \mathrm{TV}(\bar{w})^{\frac{1}{2}} h^{\frac{1}{2}} \\ &\leq j(\bar{u}, \bar{w}) + \alpha\, \mathrm{TV}(\bar{w}) + L_u \|\bar{u} - \bar{u}_h\|_U \\ &\quad + L_w \|w_u - w_\ell\|_{L^\infty}^{\frac{1}{2}} \sqrt{d}^{\frac{1}{2}}\, \mathrm{TV}(\bar{w})^{\frac{1}{2}} h^{\frac{1}{2}}, \end{aligned} \tag{7}$$

where the second inequality follows from the Lipschitz continuity of j in the first argument. □

We now employ a Γ-convergence [19] argument to prove that the locally averaged McCormick relaxations (McC_{hh}) approximate (McC) for $h \searrow 0$. This observation then implies that cluster points of sequences of minimizers to (McC_{hh}) for $h \searrow 0$ minimize (McC). To this end, we need further compatibility conditions for the bounds on the control and state variables.

Assumption 2.12 In addition to Assumption 2.9, we assume the following.

1. Let $\{\mathfrak{Q}_h\}_h$ satisfy a uniform bounded eccentricity condition; that is, there exists $C > 0$ such that for each Q_i^h there is a ball B_i^h such that $Q_i^h \subset B_i^h$ and $|B_i^h| \leq C|Q_i^h|$. See, for example, Definition 4.3 4. in [32].
2. Let $\sup_{h \searrow 0} \mathrm{TV_h}(w_h) \leq C$ and $\sup_{h \searrow 0} \|w_h\|_{L^1} \leq C$ for some $C > 0$ imply that there exists a subsequence $\{w_h\}_h$ and $w \in W$ such that $w_h \to w$ in $L^1(\Omega)$.

3. Let $\mathrm{TV_h} \circ P_h$ Γ-converge to TV for $h \searrow 0$.
4. Let $u_{\ell,h} := \sum_{i=1}^{N_h} u_\ell^i \chi_{Q_h^i}$, $u_{u,h} := \sum_{i=1}^{N_h} u_u^i \chi_{Q_h^i}$ satisfy $u_{\ell,h} \to u_\ell$ in H and $u_{u,h} \to u_u$ in H.
5. Let $w_{\ell,h} := \sum_{i=1}^{N_h} w_\ell^i \chi_{Q_h^i}$, $w_{u,h} := \sum_{i=1}^{N_h} w_u^i \chi_{Q_h^i}$ satisfy $w_{\ell,h} \to w_\ell$ in H and $w_{u,h} \to w_u$ in H.
6. Let u solve $Au + uw = f$ for some $w \in H$. Then $u_h \to u$ in U holds for the solutions u_h to (2) for $h \searrow 0$.

Remark 2.13 We note that Assumptions 2.12, 4, and 5 can generally be ensured by inferring suitable bounds $w_{\ell,h}$, $w_{u,h}$, $u_{\ell,h}$, $u_{u,h}$ from valid bounds w_ℓ, w_u, u_ℓ, u_u. The other assumptions are quite typical when approximating such problems or their solutions.

Proposition 2.14 *Let Assumption 2.12 hold. Let the* $\{(\bar{u}_{hh}, \bar{w}_h^1, \ldots, \bar{w}_h^{N_h}, \bar{z}_h^1, \ldots, \bar{z}_h^{N_h})\}_h$ *be solutions to the problems* (McC_{hh}) *for* $h \searrow 0$. *Then the sequence* $\{(\bar{u}_{hh}, \bar{w}_h, \bar{z}_h)\}_h$ *with* $\bar{w}_h := \sum_{i=1}^{N_h} \bar{w}_h^i \chi_{Q_h^i}$ *and* $\bar{z}_h := \sum_{i=1}^{N_h} \bar{z}_h^i \chi_{Q_h^i}$ *admits an accumulation point* $(\bar{u}, \bar{w}, \bar{z}) \in U \times W \times H$ *such that for a subsequence (for ease of notation denoted by the same symbol)*

$$\bar{u}_{hh} \to \bar{u} \text{ in } U \text{ and } \bar{w}_h \overset{*}{\rightharpoonup} \bar{w} \text{ in } W \text{ and } \bar{z}_h \rightharpoonup \bar{z} \text{ in } H.$$

Moreover, the point $(\bar{u}, \bar{w}, \bar{z})$ *minimizes* (McC).

Proof The sequence $\{\bar{z}_h\}_h$ is bounded in H by Assumption 2.12 4 and 5 and the McCormick inequalities. Consequently, $\{\bar{u}_{hh}\}_h$ is also bounded in U because A is continuously invertible. In turn, after possibly passing to a subsequence, $\bar{z}_h \rightharpoonup \bar{z}$ holds for some $\bar{z} \in H$ and $\bar{u}_{hh} \to \bar{u}$ in U for some $\bar{u} \in U$ because of the compact embedding $H \hookrightarrow^c U^*$ and the continuous invertibility of A. This argument also implies that the original state equation $A\bar{u} + \bar{z} = f$ holds for the limit.

Let u_{hh} solve (2) for $\bar{w}$. Then $(u_{hh}, P_h\bar{w}, (P_h\bar{w})(P_h u_{hh}))$ is feasible for (McC_{hh}) by Assumption 2.7 3. The optimality of the tuples $(\bar{u}_{hh}, \bar{w}_h^1, \ldots, \bar{w}_h^{N_h}, \bar{z}_h^1, \ldots, \bar{z}_h^{N_h})$ for the problems (McC_{hh}) and Assumption 2.9 give

$$j(\bar{u}_{hh}, \bar{w}_h) + \alpha\, \mathrm{TV_h}(\bar{w}_h) \le j(u_{hh}, P_h\bar{w}) + \alpha\, \mathrm{TV_h}(P_h\bar{w}) \le j(u_{hh}, P_h\bar{w}) + \alpha\, \mathrm{TV}(\bar{w}).$$

Because $\{\bar{w}_h\}_h$, $\{P_h\bar{w}\}_h$, and in turn $\{u_{hh}\}_h$ are bounded and j is Lipschitz and thus bounded on bounded sets, this implies that $\{\mathrm{TV_h}(\bar{w}_h)\}_h$ is bounded in $[0, \infty)$.

The properties of the set C imply that $\{\|\bar{w}_h\|_H\}_h$ is bounded so that $\{\|\bar{w}_h\|_{L^1}\}_h$ is bounded in $\mathbb{R}$, too. Thus, after passing to a subsequence (for ease of notation denoted by the same symbol), Assumption 2.12 2 implies $\bar{w}_h \to \bar{w}$ in $L^1(\Omega)$ for some $\bar{w} \in W$.

The pointwise a.e. bounds on $P_h\bar{w}_h$ and $P_h\bar{u}_h$ in the McCormick inequalities together with Assumption 2.12 4 and 5 imply boundedness of the $\bar{z}_h$ in H.

We are now concerned with feasibility of $(\bar{u}, \bar{w}, \bar{z})$ for (McC). The inclusion $\bar{w} \in C$ follows from the assumed properties of C. For the bounds on the state variable, we first observe

$$\|\bar{u} - P_h\bar{u}_{hh}\|_H \le \|\bar{u} - P_h\bar{u}\|_H + \|P_h\|_{H,H}\|\bar{u} - \bar{u}_{hh}\|_H \to 0$$

from the triangle inequality, the nonexpansiveness of P_h, and Lebesgue's differentiation theorem, which requires Assumption 2.12 1. Because, in addition, $u_{\ell,h} \to u_\ell$ and $u_{u,h} \to u_u$ in H hold, we can choose a further (subsubsub)sequence such that all three parts of the inequalities $u_{\ell,h} \leq P_h \bar{u}_{hh} \leq u_{u,h}$ converge pointwise a.e. In turn, we obtain the pointwise a.e. inequalities $u_\ell \leq \bar{u} \leq u_u$ for the limit.

It remains to show the feasibility of $\bar{z}$ for the four pointwise McCormick inequalities in (McC) and the state equation. We prove the claim only for the first one since the others follow with an analogous argument. The feasibility of the solutions for (McC$_{hh}$) gives

$$\bar{z}_h \geq u_{\ell,h}\bar{w}_h + \bar{u}_{hh}w_{\ell,h} - u_{\ell,h}w_{\ell,h} \quad \text{a.e. in } \Omega.$$

From what has been shown to this point, after passing to an appropriate subsequence, the left-hand side converges weakly to $\bar{z}$, and the right-hand side converges in H and thus also in $L^1(\Omega)$. We apply Lemma A.1 in order to obtain the desired inequality

$$\bar{z} \geq u_\ell \bar{w} + \bar{u}w_\ell - u_\ell w_\ell \quad \text{a.e. in } \Omega.$$

It remains to show that $(\bar{u}, \bar{w}, \bar{z})$ minimizes (McC). We prove this claim by using a Γ-convergence-type argument. We need to show lim inf- and lim sup-inequalities for the objectives when the iterates are restricted to the feasible sets.

For the lim inf-inequality, let $u_{hh} \to u$ in U, $w_h = \sum_{i=1}^{N_h} w_h^i \chi_{Q_h^i} \to w$ in H, and $z_h = \sum_{i=1}^{N_h} z_h^i \chi_{Q_h^i} \rightharpoonup z$ in H with $(u_{hh}, w_h^1, \ldots, w_h^{N_h}, z_h^1, \ldots, z_h^{N_h})$ being feasible for (McC$_{hh}$). Then the continuity properties of j and the lower semi-continuity of the TV-seminorm give the lim inf-inequality

$$j(u, w) + \alpha\,\mathrm{TV}(w) \leq \liminf_{h \searrow 0} j(u_{hh}, w_h) + \alpha\,\mathrm{TV}(w_h).$$

For the lim sup-inequality, let (u, w, z) be feasible for (McC). We define $w_h := P_h w$ and u_{hh} as the unique solution to (2), which exists by Assumption 2.7 2. Then the continuity properties of j, Assumption 2.12 6, and Assumption 2.12 3 imply

$$j(u, w) + \alpha\,\mathrm{TV}(w) = \lim_{h \searrow 0} j(u_{hh}, w_h) + \alpha\,\mathrm{TV}_\mathrm{h}(w_h),$$

so that the lim sup-inequality holds true if we can prove that (u_{hh}, w_h, z_h) is feasible for (McC$_{hh}$). This follows from the compatibility assumed in Assumption 2.7 3. □

As noted above, such a result requires a modification of the total variation functional dependent on h for dimensions larger than one. In the one-dimensional case it is possible prove that P_h is nonexpansive with respect to the TV-seminorm, which in turn implies the desired Γ-convergence in this case without any modification of TV beforehand. This is shown below.

Lemma 2.15 *Let Assumption 2.7 hold. Let $\Omega = (a, b)$ for some $a < b$. Let $f \in \mathrm{BV}(\Omega)$. Then*

$$\mathrm{TV}(P_h f) \leq \mathrm{TV}(f).$$

In particular,

$$\mathrm{TV} \circ P_h \ \Gamma\text{-converges to } \ \mathrm{TV}\,.$$

Proof The first claim can be shown by considering the intervals $Q_i \in \mathcal{Q}_h$ one by one and using the equivalence of TV to the pointwise variation when a *good representative* is chosen; see Definition 3.26, (3.24), and Theorem 3.28 in [1].

Let $f^n \to f$ in L^1. Because P_h is nonexpansive with respect to the L^1-norm, we have $\|P_h f - P_h f^n\|_{L^1} \leq \|f - f^n\|_{L^1}$. Moreover, $P_h f \to f$ in $L^1(\Omega)$ holds by means of the Lebesgue differentiation theorem. In combination, we obtain $P_h f^n \to f$ in $L^1(\Omega)$. Consequently, the lim inf-inequality follows from the lower semi-continuity of TV. The lim sup-inequality follows from the first claim by choosing the constant sequence that has f in every element. □

2.4 Optimization-based bound tightening

While Proposition 2.14 shows that solutions to the approximate relaxations (McC_{hh}) of (OCP) approximate solutions to the true relaxation (McC), it is of higher practical importance that the optimal objective values of (McC_h) and (McC_{hh}) are as large as possible in practice. The only way to influence this are the choices of the bounds u^i_ℓ and u^i_u, which should be as tight as possible in order to have the smallest possible feasible set and in turn the largest possible values for the optimal objectives of (McC_h) and (McC_{hh}) such that (4) and (6) still hold.

A powerful albeit computationally expensive technique in MI(N)LP is optimization-based bound tightening (OBBT), which is based on the observation that a convex set of a relaxation can be reduced (tightened) when minimizing and maximizing each variable in it over the set and then intersecting the set with this optimized bound on said variable. This is often used for variables arising in McCormick relaxations, see, for example, [12, 17, 43, 44, 57]. Further references from more general vantage points on OBBT are [37, 42] and especially [21] regarding efficient techniques.

