Available online at www.sciencedirect.com

ScienceDirect

Comput. Methods Appl. Mech. Engrg. 283 (2015) 573–598

Computer methods in applied mechanics and engineering

www.elsevier.com/locate/cma

A closer look at consistent operator splitting and its extensions for topology optimization

Cameron Talischi, Glaucio H. Paulino*

Department of Civil & Environmental Engineering, University of Illinois at Urbana–Champaign, USA

Received 17 September 2013; received in revised form 23 June 2014; accepted 5 July 2014
Available online 19 July 2014

Abstract

In this work, we explore the use of operator splitting algorithms for solving regularized structural topology optimization problems. The context is a classical structural design problem (e.g., compliance minimization and compliant mechanism design), parametrized by means of density functions, whose ill-posedness is addressed by introducing a Tikhonov regularization term. The proposed forward–backward splitting algorithm treats the constituent terms of the cost functional separately, which allows for suitable approximations of the structural objective. We will show that one such approximation, inspired by the reciprocal expansions underlying the optimality criteria method, improves the convergence characteristics and leads to an update scheme resembling the heuristic sensitivity filtering method. We also discuss a two-metric variant of the splitting algorithm that removes the computational overhead associated with bound constraints on the density field without compromising convergence and quality of optimal solutions. We present several numerical results and investigate the influence of various algorithmic parameters.

Keywords: Topology optimization; Tikhonov regularization; Forward–backward splitting; Two-metric projection; Optimality criteria method

1. Introduction

The goal of topology optimization is to find the most efficient shape of a physical system whose behavior is captured by the solution to a boundary value problem that in turn depends on the given shape. As such, optimal shape problems can be viewed as a class of optimal control problems in which the control is the shape or domain of the governing state equation. These problems may be ill-posed, that is, they may not admit solutions in the classical sense, unless additional constraints are imposed on the regularity of the admissible shapes. For example, the basic compliance minimization problem in structural design, wherein one aims to find the stiffest arrangement of a fixed volume of material, favors non-convergent sequences of shapes that exhibit progressively finer features (see, for example, [1] and references therein). A manifestation of the ill-posedness of the continuum problem is that naive finite element approximations of the problem may suffer from numerical instabilities such as spurious checkerboard patterns or exhibit mesh-dependency of the solutions, both of which can be traced back to the absence of an internal length-scale in the continuum description of the problem [2]. An appropriate regularization scheme, based on one's choice

* Corresponding author.
E-mail addresses: ktalisch@illinois.edu (C. Talischi), paulino@uiuc.edu (G.H. Paulino).

http://dx.doi.org/10.1016/j.cma.2014.07.005

of parametrization of the unknown geometry, must therefore be employed to exclude this behavior and limit the complexity of the admissible shapes.

One such restriction approach, known as the *density filtering* method, implicitly enforces a prescribed degree of smoothness on all the admissible density fields that define the topology [3]. Filtering is essentially a means to define a space of admissible densities with an embedded level of regularity (cf. [4] for a more detailed discussion). This method and its variations are consistent in their use of sensitivity information in the optimization algorithm since the sensitivity of the objective and constraint functions are computed with respect to the associated auxiliary fields whose filtering defines the densities. By contrast, the *sensitivity filtering* method [2], which precedes the density filters and is typically described at the discrete level, performs the smoothening operation directly on the sensitivity field after a heuristic scaling step. The filtered sensitivities then enter the update scheme that evolves the design despite the fact that they do not correspond to the cost function of the optimization problem. While the sensitivity filtering has proven effective in practice for certain problems (for compliance minimization, it enjoys faster convergence than the density filter counterpart), a proper justification has remained elusive. As pointed out by Sigmund [5], it is generally believed that "the filtered sensitivities correspond to the sensitivities of a smoothed version of the original objective function" even though "it is probably impossible to figure out what objective function is actually being minimized". This view is confirmed in the present work, as we will show that an algorithm with calculations similar to what is done in the sensitivity filtering can be derived in a *consistent* manner from a proper regularization of the objective. We should also mention the recent work by Sigmund and Maute [6] in which the filtered sensitivities are shown to be consistent for materials obeying a nonlocal constitutive law.

The starting point is the authors' recent work [7] on an operator splitting algorithm for solving the compliance minimization problem where a Tikhonov regularization term is introduced to address the inherent ill-posedness of the problem. The derived update expression naturally contains a particular use of Helmholtz filtering, where in contrast to density and sensitivity filtering methods, the filtered quantity is the gradient descent step associated with the original structural objective. An observation made here is that if the gradient descent step in this algorithm is replaced by the optimality criteria (OC) update, then the interim density has a similar form to that of the sensitivity filter and in fact produces similar results (cf. Fig. 3). To make such a leap rigorous, we essentially embed the same reciprocal approximation of the structural cost function that is at the heart of the OC scheme in the forward–backward splitting algorithm. This leads to a generalization of the algorithm in [7] that is consistent, computationally efficient, and demonstrably convergent. The embedding of the reciprocal expansions is carried out in the splitting framework by means of their quadratic approximations, in the same spirit as Groenwold and co-workers [8,9], who extensively studied the performance of such approximations, finding them to be comparable to the original expansions.

Within the more general framework presented here, we will examine the choice of move limits and the step size parameter more closely and discuss strategies that can improve the convergence of the algorithm while