Inspecting the arguments in the proof of Theorem 2.11 that lead to (6) ((7) in the proof), and similarly in Theorem 2.8, we observe that these arguments remain valid when the feasible set of (McC_{hh}) is shrunk as long as Assumption 2.7 3 is preserved. Then, the *typical* argument for OBBT can be applied and the infimum $m_{(\mathrm{McC}_{hh})}$ increases when the feasible set is shrunk, while the preservation of Assumption 2.7 3 yields that the approximation error bound that needs to be subtracted in order to obtain a valid bound on $m_{(\mathrm{OCP})}$ remains unaffected from this change. Consequently, the lower bound on $m_{(\mathrm{OCP})}$ is improved by this procedure.

Before proving this, we show the desired property that Assumption 2.7 3 is conserved when optimizing the bounds u^i_ℓ, u^i_u.

Lemma 2.16 *Let Assumption 2.7 hold. Let h be fixed. Let $i \in \{1, \ldots, N_h\}$ be fixed. Then*

$$\tilde{u}^i_\ell \,/\, \tilde{u}^i_u := \min / \max \{(P_h u)|_{Q_i} \mid (u, w, z_1, \ldots, z_{N_h}) \text{ is feasible for } (\mathrm{McC}_h)\}. \quad (8)$$

are well defined.

Moreover, $\tilde{u}_\ell^i \le (P_h u)|_{Q_i} \le \tilde{u}_u^i$ holds for all $w \in C \cap [w_\ell, w_u]$ and $u \in U$ that solve (2). In particular, Assumption 2.7 holds if u_ℓ^i is replaced by $\tilde{u}_\ell^i$ and u_u^i is replaced by $\tilde{u}_u^i$.

Proof We note that $u \mapsto (P_h u)|_{Q_i}$ is a weakly continuous operation from H to $\mathbb{R}$. Moreover, using the arguments for the existence of solutions to (McC_h) in the proof of Theorem 2.8, we obtain that the feasible set of (McC_h) is nonempty and sequentially compact with respect to weak convergence of u in H, weak convergence w in H, and convergence of z in H (the last one has a finite-dimensional ansatz). Consequently, (8) has a solution, and $\tilde{u}_\ell^i$ is well defined.

Inspecting Assumption 2.7 shows that it remains to prove that Assumption 2.7 3 stays valid when replacing u_ℓ^i by $\tilde{u}_\ell^i$. For all $w \in C$ and $u \in U$ that solve $Au + (P_h u)(P_h w) = f$, we define $z_i := (P_h u)|_{Q_h^i}(P_h w)|_{Q_h^i}$ for $i \in \{1, \dots, N_h\}$. Inspecting (McC_h), we obtain that $(u, w, z_1, \dots, z_{N_h})$ is feasible for (McC_h) by means of Assumption 2.7 3. Consequently, all $w \in C$ and $u \in U$ that solve $Au + (P_h u)(P_h w) = f$ satisfy

$$\begin{aligned}\tilde{u}_\ell^i &= \min\{(P_h u)|_{Q_i} \mid (u, w, z_1, \dots, z_{N_h}) \text{ is feasible for } (\mathrm{McC}_h)\} \\ &\le \inf\{(P_h u)|_{Q_i} \mid (u, w) \in U \times C,\, Au + (P_h u)(P_h w) = f\},\end{aligned}$$

and thus also $\tilde{u}_\ell^i \le (P_h u)|_{Q_i}$, which proves the first claim.

The second well-definedness and claim follow analogously. □

A similar argument can be made for (McC_{hh}).

Lemma 2.17 *Let Assumption 2.9 hold. Let h be fixed. Let $i \in \{1, \dots, N_h\}$ be fixed. Then*

$$\begin{aligned}\tilde{u}_\ell^i \,/\, \tilde{u}_u^i := \min / \max\{&(P_h u)|_{Q_i} \mid (u, w_1, \dots, w_{N_h}, z_1, \dots, z_{N_h}) \\ &\text{is feasible for } (\mathrm{McC}_{hh})\}.\end{aligned}$$

are well defined.

Moreover, $\tilde{u}_\ell^i \le (P_h u)|_{Q_i} \le \tilde{u}_u^i$ holds for all $w \in C \cap [w_\ell, w_u]$ and $u \in U$. In particular, Assumption 2.9 holds if u_ℓ^i is replaced by $\tilde{u}_\ell^i$ and u_u^i is replaced by $\tilde{u}_u^i$.

Proof The proof parallels the one of Lemma 2.16. □

We can now state a bound-tightening procedure that successively computes new bounds in order to improve the approximate lower bounds (4) or (6) respectively in Algorithm 1.

Using Lemmas 2.16 and 2.17, we can prove our main bound-tightening results for (McC_h) and (McC_{hh}).

Theorem 2.18 *Let Assumption 2.7 be satisfied. Let C be a weakly sequentially compact subset of H. Then Algorithm 1 executed with the feasible set of (McC_h) as $\mathcal{F}^0$ induces a sequence of optimization problems*

$$\min_{u, w, z_1, \dots, z_{N_h}} j(u, w) + \alpha\, \mathrm{TV}(w) \quad \text{s.t.} \quad (u, w, z_1, \dots, z_{N_h}) \in \mathcal{F}^n \tag{McC_h^n}$$

Algorithm 1 Optimization-based bound tightening (OBBT) for (McC_h) (or (McC_{hh}))

Input: Feasible set $\mathcal{F}^0$ of (McC_h) (or (McC_{hh})).

for $n = 1, \ldots$ **do**
  Choose $i \in \{1, \ldots, N_h\}$.
  Choose $s \in \{\ell, u\}$.
  if $s = \ell$ **then**
    $\tilde{u}_\ell^i \leftarrow \min\{(P_h u_h)|_{Q_i} \mid (u_h, w_h, z_h) \in \mathcal{F}^{n-1}\}$.
  else
    $\tilde{u}_u^i \leftarrow \max\{(P_h u_h)|_{Q_i} \mid (u_h, w_h, z_h) \in \mathcal{F}^{n-1}\}$.
  end if
  $\mathcal{F}^n \leftarrow \mathcal{F}^{n-1}$ with u_s^i being replaced by $\tilde{u}_s^i$.
end for

that satisfy (4) *for all* $n \in \mathbb{N}$. *Moreover, the sequence of optimal objective values* $(m_{(\mathrm{McC}_h^n)})_n$ *is monotonically nondecreasing.*

Proof By assumption, the prerequisites of Lemma 2.16 are satisfied in the first iteration and thus hold inductively for all iterations $n \in \mathbb{N}$. This implies that (4) holds for all problems (McC_h^n) because Theorem 2.8 asserts (4) under Assumption 2.7. Because the feasible set is always a subset of the previous one, the infima are monotonically nondecreasing. □

Theorem 2.19 *Let Assumption 2.9 be satisfied. Then Algorithm 1 executed with the feasible set of* (McC_{hh}) *as* $\mathcal{F}^0$ *induces a sequence of optimization problems*

$$\min_{\substack{u, w_1, \ldots, w_{N_h}, \\ z_1, \ldots, z_{N_h}}} j\left(u, \sum_{i=1}^{N_h} \chi_{Q_h^i} w_i\right) + \alpha\, \mathrm{TV}\left(\sum_{i=1}^{N_h} \chi_{Q_h^i} w_i\right) \qquad (\mathrm{McC}_{hh}^n)$$
$$\text{s.t.} \quad (u, w_1, \ldots, w_{N_h}, z_1, \ldots, z_{N_h}) \in \mathcal{F}^n$$

that satisfy (5) *for all* $n \in \mathbb{N}$. *Moreover, the corresponding sequence of optimal objective values* $\{m_{(\mathrm{McC}_{hh}^n)}\}_n$ *is monotonically nondecreasing.*

Proof The proof parallels the one of Theorem 2.19 with the applications of Lemma 2.16 and Theorem 2.8 being replaced by applications of Lemma 2.17 and Theorem 2.11. □

3 Application to an elliptic optimal control problem

We now apply our theoretical considerations from Sect. 2 to an instance of (OCP) with an elliptic PDE that is defined on the one-dimensional domain $\Omega = (0, 1)$. To this end, let $C := \{w \in L^\infty(0, 1) \mid w_\ell \le w(x) \le w_u \text{ a.e.}\}$ and $f \in H$ for real constants $-\pi^2 < w_\ell < w_u < \pi^2$ be fixed.

For these assumptions, we outline the PDE setting in Sect. 3.1 and give ellipticity estimates as well as bounds from above on the norm with constants that are as sharp as

we were able to obtain them from the literature. We provide the objective in Sect. 3.2. We then verify Assumptions 2.1, 2.7, 2.9 and 2.12 in Sect. 3.3. We prove an a priori estimate for $\|u-u_h\|_U$ in the approximate lower bounds (4), (6) in Sect. 3.4. We prove differentiability properties for the control-to-state operator of the PDE in Sect. 3.5. In particular, we verify the assumptions imposed in [29].

3.1 PDE setting

For $w \in C$, we are interested in solutions $u \in U := H_0^1(0,1)$ to the PDE in weak form

$$\int_0^1 \nabla u \nabla v + \int_0^1 wuv = \int_0^1 fv \quad \text{for all } v \in U. \tag{9}$$

The operator form is

$$-\Delta u + uw = f \text{ on } (0,1) \text{ with } u(0) = u(1) = 0,$$

that is, $A = -\Delta$ with homogeneous Dirichlet boundaries and $N(u,w) = uw$. We define the bilinear form $B : U \times U \to \mathbb{R}$ by

$$B(u,v) := \int_0^1 \nabla u \nabla v + \int_0^1 uwv$$

and obtain

$$B(u,u) \geq \|\nabla u\|_H^2 + w_\ell \|u\|_H^2 \quad \text{for all } u \in U. \tag{10}$$

The existence of (unique) solutions to (9) follows from the Lax–Milgram lemma if

$$\|\nabla u\|_H^2 + w_\ell \|u\|_H^2 \geq c_1(w_\ell)\|\nabla u\|_H^2$$

holds for some $c_1(w_\ell) > 0$, which is true if $w_\ell \geq 0$. For the case $w_\ell < 0$, we recall embedding constants for the Sobolev inequalities for the one-dimensional domain $(0,1)$:

$$\pi^2 \|u\|_H^2 \leq \|\nabla u\|_H^2 \text{ and} \tag{11}$$

$$4\|u\|_{L^\infty}^2 \leq \|\nabla u\|_H^2. \tag{12}$$

The constant in (11) is optimal and, for example, given in (1.2) in [59] (see [40] for a proof). The constant in (12) can be found in (13) in [50] (choose $a = \infty$, $b = 2$, $s = 1$ therein). By (11), we obtain that $0 < c_1(w_\ell) := 1 + w_\ell \pi^{-2}$ because (by our choice above)

$$w_\ell > -\pi^2. \tag{13}$$

The Lax–Milgram lemma also yields that the solution to (9) satisfies

$$\|\nabla u\|_H \leq \frac{1}{c_1(w_\ell)}\|f\|_H \text{ and} \tag{14}$$

$$\|u\|_{L^\infty} \leq \frac{1}{2c_1(w_\ell)}\|f\|_H, \tag{15}$$

where the second inequality follows from (12). We observe that u is also the weak solution of the PDE $\int_0^1 \nabla u \nabla v = \int_0^1 gv$ for all $v \in U$ with the choice $g = f - wu$. Because $g \in H$, this gives the improved regularity $u \in H^2(0,1)$; see, for example, Theorem 9.53 in [46].

3.2 Objective

We consider a tracking-type objective functional $j : U \times W \to \mathbb{R}$ that is defined for $(u, w) \in U \times W$ as

$$j(u, w) \coloneqq \frac{1}{2}\|u - u_d\|_H^2. \tag{16}$$

Since our specific choice for j does not depend on w, we will abbreviate it as $j(u)$ in the remainder. If the norm of the input is bounded by a constant r_u, we obtain that a feasible Lipschitz constant of j on the ball $\overline{B_{r_u}(0)}$ is given by $L_u \coloneqq r_u + \|u_d\|_H$. Clearly, $L_w = 0$.

3.3 Verification of assumptions

We now verify Assumptions 2.1, 2.7, 2.9 and 2.12 one by one.

Assumption 2.1 Let $c_2(w_\ell, w_u) \coloneqq 1 - \max\{|w_\ell|, |w_u|\}\pi^{-2}$. We choose $u_u = \frac{1}{2c_2(w_\ell,w_u)}\|f\|_H$ and $u_\ell = -\frac{1}{2c_2(w_\ell,w_u)}\|f\|_H$. Then $c_2(w_\ell, w_u) \leq c_1(w_\ell)$ gives $|u_u| = |u_\ell| \geq \frac{1}{2c_1(w_\ell)}\|f\|_H$ so that together with the bounds w_ℓ, w_u, we obtain that Assumption 2.1 is satisfied because of (15). (We could of course choose the tighter bounds $u_u = \frac{1}{2c_1(w_\ell)}\|f\|_H$ and $u_\ell = -\frac{1}{2c_1(w_\ell)}\|f\|_H$ here, but the relaxed bounds using $c_2(w_\ell, w_u)$ instead of $c_1(w_\ell)$ will be used to assert the other assumptions below.)

Assumption 2.7 We consider a sequence of uniform partitions $\{\mathcal{Q}_{h_n}\}_{n\in\mathbb{N}}$ of the domain $(0, 1)$ into 2^{n+1} intervals with $h_n = 2^{-1-n}$ for $n \in \mathbb{N}$ and $N_{h_n} = 2^{n+1}$, which gives Assumption 2.7 1.

Let $n \in \mathbb{N}$, and abbreviate $h \coloneqq h_n$. For each interval $I \in \mathcal{Q}_h$, the projection P_I is defined as $P_I g \coloneqq \chi_I \frac{1}{h}\int_I g$ for $g \in L^1(0, 1)$. Clearly, P_I is nonexpansive in $L^1(0, 1)$, H, and $L^\infty(0, 1)$; that is,

$$\|P_I g\|_{L^\infty} \leq \|g\|_{L^\infty} \text{ and } \|P_I g\|_H \leq \|g\|_H.$$

The linear projection operator P_h from (1) satisfies $P_h(g) = \sum_{I \in \mathcal{Q}_h} P_I g$ and

$$\|g - P_h g\|_H \leq \pi h \|\nabla g\|_H, \tag{17}$$

where the constant π is due to the application of the one-dimensional Poincaré inequality into the interpolation estimate for piecewise constant functions; see, for example, §3.1 in [16]. From (12.24) in Theorem 12.26 in [31], we also obtain

$$\|w - P_h w\|_{L^1} \leq h \,\mathrm{TV}(w), \tag{18}$$

which we have already used in the proof of Theorem 2.11.

For arbitrary, fixed w, we define the bilinear form $B_h : U \times U \to \mathbb{R}$ as

$$B_h(u, v) \coloneqq \int_0^1 \nabla u \nabla v + \int_0^1 (P_h w)(P_h u) v$$

for $u, v \in U$ and obtain the coercivity

$$\begin{aligned} B_h(u, u) &\geq \|\nabla u\|_H^2 - \max\{|w_\ell|, |w_u|\} \|u\|_H \|P_h(u)\|_H \\ &\geq \|\nabla u\|_H^2 - \max\{|w_\ell|, |w_u|\} \|u\|_H^2. \end{aligned}$$

Consequently, the existence of unique solutions to the PDE

$$\int_0^1 \nabla u \nabla v + \int_0^1 (P_h w)(P_h u) v = \int_0^1 f v \quad \text{for all } v \in U. \tag{19}$$

follows from the Lax–Milgram lemma with analogous estimates to (14) and (15), where $c_1(w_\ell)$ is replaced by $c_2(w_\ell, w_u) = 1 - \max\{|w_\ell|, |w_u|\}\pi^{-2}$ because $0 < c_2(w_\ell, w_u)$. Specifically, we have

$$\|\nabla u\|_H \leq \frac{1}{c_2(w_\ell, w_u)} \|f\|_H \text{ and} \tag{20}$$

$$\|u\|_{L^\infty} \leq \frac{1}{2c_2(w_\ell, w_u)} \|f\|_H, . \tag{21}$$

This proves Assumption 2.7 2. As for (9), we observe that u solves $\int_0^1 \nabla u \nabla v = \int_0^1 g v$ for all $v \in U$ with the choice $g = f - (P_h w)(P_h u)$. Because $g \in H$, this gives the improved regularity $u \in H^2(0, 1)$; see, for example, Theorem 9.53 in [46].

Because P_h is nonexpansive with respect to $L^\infty(0, 1)$, we can reuse the bounds w_ℓ and w_u for w_ℓ^i and w_u^i for all $i \in \{1, \ldots, N_h\}$. With the same argument as for Assumption 2.1, we choose $u_u^i = \frac{1}{2c_2(w_\ell, w_u)} \|f\|_H$ and $u_\ell^i = -\frac{1}{2c_2(w_\ell, w_u)} \|f\|_H$ for all $i \in \{1, \ldots, N_h\}$. In combination, Assumption 2.7 3 is satisfied.

Assumption 2.9 In our one-dimensional setting, Lemma 2.15 implies that Assumption 2.9 holds with the choice $\mathrm{TV_h} \coloneqq \mathrm{TV}$.

Assumptions 2.9 and 2.12 Assumption 2.12 1 is satisfied because of the uniform discretization of the computational domain $\Omega = (0, 1)$ into intervals. Assumption 2.12 2 follows from the properties of the total variation seminorm with the choice $\mathrm{TV_h} \coloneqq \mathrm{TV}$; see Theorem 3.23 in [1]. Assumption 2.12 3 follows from Lemma 2.15. Assumption 2.12 4 follows because we have chosen $u_\ell^i = u_\ell$ and $u_u^i = u_u$ for all i and h. Assumption 2.12 5 follows because we have chosen $w_\ell^i = w_\ell$ and $w_u^i = w_u$ for all i and h. Assumption 2.12 6 follows from Lemma 3.1 that is proven in Sect. 3.4 and a bootstrapping argument / the continuous invertibility of A.

3.4 Estimate on $\|u - u_h\|_H$ and a priori estimates on $m_{(\mathrm{OCP})}$

For our example PDE setting and the tracking-type objective in H, we are able to obtain lower bounds on (4) and (6) that are quadratic in h because we can improve over $\|u - u_h\|_U$ when considering $\|\cdot\|_H$ instead of $\|\cdot\|_U$. We first show the necessary estimates on $\|u - u_h\|_H$ and subsequently prove the lower bounds on $m_{(\mathrm{OCP})}$.

Lemma 3.1 *Let Assumption 2.7 hold. Let a mesh size h be fixed. Let u solve (9) and u_h solve (19) for the same fixed $w \in W$. Then the estimates*

$$\|u - u_h\|_H \le C^a_{3/2}(w_\ell, w_u, f, w)h^{\frac{3}{2}} + C^b_{3/2}(w_\ell, w_u, f)h^2$$

and

$$\|u - u_h\|_H \le C_2(w_\ell, w_u, f, w, u_h)h^2 \tag{22}$$

hold for constants $C^a_{3/2}(w_\ell, w_u, f, w)$, $C^b_{3/2}(w_\ell, w_u, f)$, $C_2(w_\ell, w_u, f, w, u_h) > 0$.

Proof We observe that the local averaging of w and u in the lower-order term in (19) satisfies Galerkin-type orthogonality properties. Specifically, for given $\phi, \psi, \theta \in H$, we have

$$\int_0^1 (\phi - P_h\phi)(P_h\psi)(P_h\theta) = \sum_{i=1}^{N_h} \left(\frac{1}{|Q_h^i|}\int_{Q_h^i}\psi\right)\left(\frac{1}{|Q_h^i|}\int_{Q_h^i}\theta\right)\underbrace{\int_{Q_h^i}\phi - P_h\phi}_{=0} = 0. \tag{23}$$

Thus, we use an Aubin–Nitsche duality argument in order to obtain an estimate on the right-hand side of

$$\|u - u_h\|_H = \sup_{g\neq 0}\frac{(u - u_h, g)_H}{\|g\|_H}. \tag{24}$$

To this end, we consider the solution $p \in U$ to the adjoint PDE

$$\int_0^1 \nabla p \nabla v + \int_0^1 pwv = \int_0^1 gv \quad \text{for all } v \in U \tag{25}$$

for arbitrary $g \in H$, which has the same properties as (9) (the only difference is the source term g instead of f). Moreover, we observe with the choice $v = p$ in (9) and (19) and the insertion of a suitable zero that

$$\int_0^1 \nabla(u - u_h)\nabla p = -\int_0^1 (u - P_h u_h)wp - \int_0^1 (P_h u_h)(w - P_h w)p. \tag{26}$$

We choose $v = u - u_h$ in (25) and apply (26) so that we obtain

$$\begin{aligned}(u - u_h, g)_H &= \int_0^1 \nabla p \nabla(u - u_h) + \int_0^1 pw(u - u_h) \\ &= \int_0^1 (u - u_h)wp - \int_0^1 (u - P_h u_h)wp - \int_0^1 (P_h u_h)(w - P_h w)p.\end{aligned}$$

Combining the first two terms on the right-hand side and subtracting $0 = \int_0^1 (P_h u_h)(w - P_h w)(P_h p)$, which holds because of (23), we obtain

$$(u - u_h, g)_H = \int_0^1 (P_h u_h - u_h)wp - \int_0^1 (P_h u_h)(w - P_h w)(p - P_h p).$$

Inserting another zero and applying $0 = \int_0^1 (P_h u_h)(w - P_h w)(P_h p)$ again give

$$\begin{aligned}(u - u_h, g)_H = &\int_0^1 (P_h u_h - u_h)(w - P_h w)p + \int_0^1 (P_h u_h - u_h)(P_h w)(p - P_h p) \\ &- \int_0^1 (P_h u_h)(w - P_h w)(p - P_h p).\end{aligned}$$

Now, we estimate the terms on the right-hand side and obtain with the same estimates that have been used in the preceding subsections the following estimates:

$$\begin{aligned}&\int_0^1 (P_h u_h - u_h)(w - P_h w)p \\ &\quad\le h\pi \|\nabla u_h\|_H \|w - P_h w\|_H \|p\|_{L^\infty} && \text{Hölder, (17)} \\ &\quad\le h \frac{\pi}{c_2(w_\ell, w_u)} \|f\|_H |w_u - w_\ell|^{\frac{1}{2}} \|w - P_h w\|_{L^1}^{\frac{1}{2}} \frac{1}{2c_1(w_\ell)} \|g\|_H && \text{Hölder, (15), (20)} \\ &\quad\le \frac{\pi |w_u - w_\ell|^{\frac{1}{2}}}{2c_1(w_\ell)c_2(w_\ell, w_u)} \mathrm{TV}(w)^{\frac{1}{2}} \|f\|_H h^{\frac{3}{2}} \|g\|_H, && (18)\end{aligned}$$

$$\int_0^1 (P_h u_h - u_h)(P_h w)(p - P_h p) \le \frac{\pi^2 \max\{|w_u|, |w_\ell|\}}{c_1(w_\ell)c_2(w_\ell, w_u)} \|f\|_H h^2 \|g\|_H, \quad \text{Hölder, (15), (17), (20)}$$

and

$$\int_0^1 (P_h u_h)(w - P_h w)(p - P_h p)$$

$$\leq \frac{\pi |w_u - w_\ell|^{\frac{1}{2}}}{2c_1(w_\ell)c_2(w_\ell, w_u)} \mathrm{TV}(w)^{\frac{1}{2}} \|f\|_H h^{\frac{3}{2}} \|g\|_H. \qquad \text{Hölder, (15), (17), (18), (20)}$$

Combining these considerations, we obtain from (24)

$$\|u - u_h\|_H \leq \underbrace{\frac{\pi |w_u - w_\ell|^{\frac{1}{2}}}{c_1(w_\ell)c_2(w_\ell, w_u)} \mathrm{TV}(w)^{\frac{1}{2}} \|f\|_H}_{=: C^a_{3/2}(w_\ell, w_u, f, w)} h^{\frac{3}{2}} + \underbrace{\frac{\pi^2 \max\{|w_u|, |w_\ell|\}}{c_1(w_\ell)c_2(w_\ell, w_u)} \|f\|_H}_{=: C^b_{3/2}(w_\ell, w_u, f)} h^2,$$

which is the first of the claimed estimates.

We next prove the second estimate. Because we know that u, u_h, p are also $H^2(0, 1)$-functions (see again Theorem 9.53 in [46]), their derivatives are uniformly bounded, and we can estimate them pointwise. Let $(\eta_\varepsilon)_\varepsilon$ be a family of standard mollifiers. For $y \in (0, 1)$, we define the antiderivative $N^y_\varepsilon(x) := \int_{-\infty}^x \eta_\varepsilon(z - y)\,\mathrm{d}\,z$. We test (19) with N^y_ε and deduce

$$\begin{aligned}
\int_0^1 \nabla u_h(x)\eta_\varepsilon(x - y)\,\mathrm{d}\,x &= \int_0^1 f(x)N^y_\varepsilon(x)\,\mathrm{d}\,x - \int_0^1 (P_h u_h)(x)(P_h w)(x)N^y_\varepsilon(x)\,\mathrm{d}\,x \\
&\leq \int_0^1 |f(x) - (P_h u_h)(x)(P_h w)(x)| \int_{-\infty}^\infty \eta_\varepsilon(x - y)\,\mathrm{d}\,y\,\mathrm{d}\,x \\
&= \|f - (P_h u_h)(P_h w)\|_{L^1}.
\end{aligned}$$

Driving $\varepsilon \searrow 0$ and supremizing over the left-hand side, we obtain

$$\|\nabla u_h\|_{L^\infty} \leq \|f - (P_h u_h)(P_h w)\|_{L^1}.$$

An analogous argument and applications of the triangle inequality, Hölder's inequality, (11), and (14) give

$$\|\nabla p\|_{L^\infty} \leq \|g - pw\|_{L^1} \leq \left(1 + \frac{\max\{|w_\ell|, |w_u|\}}{\pi c_1(w_\ell)}\right) \|g\|_H.$$

We also obtain for $\phi \in W^{1,\infty}(0, 1)$ that

$$\|\phi - P_h\phi\|_{L^\infty} \leq \frac{1}{2}h\|\nabla\phi\|_{L^\infty},$$

where $\frac{1}{2}$ is the Wirtinger–Sobolev constant that can be found in (13) in [50] (choose $a = \infty, b = \infty, s = 1$ therein).

Then, using Hölder's inequality with different exponents, we can estimate the two $h^{\frac{3}{2}}$-terms similarly as above but we replace the Hölder conjugates $(2, 2)$ by $(1, \infty)$ in order to obtain h^2-terms, specifically

$$\int_0^1 (P_h u_h - u_h)(w - P_h w)p \leq \frac{1}{2}h\|\nabla u_h\|_{L^\infty}\|w - P_h w\|_{L^1}\|p\|_{L^\infty}$$

$$\leq \frac{1}{4c_1(w_\ell)c_2(w_\ell, w_u)} \operatorname{TV}(w) \times \|f - (P_h u_h)(P_h w)\|_{L^1} h^2 \|g\|_H$$

and

$$\int_0^1 (P_h u_h)(w - P_h w)(p - P_h p) \leq \frac{c_1(w_\ell) + \max\{|w_\ell|, |w_u|\}\pi^{-1}}{4c_1(w_\ell)c_2(w_\ell, w_u))} \operatorname{TV}(w) \|f\|_H h^2 \|g\|_H.$$

In summary, we obtain

$$\|u - u_h\|_H \leq C_2(w_\ell, w_u, f, w, u_h)h^2$$

with

$$C_2(w_\ell, w_u, f, w, u_h) := \frac{1}{4c_1(w_\ell)c_2(w_\ell, w_u)} \Big(4\pi^2 \max\{|w_u|, |w_\ell|\}\|f\|_H + \operatorname{TV}(w)\|f - (P_h u_h)(P_h w)\|_{L^1} + \big(c_1(w_\ell) + \max\{|w_\ell|, |w_u|\}\pi^{-1}\big) \operatorname{TV}(w)\|f\|_H\Big), \tag{27}$$

which is the second estimate. □

We now provide a lower bound on $m_{(\mathrm{OCP})}$ in terms of h^2 for our guiding example.

Lemma 3.2 *Let $j_0 \geq 0$ be a lower bound on $j(\bar{u})$ if $(\bar{u}, \bar{w})$ minimizes (OCP). Let $\hat{w} \in W$ be feasible for (OCP). Then*

$$m_{(\mathrm{OCP})} \geq m_{(\mathrm{McC}_{hh})} - c_{quad}(w_\ell, w_u, f, \hat{w}, j_0, u_{\ell,h}, u_{u,h})h^2. \tag{28}$$

Proof For the objective $j(u) = \frac{1}{2}\|u - u_d\|_H^2$, we obtain $\nabla_u j(u) = u - u_d$. Since we have pointwise lower and upper bounds u_ℓ and u_u on u, for example, from (21), we can overestimate the Lipschitz constant of j on the feasible set with respect to the L^2-norm by setting $\tilde{d}_u(x) := \max\{|u_u(x) - u_d(x)|, |u_d(x) - u_\ell(x)|\}$ for a.a. $x \in (0, 1)$. Then the Lipschitz constant can be estimated as $L_u \leq \|\tilde{d}_u\|_H$.

Using (22) from Lemma 3.1, we obtain

$$m_{(\mathrm{OCP})} \geq m_{(\mathrm{McC}_{hh})} - \|\tilde{d}_u\|_H C_2(w_\ell, w_u, f, \bar{w}, \bar{u}_h)h^2.$$

Inspecting $C_2(w_\ell, w_u, f, \bar{w}, \bar{u}_h)$ in Lemma 3.1, we observe that we can overestimate $\operatorname{TV}(\bar{w}) \leq \frac{j(\hat{u}) + \alpha \operatorname{TV}(\hat{w}) - j_0}{\alpha}$, where $\hat{u}$ denotes the unique solution to (9) for $\hat{w}$. Moreover, using available bounds on $P_h\bar{u}_h$, we can overestimate $|f(x) - (P_h\bar{u}_h)(x)(P_h\bar{w})(x)|$ pointwise by setting

$$\tilde{d}_f(x) := \max\{|f(x) - u_\ell^i w_\ell|, |f(x) - u_\ell^i w_u|, |f(x) - u_u^i w_\ell|, |f(x) - u_u^i w_u|\}$$

for a.a. $x \in Q_i$ for all $Q_i \in \mathcal{Q}_h$. Thus we obtain the constant $c_{quad}(w_\ell, w_u, f, \hat{w}, j_0, u_{\ell,h}, u_{u,h})$ by replacing $\|f-(P_h\bar{u}_h)(P_h\bar{w})\|_{L^1}$ by $\|\tilde{d}_f\|_{L^1}$ and $\mathrm{TV}(\bar{w})$ by $\frac{j(\hat{u})+\alpha\,\mathrm{TV}(\hat{w})-j_0}{\alpha}$ in (27) and multiplying the resulting constant by $\|\tilde{d}_u\|_H$. □

Remark 3.3 We note that the estimate $L_u \le \|\tilde{d}_u\|_H$ can be sharpened by using a posteriori information from the approximate McCormick relaxations in order to bootstrap improved pointwise bounds u_ℓ and u_u on u.

3.5 Properties of the control-to-state operator for (19)

We denote the nonlinear control-to-state operator that maps $w \in C$ to the solution of the PDE (19) with the locally averaged state in the bilinearity as $S : C \to U$. It is well defined and uniformly bounded due to the considerations in Sects. 3.1 and 3.3. We briefly state Lipschitz continuity and differentiability properties of the control-to-state operator so that we verify the assumptions in [29]. Consequently, the algorithmic strategy from [29, 36] can be used to obtain stationary feasible (primal points) when additional integrality restrictions are imposed on w in (OCP) while the analysis carried out above allows us to obtain valid (approximate) lower bounds. While the arguments are standard by following, for example, the considerations on control-to-state operators in [58], we provide these proofs here because we were not able to find good references for the specific Lipschitz, embedding,... constants of the example in this article. We believe, however, that they help make informed assessments of the possible quality and performance of our algorithmic approaches, although some of them may not be tight.

Lemma 3.4 *S is Lipschitz continuous on the set C as a function $L^1(0,1) \to U$.*

Proof Let $w_1, w_2 \in C$. Let $u_1 = S(w_1)$, $u_2 = S(w_2)$. We subtract the weak two formulations, insert a suitable zero, and deduce by means of Hölder's inequality the triangle inequality, (11), (12), and the nonexpansiveness of P_h:

$$\begin{aligned}
\|\nabla(u_1-u_2)\|_H^2 &= ((P_hw_1)(P_hu_1)-(P_hw_2)(P_hu_1), u_1-u_2)_H \\
&\quad + ((P_hw_2)(P_hu_1)-(P_hw_2)(P_hu_2), u_1-u_2)_H \\
&\le \|u_1-u_2\|_{L^\infty}\|u_1\|_{L^\infty}\|w_1-w_2\|_{L^1} + \max\{|w_\ell|,|w_u|\}\|u_1-u_2\|_H^2 \\
&\le \left(\frac{\|f\|_H}{4c_2(w_\ell,w_u)}\|\nabla(u_1-u_2)\|_H\|w_1-w_2\|_{L^1}\right. \\
&\quad \left. + \frac{\max\{|w_\ell|,|w_u|\}}{\pi^2}\|\nabla(u_1-u_2)\|_H^2\right).
\end{aligned}$$

An equivalent reformulation of this estimate gives

$$\|\nabla(u_1-u_2)\|_H \le \underbrace{\frac{\|f\|_H}{4c_2(w_\ell,w_u)^2}}_{=:L_S}\|w_1-w_2\|_{L^1}. \tag{29}$$

□

Let u solve (19) for $w \in C$. For a given direction $s \in L^1(0,1)$ so that $w+s \in C$ and with $u = S(w)$, we denote the (weak) solution to

$$\int_0^1 \nabla q \nabla v + \int_0^1 (P_h q)(P_h w)v = -\int_0^1 (P_h u)(P_h s)v \quad \text{for all } v \in U$$

by q. We obtain that $q \in U$ is uniquely defined because $s \in C - w$ implies $L^\infty(0,1)$-bounds on s, namely, $w_\ell - w_u \le s \le w_u - w_\ell$. This gives $\|(P_h s)(P_h u)\|_H \le \|w_u - w_\ell|\|u\|_H$ and a similar analysis as for (19) applies.

Lemma 3.5 *The operator $S : C \to U$ is Fréchet differentiable with respect to $L^1(0,1)$ in C. Let q be as above for a given s such that $w+s \in C$. Then $S'(w)s = q$. For a fixed $s \in C$, the mapping $w \mapsto S'(w)s$ is Lipschitz continuous in C with respect to $L^1(0,1)$.*

Proof Clearly, the mapping $s \mapsto q$ is a bounded linear operator. We need to show

$$\lim_{\substack{\|s\|_{L^1} \to 0 \\ s \in C}} \frac{\|S(w+s) + S(w) - q\|_U}{\|s\|_{L^1}} \to 0.$$

Let $u_s = S(w+s)$, $u = S(w)$. Let $d := u_s - u - q$. Then, we deduce with the considerations above and Lemma 3.4 and the Lipschitz constant L_S from (29) that

$$\begin{aligned}\|\nabla d\|_H^2 &= \big(P_h(u-u_s)(P_h s), d\big)_H - \big((P_h d)(P_h w), d\big)_H \\ &\le \frac{L_S}{2}\|\nabla d\|_H \|s\|_{L^1}\|s\|_{L^1} + \frac{\max\{|w_\ell|, |w_u|\}}{\pi^2}\|\nabla d\|_H^2,\end{aligned}$$

which yields

$$\|\nabla d\|_H \le \frac{L_S}{2c_2(w_\ell, w_u)}\|s\|_{L^1}^2.$$

Since $\|\cdot\|_U$ is equivalent to $\|\nabla \cdot\|_H$ with the embedding (11), this proves the claim.

For the Lipschitz continuity, let $w_1, w_2 \in C$, $u_1 = S(w_1)$, $u_2 = S(w_2)$, and $q_1 = S'(w_1)s$, $q = S'(w_2)s$. Then, we obtain with $r = q_1 - q_2$ that

$$\begin{aligned}\|\nabla r\|_H^2 &= \big(P_h(u_2-u_1)(P_h s), r\big)_H - \big((P_h r)(P_h w), r\big)_H - \big((P_h q_2)P_h(w_2-w_1), r\big)_H \\ &\le \frac{L_S}{2}\|w_1 - w_2\|_{L^1}\|s\|_{L^1}\|\nabla r\|_H + \frac{\max\{|w_\ell|, |w_u|\}}{\pi^2}\|\nabla r\|_H^2 \\ &\quad + \frac{1}{2}\|\nabla q_2\|_H \|w_1 - w_2\|_{L^1}\|\nabla r\|_H.\end{aligned}$$

With the arguments from the preceding subsections, we obtain

$$\|\nabla q_2\|_H \le \frac{1}{c_2(w_\ell, w_u)}\|s\|_{L^\infty}\|u_2\|_H \le \frac{|w_u - w_\ell|}{\pi c_2(w_\ell, w_u)^2}\|f\|_H,$$

so that we obtain the estimate

$$\|\nabla r\|_H \leq \underbrace{\frac{|w_u - w_\ell|}{2c_2(w_\ell, w_u)}\left(L_S + \frac{\|f\|_H}{\pi c_2(w_\ell, w_u)^2}\right)}_{=:L_{S'}} \|w_1 - w_2\|_{L^1}.$$

□

For given directions $\phi, \psi \in L^1(0,1)$ so that $w+\phi, w+\psi \in C$, we denote the (weak) solution to

$$\begin{aligned}\int_0^1 \nabla\xi\nabla v + \int_0^1 (P_h\xi)(P_h w)v \\ = -\int_0^1 P_h(S'(w)\phi)(P_h\psi)v - \int_0^1 P_h(S'(w)\psi)(P_h\phi)v \quad \text{for all } v \in U \end{aligned} \tag{30}$$

by ξ. As above, we obtain that $\xi \in U$ is uniquely defined because $\phi, \psi \in C - w$ imply $L^\infty(0,1)$-bounds on ϕ, ψ, that is, $w_\ell - w_u \leq \phi \leq w_u - w_\ell$ and $w_\ell - w_u \leq \psi \leq w_u - w_\ell$. This implies bounds on the lower-order order terms and a similar a similar analysis as for (19) applies.

Proposition 3.6 *Let $\phi \in C - w$ be given. The operator $w \mapsto S'(w)\phi$ is continuously Fréchet differentiable with respect to $L^1(0,1)$ in C. Its derivative in direction $\psi \in C - w$ is given by $S''(w)[\psi, \phi] = \xi$, where ξ is the solution to (30) above. Under these assumptions, it holds that*

$$\|S''(w)[\psi, \phi]\|_H \leq \kappa \|\psi\|_{L^1}\|\phi\|_{L^1}$$

for some $\kappa > 0$.

Proof For $\phi \in C - w$ and $\psi \in C - w$, let $q := S'(w)\phi$, $q_\psi := S'(w+\psi)\phi$, $\tilde{q}_\psi := S'(w)\psi$, $u_\psi = S(w+\psi)$, $u = S(w)$, ξ be the solution to (30). Let $r := q_\psi - q - \xi$. Then, we obtain from Hölder's inequality and the embedding constants (11), (12)

$$\begin{aligned}\|\nabla r\|_H^2 &= \big(P_h(u + \tilde{q}_\psi - u_\psi)(P_h\phi), r\big)_H + \big(P_h(q_\psi - q)(P_h\psi), r\big)_H - \big((P_h r)(P_h w), r\big)_H \\ &\leq \frac{1}{4}\|\nabla(u_\psi - u - \tilde{q}_\psi)\|_H\|\phi\|_{L^1}\|\nabla r\|_H + \frac{1}{4}\|\nabla(q_\psi - q)\|_H\|\psi\|_{L^1}\|\nabla r\|_H \\ &\quad + \frac{\max\{|w_\ell|, |w_u|\}}{\pi^2}\|\nabla r\|_H^2.\end{aligned}$$

Using the constants defined above and the estimates from the proof of Lemma 3.5, we obtain

$$4c_2(w_\ell, w_u)\|\nabla r\|_H \leq \frac{L_S}{2c_2(w_\ell, w_u)}\|\psi\|_{L^1}^2\|\phi\|_{L^1} + L_{S'}\|\psi\|_{L^1}^2.$$

Consequently, we obtain

$$\lim_{\|\psi\|_{L^1} \underset{\psi\in C}{\to} 0} \frac{\|S(w+\psi) + S(w) - \xi\|_U}{\|\psi\|_{L^1}} \to 0,$$

which proves the desired differentiability.

Testing (30) with $\xi = S''(w)[\psi, \phi]$ and using (11) and similar computations as in the previous arguments, we obtain

$$\|\nabla\xi\|_H^2 \leq \frac{\max\{|w_\ell|, |w_u|\}}{\pi^2}\|\nabla\xi\|_H^2 + \frac{\|f\|_H}{2c_2(w_\ell, w_u)^2}\|\psi\|_{L^1}\|\phi\|_{L^1}\|\nabla\xi\|_H$$

and in turn

$$\|\xi\|_H \leq \underbrace{\frac{\|f\|_H}{2\pi c_2(w_\ell, w_u)^3}}_{=:\kappa}\|\psi\|_{L^1}\|\phi\|_{L^1}.$$

The (Lipschitz) continuity of $w \mapsto S''(w)[\phi, \psi]$ can be shown with arguments similar to but more lengthy than that for the Lipschitz continuity of the mapping $w \mapsto S'(w)s$ in Lemma 3.5. □

4 Computational experiments for a 1D example

We start from the example given in Sect. 3 and provide computational experiments that will serve several purposes. We compare the approximate lower bounds obtained with approximate McCormick relaxations before and after applying the OBBT procedure to a feasible primal point obtained with a gradient-based optimization of (OCP) and a primal point obtained with the SLIP algorithm from [29, 34] in the presence of an additional integrality restriction on w. We also compare the bounds with lower bounds on (OCP) that can be obtained with much less effort than the OBBT procedure.

We analyze how the bounds we obtain from (McC_{hh}) behave when the partitions of the domain assumed in Assumption 2.7 1 are uniformly refined, thereby bringing the locally averaged McCormick relaxations closer to a pointwise limit. Moreover, we apply the OBBT procedure Algorithm 1 to assess whether and how much the OBBT procedure tightens the bounds for a prescribed termination tolerance as well as to assess its computational effort.

This section is organized as follows. We describe our experiments in Sect. 4.1. In Sect. 4.2 we describe how we approximate (9) and (19) with a very fine finite-element discretization and use the resulting discretized equations as state equations for (OCP), (McC_h), and (McC_{hh}). We then give details on the practical implementation of the OBBT procedure Algorithm 1 in Sect. 4.3. The results are provided in Sect. 4.4.

4.1 Experiment description

We consider the instance of (OCP) from Sect. 3 with the following choices of the parameters and fixed inputs. The source term f of the PDEs (9) and (19) is chosen as $f(x) := 6$ for all $x \in [0, 1]$. Regarding the bounds, we set $w_\ell := -4$, $w_u := 4$. For the penalty parameter, we choose $\alpha := 2.5 \cdot 10^{-4}$. For the function u_d in the tracking-type

objective (see (16), we choose

$$\begin{aligned} u_d(x) \coloneqq 3\Big(&1.5x(1-x)\chi_{[0,0.25]}(x) + 1.5x(1-x)\chi_{[0.75,1]} \\ &+ (0.28125 + 3(x-0.25))\chi_{(0.25,0.4]}(x) \\ &+ (0.73125 - 3(x-0.6))\chi_{[0.6,0.75)}(x) + 2\chi_{(0.4,0.6)}(x)\Big). \end{aligned}$$

Our main object of interest is to assess the quality of the approximate McCormick relaxations, the effect of the OBBT procedure Algorithm 1. To this end, we perform the following computations.

- We execute a local gradient-based NLP solver (L-BFGS-B) using Scipy's implementation [60] in order to obtain a stationary point for (OCP) and thus a low upper bound on $m_{(\mathrm{OCP})}$. In order to be able to apply this method to the nondifferentiable total variation seminorm, the latter is smoothed by using an overestimating Huber regularization for the absolute value in the integrand with smoothing parameter 10^{-3}.
- We add the additional integrality constraint $w(x) \in \mathbb{Z}$ and execute the SLIP algorithm [29, 36] using the subproblem solver described in [55] in order to obtain a low upper bound for the integrality-constrained version of (OCP).
- We use monotonicity properties of the PDE that we do not exploit in our computations elsewhere in order to compute a pointwise McCormick envelope with bounds u_ℓ, u_u that are as tight as possible in order to assess the quality of the approximate McCormick relaxations a posteriori.
- We compute approximate McCormick relaxations, that is, solutions to (McC_{hh}), with and without bound tightening (and subsequent improvement of the involved constants) as well as the induced lower bounds on (OCP) for decreasing values of h using the estimates from Sect. 3.4.

All experiments were executed on a laptop computer with Intel (TM) i7-11850 H CPU clocked at 2.5 GHz and 64 GB main memory.

4.2 Baseline PDE discretization with Ritz–Galerkin ansatz

We consider conforming finite elements and thus a finite-dimensional subspace $U_N \subset U$ with dimension $N \in \mathbb{N}$. By means of the Lax–Milgram lemma, we obtain that there exists a unique solution $u_N \in U_N$, the Ritz approximation, of the weak formulation on U_N of the PDEs (9):

$$\int_0^1 \nabla u_N \nabla v_N + \int_0^1 w u_N v_N = \int_0^1 f v_N \quad \text{for all } v_N \in U_N. \tag{31}$$

and (19)

$$\int_0^1 \nabla u_N \nabla v_N + \int_0^1 (P_h w)(P_h u_N) v_N = \int_0^1 f v_N \quad \text{for all } v_N \in U_N. \tag{32}$$

The solutions to (31) satisfy

$$\|\nabla u_N\|_H \leq \frac{1}{c_1(w_\ell)}\|f\|_H,$$

and the solutions to (32) satisfy

$$\|\nabla u_N\|_H \leq \frac{1}{c_2(w_\ell, w_u)}\|f\|_H.$$

The whole analysis from Sects. 2 and 3 and, in particular, all of the estimates derived for our example PDE in Sect. 3 still hold when discretizing the variational form of the PDE with a conforming finite-element setting and using a piecewise constant control ansatz for w on the same or a coarser grid. The only required change is to insert U_N as the state space for U. We use first-order Lagrange elements and a discretization of $(0, 1)$ into $N = 2,048$ intervals. Regarding the additional control variable z in the state equation $Au + z = f$ of the pointwise envelope in (McC), we also choose a piecewise constant ansatz on the 2,048 intervals, which introduces a small approximation error because a piecewise constant function cannot completely capture the product of a first-order Lagrange element and a piecewise constant function.

Consequently, in our implementation of (OCP) and (McC), we use following discretization that serves as a baseline and substitutes the infinite-dimensional setting in our experiments:

- discretization of u with first-order Lagrange elements on $N = 2,048$ intervals,
- discretization of w with piecewise constant functions on $N = 2,048$ intervals, and
- discretization of z with piecewise constant functions on $N = 2,048$ intervals.

Remark 4.1 In this *discretize-then-optimize* setting, it is possible to directly employ a branch-and-bound procedure on the fully discretized version of (OCP) when additional integrality restrictions are imposed on w such as $w(x) \in \mathbb{Z}$. Then (9) is just replaced by (31); the lower bounds are computed by means of McCormick relaxations using (32), and the upper bounds are, for example, computed by means of the SLIP algorithm [29, 34].

Remark 4.2 Since the problem considered in this section is defined on an interval, that is in one dimension, it is possible to reformulate it as an optimal control problem with a multiple point boundary-value problem as constraint and apply optimal control techniques like single- or multiple-shooting or collocation for discretization and solution, see the articles [5, 45] for overviews over such techniques. Due to our focus on applicability to PDEs, we have decided to use a finite-element discretization in this work.

4.3 Practical implementation of OBBT / Algorithm 1

In our implementation of Algorithm 1, we first minimize the lower bounds u_ℓ^i one by one along the order of the intervals $i = 1, \ldots, N_h$. Then we maximize the upper

bounds u_u^i one by one along the order of the intervals $i = 1, \ldots, N_h$. Then we repeat this process. We terminate when a complete run of minimization of the u_ℓ^i (or maximization of the u_u^i) does not produce a tightening of any of the bounds by more than a prescribed tolerance of 10^{-6}.

Moreover, in order to avoid numerical problems and in turn incorrect results, additional safeguarding was necessary. Specifically, when bounds are set to the computed value in Algorithm 1 ln. 5 or ln. 7, the value might be slightly too sharp because of the numerical precision of the subproblem solver for the OBBT problems. This, however, can cause the lower bounds to increase and the upper bounds to decrease to incorrect values in further tightenings and the feasible set even contracting to an empty set (when the lower and upper bounds cross each other). In order to avoid this situation, the value 10^{-7} is added to the upper bounds and subtracted from the lower bounds computed by the subproblem solver for the OBBT problems before they are assigned as new bounds and Algorithm 1 moves on to tighten the next bound.

Because Algorithm 1 requires valid initial bounds, we use the bounds that can be inferred from (15) and our choice of f. Specifically, we use the initial bounds $u_\ell^i = -5.0444$ and $u_u^i = 5.0444$ for all $i \in \{1, \ldots, N_h\}$.

4.4 Results

We first ran Scipy's implementation [60] of L-BFGS-B on (OCP), where we used the unique solvability of the (discretized) PDE constraint to integrate it into the objective and used adjoint calculus in order to compute the derivative. To use this gradient-based solver, we have smoothed the total variation seminorm using a Huber regularization with smoothing parameter 10^{-3}. We inserted the solution into the nonsmooth objective and obtained an upper bound of 8.3808×10^{-2} on the optimal solution to (OCP).

Then we added the integrality constraint $w(x) \in \mathbb{Z}$ to (OCP) and executed the SLIP algorithm [29, 36] with the subproblem solver from [55] to compute an upper bound on the optimal solution to (OCP) with the additional integrality constraint $w(x) \in \mathbb{Z}$. The resulting upper bound was 8.5551×10^{-2}.

Regarding the lower bounds, we first compute pointwise lower bounds for comparison. For our computational example, we compute the optimal objective value for the uniform bounds $u_\ell \equiv -5.0444$ and $u_u \equiv 5.0444$ that are enforced on the nodes of our first-order Lagrange elements. We first compute a lower bound by omitting the total variation term from the objective ($\alpha = 0$), which also gives us an overall lower bound on the tracking-type term in the objective. The resulting linearly constrained convex quadratic program is solved by using Gurobi [24], and the computed lower bound is 6.7701×10^{-2}. Then we consider (McC) with the same bounds but including the total variation term. Again, the resulting linearly constrained convex quadratic program is solved by using Gurobi [24], and the computed lower bound is 6.8649×10^{-2}. Moreover, we can characterize the tightest possible pointwise McCormick relaxation exactly. Specifically, the monotonicity properties of the Laplacian yield that the discretization of (9) attains its pointwise minimum $u_{\min}$ for the choice $w(x) = w_u = 4$ for all $x \in \Omega$ and its pointwise maximum $u_{\max}$ for the choice $w(x) = w_\ell = -4$ for all $x \in (0, 1)$. Consequently, we execute (McC) with the bounds $u_\ell = u_{\min}$ and

Table 1 Exact upper and lower bounds and relative gaps (ratio of the difference between upper and lower bound to the lower bound) for (OCP) and its counterpart with integrality restriction $w(x) \in \mathbb{Z}$

	Upper bounds		Lower bounds		
	SLIP ($w(x) \in \mathbb{Z}$)	L-BFGS-B	(McC) (tightest)	(McC)	(McC) ($\alpha = 0$)
Value	8.555×10^{-2}	8.381×10^{-2}	8.368×10^{-2}	6.865×10^{-2}	6.770×10^{-2}
Rel. gap (MINLP)			2.237×10^{-2}	2.462×10^{-1}	2.637×10^{-1}
Rel. gap (NLP)			1.542×10^{-3}	2.208×10^{-1}	2.379×10^{-1}

$u_u = u_{\max}$ to obtain the tightest lower bound the McCormick relaxation can possibly achieve. We obtain a higher optimal objective value of 8.3679×10^{-2}.

This means that if this perfect information on the bounds u_u and u_ℓ is available when solving (McC), the gap between upper and lower bounds gets reduced from 1.690×10^{-2} to 1.872×10^{-3} in the presence of the constraint $w(x) \in \mathbb{Z}$ (MINLP case) and from 1.516×10^{-2} to 1.290×10^{-4} in the continuous case. In both cases, the relative gap, which is generally computed as the difference between upper and lower bound divided by the lower bound, gets reduced by an order of magnitude. These results are tabulated in Table 1.

In the remaining experiments, we use these results as a baseline to compare them with the (approximate) lower bounds obtained by solving instances of (McC_{hh}). We executed our experiments on (McC_{hh}) on uniform partitions of the domain (0, 1) into $N_h \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$ intervals with the corresponding values $h = 2^{-N_h}$.

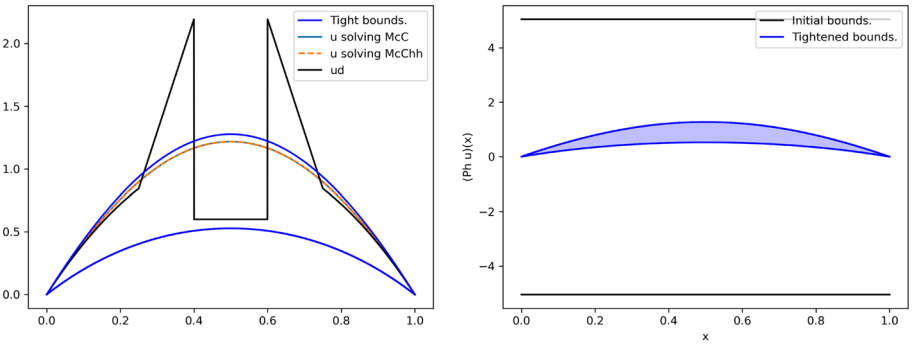
We now consider the approximate lower bounds that are obtained by solving (McC_{hh}) for the aforementioned values of h. In particular, we assess their approximation quality and the running times that are required to compute them by means of the OBBT procedure for the different values of h. To obtain a valid lower bound on (OCP), we need to subtract $c_{quad}h^2$ from the optimal objective value of (McC_{hh}); see (28). In practice, the values j_0, $\|\tilde{d}_u\|_H$, and $\|\tilde{d}_f\|_H$ are not known precisely because this would imply that tight pointwise McCormick relaxations have already been computed.

We therefore use two constants in our considerations that allow us to present a range where we expect that the a priori estimates lie when all of the other constants are known. The larger and thus more conservative constant is obtained by choosing $j_0 = 0$ and computing $\tilde{d}_u$ and $\tilde{d}_f$ with the conservative bounds $u_\ell \equiv -5.0444$ and $u_u \equiv 5.0444$. The smaller constant is computed by choosing $j_0 = 6.7701 \times 10^{-2}$ as well as the optimal bounds $u_\ell = u_{\min}$ and $u_u = u_{\max}$, which are of course not available in a realistic setting. In our setting, the conservative bound is almost 50 times larger than the tight one. Both values are given in Table 2.

For our example, the initial bounds that can be deduced from (15) are $u_\ell^i = -5.0444$ and $u_u^i = 5.0444$ for all $i \in \{1, \ldots, N_h\}$. After executing our implementation of

Table 2 Values for c_{quad} in (28)

conservative	tight
5.603×10^4	1.132×10^3



(a) Solutions to (McC) and (McC$_{hh}$).

(b) PDE solution envelope after OBBT.

Fig. 2 **a** Solutions to (McC) (cyan, solid) and (McC$_{hh}$) for $h = 2^{-10}$, (orange, dashed) with $u_{\min}$, $u_{\max}$ (dark blue, solid), and tracking function u_d (black, solid). **b** Initial (black) and tightened (blue) lower and upper bounds on $P_h u$ for solutions u to (19) for $h = 2^{-10}$. The area between the tightened lower and upper bounds, where the solution of the PDE can attain values, is colored with a less intense blue

Algorithm 1, the lower bounds and upper bounds are much closer to each other and reflect the structure of the possible PDE solutions; that is, they tend to zero toward the boundaries of the computational domain. For the finest computed case with $h = 2^{-10}$, $N_h = 1024$, and a termination tolerance of 10^{-6} for our implementation of the OBBT procedure, the lower bounds u_ℓ^i for $P_h u$ vary between 0 and 5.280×10^{-1} and the upper bounds u_u^i vary between 0 and 1.276, thereby giving a much tighter envelope in which $P_h u$ for the solutions to (19) can lie. Moreover, the optimal solution computed for (McC$_{hh}$) is close to the optimal solution computed for (McC) for fine values of h. Both findings are visualized in Fig. 2.

Generally, the increase in compute times for the bounds is worse than linear with decreasing values of h. The runtime for the OBBT procedure is 8.212 seconds for $h = 2^{-3}$ and increases to 1.080×10^4 seconds for $h = 2^{-10}$ for (McC$_{hh}$) compared with a pointwise bound computation with 2.142×10^4 seconds (mesh size 2^{-11}) for (McC). All running times of the OBBT algorithm are tabulated in Table 3.

Next, we assess the approximation quality of the bounds induced by the solutions to (McC$_{hh}$) with and without bound tightening. In both cases we observe that the infima converge numerically to their pointwise counterparts of (McC) that are tabulated in Table 1. When subtracting $c_{quad} h^2$ from each of these values for the conservative and tight choices of c_{quad} from Table 2, we obtain valid lower bounds and observe that only for the smallest grid with $h = 2^{-10}$ all of the lower bounds are positive and thus beat the trivial lower bound 0 that can be found by inspecting the objective. If the OBBT procedure is applied and a tight choice of c_{quad} is used, the lower bound

Table 3 Running times (in seconds) for the OBBT algorithm for different values of h on instances of (McC_{hh}) and (McC) (mesh size 2^{-11})

	(McC_{hh}) with $h =$								(McC)
	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}	
Time [s]	8.212	1.612×10^{1}	3.174×10^{1}	7.026×10^{1}	1.652×10^{2}	4.565×10^{2}	1.594×10^{3}	1.080×10^{4}	2.142×10^{4}

Table 4 Optimal objective values achieved for (McC_{hh}) and induced bounds for (OCP) using the estimate (28) with and without OBBT with conservative (c) and tight (t) estimate on c_{quad}

OBBT	h	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}
No	$m_{(\mathrm{McC}_{hh})}$	$+7.153\times10^{-2}$	$+7.028\times10^{-2}$	$+6.868\times10^{-2}$	$+6.867\times10^{-2}$	$+6.866\times10^{-2}$	$+6.865\times10^{-2}$	$+6.865\times10^{-2}$	$+6.865\times10^{-2}$
Yes	$m_{(\mathrm{McC}_{hh})}$	$+8.373\times10^{-2}$	$+8.370\times10^{-2}$	8.368×10^{-2}	$+8.368\times10^{-2}$	$+8.368\times10^{-2}$	$+8.368\times10^{-2}$	$+8.368\times10^{-2}$	$+8.368\times10^{-2}$
No	LB (28) (c)	$-8.753\times10^{+2}$	$-2.188\times10^{+2}$	$-5.464\times10^{+1}$	$-1.361\times10^{+1}$	-3.351	-7.862×10^{-1}	-1.451×10^{-1}	$\mathbf{1.522\times10^{-2}}$
Yes	LB (28) (c)	$-8.753\times10^{+2}$	$-2.188\times10^{+2}$	$-5.463\times10^{+1}$	$-1.360\times10^{+1}$	-3.336	-7.712×10^{-1}	-1.300×10^{-1}	$\mathbf{+3.025\times10^{-2}}$
No	LB (28) (t)	$-1.761\times10^{+1}$	-4.351	-1.037	-2.076×10^{-1}	-4.106×10^{-4}	$\mathbf{+5.139\times10^{-2}}$	$\mathbf{+6.433\times10^{-2}}$	$\mathbf{+6.757\times10^{-2}}$
Yes	LB (28) (t)	$-1.760\times10^{+1}$	-4.337	-1.022	-1.926×10^{-1}	$\mathbf{+1.460\times10^{-2}}$	$\mathbf{+6.641\times10^{-2}}$	$\mathbf{+7.936\times10^{-2}}$	$\mathbf{+8.260\times10^{-2}}$

The bounds that are able to beat the trivial bound 0 that can be directly inferred from the nonnegative objective terms are bold

Table 5 Difference and relative difference between $m_{(\mathrm{McC})}$ for $u_\ell = u_{\min}$, $u_u = u_{\max}$ and $m_{(\mathrm{McC}_{hh})}$ for u_ℓ, u_u computed using the OBBT procedure

h	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}
$\lvert m_{(\mathrm{McC}_{hh})} - m_{(\mathrm{McC})}\rvert$	$+5.106\times10^{-5}$	$+2.2474\times10^{-5}$	$+2.292\times10^{-6}$	$+5.740\times10^{-7}$	$+1.450\times10^{-7}$	$+3.820\times10^{-8}$	$+1.176\times10^{-8}$	$+3.436\times10^{-8}$
$\frac{\lvert m_{(\mathrm{McC}_{hh})} - m_{(\mathrm{McC})}\rvert}{m_{(\mathrm{McC})}}$	$+6.102\times10^{-4}$	$+2.686\times10^{-4}$	$+2.738\times10^{-5}$	$+6.859\times10^{-6}$	$+1.733\times10^{-6}$	$+4.565\times10^{-7}$	$+1.405\times10^{-7}$	$+4.106\times10^{-7}$

is positive for the choices of $h \leq 2^{-7}$, and the computed bound beats the pointwise bound without OBBT for $h = 2^{-9}$ and $h = 2^{-10}$, thereby yielding the second and third best lower bounds of all computed lower bounds. All computed values are given in Table 4.

We assess the quality of the approximate bounds $m_{(\mathrm{McC}_{hh})}$ after optimization-based bound tightening with respect to $m_{(\mathrm{McC})}$. We observe that the difference as well as the relative difference becomes very small for small values of h, in particular several magnitudes smaller than the error margin of the a priori estimate. The obtained values are given in Table 5.

5 Computational experiments for an example in 2D

While an analysis that verifies all of our assumptions for a PDE on a multi-dimensional domain is beyond the scope of this article, we still present an example for which we are certain that we can verify them by means of the techniques mentioned in Sect. 2.2. Moreover, the techniques from Sect. 3.4 can be combined with regularity estimates from [22] to derive a priori estimates as well. Specifically, we consider a convection-diffusion boundary value problem that leans on the setting in [3].

This section is organized as follows. We describe our experiments in Sect. 5.1. In Sect. 5.2 we describe how we approximate (33) and its counterpart with u and w

replaced by $P_h u$ and $P_h w$ with a finite-element discretization and use the resulting discretized equations as state equations for (OCP), (McC$_h$), and (McC$_{hh}$). We give details on the implementation of the OBBT procedure in Sect. 4.3. The results are provided in Sect. 4.4.

5.1 Experiment description

Regarding the state equation, we set $\Omega = (0, 1)^2$ and $W = \{0, 1\}$. Let w with $w(x) \in W$ a.e. be given. Then the state vector u is given by the solution to

$$\begin{aligned} -\varepsilon \Delta u + c_1 \cdot \nabla u + c_2 u w &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \{0,1\} \times (0,1) \cup ((0,0.25) \cup (0.75,1)) \times \{1\} \\ u &= \sin(2\pi(x_1 - 0.25)) \quad \text{on } (0.25, 0.75) \times \{1\} \\ \partial_n u &= 0 \quad \text{on } (0,1) \times \{0\}, \end{aligned} \tag{33}$$

where $\varepsilon = 0.04$, $c_2 = 4$, $c_1(x) = (\sin(\pi x_1)\cos(2\pi x_2))^T$ for $x \in \Omega$, $f(x) = \sin(2\pi x_1 + 2\pi x_2) + 3$ for $x \in \Omega$. Let S be the control-to-state operator of (33). We choose the objective

$$j(u) := \frac{1}{2}\|u - u_d\|_{L^2}^2,$$

where we compute u_d as follows. We replace the coefficient c_1 in (33) by $\tilde{c}_1(x) = (-x_2\ 2x_1)^T$ and solve this modified boundary value problem for the control $w = 2.5\chi_A - 4(x_1 - 0.35)^3\chi_A - 6(x_2 - 0.35)^3\chi_B$, where we have $A = (0, 0.35)^2$ and $B = \Omega \backslash (0, 0.35)^2$.

Since our main object of interest is to assess the quality of the approximate McCormick relaxations and the effect of the OBBT procedure, we perform the following computations, where h denotes the mesh size.

- We execute a local gradient-based NLP solver in order to obtain a stationary point for (OCP) and thus a low upper bound on $m_{(\text{OCP})}$.
- We add the additional integrality constraint $w(x) \in \mathbb{Z}$ and execute the SLIP algorithm [29, 34] to obtain a low upper bound for the integrality-constrained version of (OCP).
- Similar to the 1D experiment, we compute a pointwise McCormick envelope with bounds u_ℓ, u_u using OBBT on the nodal basis of the finest discretization we have available. This then serves as a baseline in order to assess the quality of the approximate McCormick relaxations.
- We compute approximate McCormick relaxations, that is, solutions to (McC$_{hh}$), with and without bound tightening for decreasing values of h, where h is the mesh size of a uniform grid of squares.

We have carried out the experiments on a node of the Linux HPC cluster LiDO3 with two AMD EPYC 7542 32-Core CPUs and 64 GB RAM.

5.2 Baseline PDE discretization with Ritz–Galerkin ansatz

We decompose the domain into a 48×48 grid of squares, which consist of four triangles each, on which the state vector of (33) is defined. We consider conforming finite elements and thus a finite-dimensional subspace $U_N \subset U$ with dimension $N \in \mathbb{N}$, specifically, we use first-order Lagrange elements. The term pointwise envelope refers to the pointwise constraint bounds are enforced on the nodes of the first-order Lagrange ansatz. Regarding the additional control variable z in the state equation of the pointwise envelope in (McC), we also choose a piecewise constant ansatz on all triangles similar to the one-dimensional case.

Consequently, in our implementation of (OCP) and (McC), we use following discretization that serves as a baseline and substitutes the infinite-dimensional setting in our experiments:

- discretization of u with first-order Lagrange elements on $4 \times 48 \times 48$ triangles,
- discretization of w with piecewise constant functions on 48×48 squares, and
- discretization of z with piecewise constant functions on $4 \times 48 \times 48$ triangles.

5.3 Practical implementation of OBBT

Our implementation Algorithm 1 differs from the one described in Sect. 4.3 to be able to parallelize over bound computations due to the very long compute times. We initialize the lower bounds u_ℓ both for the pointwise and the locally averaged averaged McCormick envelopes by -10^3 and the upper bounds u_u by 10^3, which are high enough for this example to safely assume that the true bounds lie within them.

For a given set of bounds, we minimize the lower bounds u_ℓ^i and maximize the upper bounds u_u^i for all nodes/grid cells in parallel. Then we use this set of bounds as the new set of bounds and compute another round of tightened bounds. We repeat this seven times so that eight rounds of bound tightening are executed in total. Again, in order to avoid numerical problems and obtain correct results, we applied the additional safeguarding from Sect. 4.3 but note that a coarser safety tolerance of 10^{-2} was necessary here and we needed to set the parameter *BarHomogeneous* to one in Gurobi to prevent incorrect identification of infeasibility in a few cases that occurred in the eighth round of OBBT on the finest discretization.

5.4 Results

We first ran the local gradient-based NLP solver, in which we employ an anisotropic discretization of the total variation seminorm based on the piecewise constant control function ansatz, see also Appendix B in [34] and section 2 in [35]. In the trust-region method, we use a linear model for the first part of the objective and handle the trust-region seminorm as in [34] so that our trust-region subproblems are linear programs after discretization. The resulting objective value was +3.2337.

Then we added the integrality constraint $w(x) \in \mathbb{Z}$ to (OCP) and executed the SLIP algorithm [29, 34] to compute an upper bound on the optimal solution to (OCP)

Table 6 Upper and lower bounds and relative gaps (ratio of the difference between upper and lower bound to the lower bound) for (OCP) and its counterpart with integrality restriction $w(x) \in \mathbb{Z}$

	Upper bounds		Lower bounds	
	SLIP ($w(x) \in \mathbb{Z}$)	NLP solver	(McC) (OBBT)	(McC) (initial)
Value	+3.2340	+3.2337	+3.1347	$+3.0255 \times 10^{-11}$
Rel. gap (MINLP)			$+3.1689 \times 10^{-2}$	$+1.0689 \times 10^{+11}$
Rel. gap (NLP)			$+3.1596 \times 10^{-2}$	$+1.0688 \times 10^{+11}$

with the additional integrality constraint $w(x) \in \mathbb{Z}$. The resulting upper bound was +3.2340.

Regarding the lower bounds, we first compute pointwise lower bounds using OBBT as described above. Then we solve (McC) with these bounds using Gurobi [24] as well as with the initial bounds (-10^3 for lower and 10^3 for upper bounds). The computed lower bound for the initial bounds is $+3.025510^{-11}$ and the computed lower bound for the tightened bounds +3.1347. These results are tabulated in Table 6.

In the remaining experiments, we use these results as a baseline to compare them with the (approximate) lower bounds obtained by solving instances of (McC_{hh}). We executed our experiments on (McC_{hh}) on uniform partitions of the domain Ω into $N_h \in \{3 \times 3, 6 \times 6, 12 \times 12, 24 \times 24, 48 \times 48\}$ squares with mesh sizes $h = \{3^{-1}, 3^{-1}2^{-1}, 3^{-1}2^{-2}, 3^{-1}2^{-3}, 3^{-1}2^{-4}\}$.

We now consider the approximate lower bounds that are obtained by solving (McC_{hh}) for the aforementioned values of h. In particular, we assess their approximation quality and the running times that are required to compute them by means of the OBBT procedure for the different values of h. To obtain a valid lower bound on (OCP), we need to subtract an a priori bound. Since the a priori analysis for this PDE is beyond of this work, we omit this step but note that all bounds are already valid lower bounds (see below).

Generally, the increase in compute times for the bounds is worse than linear with decreasing values of h. The runtime for the OBBT procedure is $+2.800 \times 10^{+1}$ seconds for $h = 3^{-1}$ and increases to $+9.552 \times 10^{+4}$ seconds for $h = 3^{-1}2^{-4}$ for (McC_{hh}) compared with a pointwise bound computation with $+4.008 \times 10^{+5}$ seconds for (McC) (mesh size $h = 3^{-1}2^{-5}$).

Next, we assess the approximation quality of the bounds induced by the solutions to (McC_{hh}) with and without bound tightening. After 8 rounds of OBBT, the quality of the bounds does not differ much for $h \leq 3^{-1}2^{-1}$ and are close to the pointwise/baseline counterpart of (McC) so that a similar bound quality can be achieved with 127 s of compute time as with 95522 s and 400780 s of compute time. Since they are also always lower, they are actually true lower bounds for the baseline. Since we have used very conservative initial guesses to prevent initializing from bounds that cut off feasible points, the quality without bound tightening is extremely bad. It is beneficial to make several rounds of bound tightening as is demonstrated by the quality improvement of

Table 7 Running times (in seconds) for the OBBT algorithm and optimal objective values for different values of h of (McC_{hh}) and (McC) (mesh size $h = 3^{-1}2^{-5}$)

	(McC_{hh}) with $h =$ (McC)					
	3^{-1}	$3^{-1}2^{-1}$	$3^{-1}2^{-2}$	$3^{-1}2^{-3}$	$3^{-1}2^{-4}$	
Time (8 OBBT rounds) [s]	$+2.8 \times 10^{+1}$	$+1.27 \times 10^{+2}$	$+7.9 \times 10^{+2}$	$+6.489 \times 10^{+3}$	$+9.5522 \times 10^{+14}$	$+4.0078 \times 10^{+5}$
Objective (8 OBBT rounds)	+3.0002	+3.1314	+3.1311	+3.1301	+3.1298	+3.1347
Objective (7 OBBT rounds)	+2.5795	+2.7020	+2.6967	+2.6923	+2.6912	+2.6999
Objective (0 OBBT rounds)	+1.2734	$+4.8655 \times 10^{-2}$	$+6.5599 \times 10^{-4}$	$+2.7435 \times 10^{-6}$	$+8.3494 \times 10^{-9}$	$+3.0255 \times 10^{-11}$

the bounds from the 7th to the 8th round. All running times of the OBBT algorithm and the induced achieved optimal objective values for (McC_{hh}) and (McC) are tabulated in Table 7.

6 Conclusion

We show how to replace nonlinearities that are given in a pointwise fashion in the PDE of an (integer) optimal control problem on the example of bilinear terms by analyzing the idea of McCormick envelopes and relaxations in the infinite-dimensional setting. To keep the computational effort manageable, in particular with respect to a bound-tightening procedure that tightens the convex relaxations, we introduce a two-level approximation scheme by means of a grid that decomposes the computational domain, on which the inequalities that yield the convex relaxation are averaged.

Our computational experiments for the 1D and the 2D example validate that the computational effort of such a procedure can indeed be significantly reduced by using coarser grids (two to three orders of magnitude for the 2D example). While an a posteriori comparison with the baseline solution shows that the approximation quality is very good, the a priori approximation quality of the lower bound only improves over trivial bounds when the mesh size is already quite fine and the involved constants can be estimated well. For the one-dimensional test case, the OBBT procedure for $h = 2^{-9}$ yields bounds with (relative) gaps to the upper bounds of +6.338 % and +8.549 % in the NLP and MINLP case compared with theoretically optimal bounds 0.1542 % and +2.237 % at +7.445 % of the computational cost for the pointwise OBBT procedure.

Consequently, in order to use the McCormick relaxations together with a priori estimates in a branch-and-bound procedure as sketched in Remark 4.1, it is necessary to obtain estimates on $\|u - u_h\|_H$ of higher order, and a deliberate analysis of the underlying PDE and its discretization are crucial when applying this method. Similarly, good estimates for the constant L_u in (6) are important, too.

As mentioned before, we have used a priori error estimates in our analysis so far. Because the actual optimal objective values for (McC_{hh}) were very close to a true lower bound on (OCP) for relatively large values of h (coarser local averaging) both in 1D and 2D, we are convinced that the analysis and integration of *a posteriori error estimates* into the procedure is key for future research to be able to use relatively coarse grids for (McC_{hh}) and scale this methodology.

One issue that might arise in practice, in particular for PDEs defined on multidimensional domains, is that L^∞-bounds as asserted in (12) for our example might not readily be available. However, we believe that additional interior regularity (see, e.g., Theorem 9.51 in [46]) and nonstandard regularity theory (see, e.g., [23]) can help establish such bounds.

Although this goes significantly beyond the scope of this article, we note that the McCormick relaxations suggest defining a branch-and-bound algorithm in function space in the spirit of the recent article [7], where—depending on current fixations and the overlapping of the approximate upper and lower bounds—one refines the mesh that is used for the locally averaged McCormick relaxations and the control ansatz adaptively until an acceptable gap is reached.

For more general nonlinearities than uw, it may be possible to combine the local averaging on the computational domain with the successive refinement procedure (*nested intervals*) of the parameter space for factorable functions that is analyzed in [54], see in particular section 7 therein for an application to obtain relaxations of ODEs.

A Auxiliary results

Lemma A.1 *Let $f^n \to f$ in $L^1(\Omega)$. Let $g^n \rightharpoonup g$ in $L^1(\Omega)$. If $f^n \leq g^n$ (or $f^n \geq g^n$) holds pointwise a.e. for all $n \in \mathbb{N}$. Then $f \leq g$ (or $f \geq g$, respectively).*

Proof We prove only the case for the $\leq$-inequalities because the other case follows by symmetry of the argument. By way of contradiction, we assume that there exist a measurable set $A \subset \Omega$, $\varepsilon_1 > 0$, and $\varepsilon_2 > 0$ such that $|A| > \varepsilon_1$ and

$$f(x) > g(x) + \varepsilon_2 \text{ for a.a. } x \in A.$$

Egorov's theorem gives that there exists a measurable set $B \subset A$ and $n_0 \in \mathbb{N}$ such that $|B| > \frac{\varepsilon_1}{2}$, and for all $n \geq n_0$, we obtain

$$f^n(x) > g(x) + \frac{\varepsilon_2}{2} \text{ for a.a. } x \in B.$$

Integrating over B yields

$$\int_B f^n(x) > \int_B g(x) + |B| \frac{\varepsilon_2}{2}.$$

for all $n \geq n_0$. Because $g^n \rightharpoonup g$, there exists $n_1 \geq n_0$ such that

$$\int_B f^n(x) > \int_B g^n(x) + |B|\frac{\varepsilon_2}{4}$$

holds for all $n \geq n_1$. Consequently, there must exist a measurable subset $C \subset B$ of strictly positive measure that $f^n(x) > g^n(x)$ holds for $x \in C$, which contradicts the assumption that $f^n \leq g^n$ holds pointwise a.e. □

Acknowledgements The authors thank Joachim Stöckler (TU Dortmund) for the pointer to the proof of (12) in [50]. The authors thank Mariia Pokotylo (TU Dortmund) as well as two anonymous referees for helpful comments on the manuscript. The authors gratefully acknowledge computing time on the LiDO3 HPC cluster at TU Dortmund, partially funded in the Large-Scale Equipment 796 Initiative by the Deutsche Forschungsgemeinschaft (DFG) as project 271512359. Paul Manns acknowledges funding by Deutsche Forschungsgemeinschaft (DFG) under Project No. 540198933. This work was also supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through the Advanced Computing (SciDAC) Program through the FASTMath Institute under Contract No. DE-AC02-06CH11357.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was financially supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through the Advanced Computing (SciDAC) Program through the FASTMath Institute under Contract No. DE-AC02-06CH11357.
Paul Manns is partially funded by Deutsche Forschungsgemeinschaft (DFG) under Project Nos. 515118017 and 540198933.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Ambrosio, L., Fusco, N., Pallara, D.: Functions of Bounded Variation and Free Discontinuity Problems. Oxford Mathematical Monographs, vol. 254. Clarendon Press, Oxford (2000)
2. Azunre, P.: Bounding the solutions of parametric weakly coupled second-order semilinear parabolic partial differential equations. Optimal Control Appl. Methods **38**(4), 618–633 (2017)
3. Baraldi, R., Manns, P.: Domain decomposition for integer optimal control with total variation regularization. arXiv preprint arXiv:2410.15672 (2024)
4. Bendsoe, M.P., Sigmund, O.: Topology Optimization: Theory, Methods, and Applications. Springer, Berlin (2013)
5. Betts, J.T.: Survey of numerical methods for trajectory optimization. J. Guid. Control. Dyn. **21**(2), 193–207 (1998)
6. Buchheim, C.: Compact extended formulations for binary optimal control problems. arXiv preprint arXiv:2401.03942 (2024)
7. Buchheim, C., Grütering, A., Meyer, C.: Parabolic optimal control problems with combinatorial switching constraints–Part III: Branch-and-bound algorithm. arXiv preprint arXiv:2401.10018 (2024)

8. Buchheim, C., Grütering, A., Meyer, C.: Parabolic optimal control problems with combinatorial switching constraints, part I: convex relaxations. SIAM J. Optim. **34**(2), 1187–1205 (2024)
9. Buchheim, C., Grütering, A., Meyer, C.: Parabolic optimal control problems with combinatorial switching constraints, part II: outer approximation algorithm. SIAM J. Optim. **34**(2), 1295–1315 (2024)
10. Buchheim, C., Hügging, M.: Bounded variation in binary sequences. In: International Symposium on Combinatorial Optimization, pp. 64–75. Springer, Berlin (2022)
11. Burger, M., Dong, Y., Hintermüller, M.: Exact relaxation for classes of minimization problems with binary constraints. arXiv preprint arXiv:1210.7507 (2012)
12. Bynum, M., Castillo, A., Watson, J.-P., Laird, C.D.: Tightening McCormick relaxations toward global solution of the ACOPF problem. IEEE Trans. Power Syst. **34**(1), 814–817 (2018)
13. Chambolle, A., Darbon, J.: A parametric maximum flow approach for discrete total variation regularization. In: Theory and Practice, Image Processing and Analysis with Graphs (2012)
14. Chambolle, A., Pock, T.: Approximating the total variation with finite differences or finite elements. In: Handbook of Numerical Analysis, vol. 22, pp. 383–417. Elsevier, New York (2021)
15. Chambolle, A., Tan, P., Vaiter, S.: Accelerated alternating descent methods for Dykstra-like problems. J. Math. Imaging Vis. **59**, 481–497 (2017)
16. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. SIAM, Philadelphia (2002)
17. Coffrin, C., Hijazi, H.L., Van Hentenryck, P.: Strengthening convex relaxations with bound tightening for power network optimization. In: International Conference on Principles and Practice of Constraint Programming, pp. 39–57. Springer, Berlin (2015)
18. Cristinelli, G., Iglesias, J., Walter, D.: Conditional gradients for total variation regularization with PDE constraints: a graph cuts approach. arXiv preprint arXiv:2310.19777 (2023)
19. Dal Maso, G.: An Introduction to Γ-Convergence, vol. 8. Springer, Berlin (2012)
20. D'Ambrosio, C., Lodi, A., Wiese, S., Bragalli, C.: Mathematical programming techniques in water network optimization. Eur. J. Oper. Res. **243**(3), 774–788 (2015)
21. Gleixner, A.M., Berthold, T., Müller, B., Weltge, S.: Three enhancements for optimization-based bound tightening. J. Global Optim. **67**, 731–757 (2017)
22. Grisvard, P.: Elliptic Problems in Nonsmooth Domains. SIAM, Philadelphia (2011)
23. Gröger, K.: A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations. Math. Ann. **283**, 679–687 (1989)
24. Gurobi Optimization, LLC.: Gurobi Optimizer Reference Manual (2020)
25. Haslinger, J., Mäkinen, R.A.E.: On a topology optimization problem governed by two-dimensional Helmholtz equation. Comput. Optim. Appl. **62**, 517–544 (2015)
26. Hochbaum, D.S.: An efficient algorithm for image segmentation, Markov random fields and related problems. J. ACM (JACM) **48**(4), 686–701 (2001)
27. Houska, B., Chachuat, B.: Branch-and-lift algorithm for deterministic global optimization in nonlinear optimal control. J. Optim. Theory Appl. **162**, 208–248 (2014)
28. Kinderlehrer, D., Stampacchia, G.: An Introduction to Variational Inequalities and their Applications. SIAM, Philadelphia (2000)
29. Leyffer, S., Manns, P.: Sequential linear integer programming for integer optimal control with total variation regularization. ESAIM Control Optim. Calc. Var. **28**, 66 (2022)
30. Leyffer, S., Manns, P., Winckler, M.: Convergence of sum-up rounding schemes for cloaking problems governed by the Helmholtz equation. Comput. Optim. Appl. **79**, 193–221 (2021)
31. Maggi, F.: Sets of Finite Perimeter and Geometric Variational Problems: An Introduction to Geometric Measure Theory. Number 135. Cambridge University Press, Cambridge (2012)
32. Manns, P., Kirches, C.: Multidimensional sum-up rounding for elliptic control systems. SIAM J. Numer. Anal. **58**(6), 3427–3447 (2020)
33. Manns, P., Nikolić, V.: Homotopy trust-region method for phase-field approximations in perimeter-regularized binary optimal control. ESAIM Control Optim. Calc. Var. (2024) (accepted)
34. Manns, P., Schiemann, A.: On integer optimal control with total variation regularization on multi-dimensional domains. SIAM J. Control. Optim. **61**(6), 3415–3441 (2023)
35. Manns, P., Severitt, M.: On discrete subproblems in integer optimal control with total variation regularization in two dimensions. INFORMS J. Comput. (2024)
36. Manns, P., Surowiec, T.M.: On binary optimal control in $H^s(0,T)$, $s < 1/2$. Comptes Rendus. Mathématique **361**(G9), 1531–1540 (2023)
37. Maranas, C.D., Floudas, C.A.: Global optimization in generalized geometric programming. Comput. Chem. Eng. **21**(4), 351–369 (1997)

38. McCormick, G.P.: Computability of global solutions to factorable nonconvex programs: Part I-convex underestimating problems. Math. Program. **10**(1), 147–175 (1976)
39. Papamichail, I., Adjiman, C.S.: Proof of convergence for a global optimization algorithm for problems with ordinary differential equations. J. Global Optim. **33**, 83–107 (2005)
40. Payne, L.E., Weinberger, H.F.: An optimal Poincaré inequality for convex domains. Arch. Ration. Mech. Anal. **5**(1), 286–292 (1960)
41. Pfetsch, M.E., Fügenschuh, A., Geißler, B., Geißler, N., Gollmer, R., Hiller, B., Humpola, J., Koch, T., Lehmann, T., Martin, A., et al.: Validation of nominations in gas network optimization: models, methods, and solutions. Optim. Methods Softw. **30**(1), 15–53 (2015)
42. Puranik, Y., Sahinidis, N.V.: Domain reduction techniques for global NLP and MINLP optimization. Constraints **22**(3), 338–376 (2017)
43. Quesada, I., Grossmann, I.E.: Global optimization algorithm for heat exchanger networks. Ind. Eng. Chem. Res. **32**(3), 487–499 (1993)
44. Quesada, I., Grossmann, I.E.: A global optimization algorithm for linear fractional and bilinear programs. J. Global Optim. **6**, 39–76 (1995)
45. Rao, A.V.: A survey of numerical methods for optimal control. Adv. Astronaut. Sci. **135**(1), 497–528 (2009)
46. Renardy, M., Rogers, R.C.: An Introduction to Partial Differential Equations, vol. 13. Springer, Berlin (2006)
47. Sahlodin, A.M., Chachuat, B.: Discretize-then-relax approach for convex/concave relaxations of the solutions of parametric ODEs. Appl. Numer. Math. **61**(7), 803–820 (2011)
48. Sahlodin, A.M., Chachuat, B.: Convex/concave relaxations of parametric ODEs using Taylor models. Comput. Chem. Eng. **35**(5), 844–857 (2011)
49. Schiemann, A., Manns, P.: Discretization of total variation in optimization with integrality constraints. SIAM J. Numer. Anal. (2024) (accepted)
50. Schmidt, E.: Über die ungleichung, welche die integrale über eine potenz einer funktion und über eine andere potenz ihrer ableitung verbindet. Math. Ann. **117**(1), 301–326 (1940)
51. Scott, J.K., Barton, P.I.: Convex relaxations for nonconvex optimal control problems. In: 2011 50th IEEE Conference on Decision and Control and European Control Conference, pp. 1042–1047. IEEE (2011)
52. Scott, J.K., Barton, P.I.: Improved relaxations for the parametric solutions of ODEs using differential inequalities. J. Global Optim. **57**(1), 143–176 (2013)
53. Scott, J.K., Chachuat, B., Barton, P.I.: Nonlinear convex and concave relaxations for the solutions of parametric ODEs. Optimal Control Appl. Methods **34**(2), 145–163 (2013)
54. Scott, J.K., Stuber, M.D., Barton, P.I.: Generalized McCormick relaxations. J. Global Optim. **51**(4), 569–606 (2011)
55. Severitt, M., Manns, P.: Efficient solution of discrete subproblems arising in integer optimal control with total variation regularization. INFORMS J. Comput. **35**(4), 869–885 (2023)
56. Singer, A.B., Barton, P.I.: Global optimization with nonlinear ordinary differential equations. J. Global Optim. **34**, 159–190 (2006)
57. Sundar, K., Nagarajan, H., Misra, S., Lu, M., Coffrin, C., Bent, R.: Optimization-based bound tightening using a strengthened qc-relaxation of the optimal power flow problem. In: 2023 62nd IEEE Conference on Decision and Control (CDC), pp. 4598–4605. IEEE (2023)
58. Tröltzsch, F.: Optimal Control of Partial Differential Equations: Theory, Methods, and Applications, vol. 112. American Mathematical Society, Providence (2010)
59. Veeser, A., Verfürth, R.: Poincaré constants for finite element stars. IMA J. Numer. Anal. **32**(1), 30–47 (2012)
60. 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., Jarrod Millman, K., 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.: SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat. Methods **17**, 261–272 (2020)
61. Wilhelm, M.E., Le, A.V., Stuber, M.D.: Global optimization of stiff dynamical systems. AIChE J. **65**(12), e16836 (2019)
62. Ye, J., Scott, J.K.: Modification and improved implementation of the RPD method for computing state relaxations for global dynamic optimization. J. Global Optim. 1–29 (2024)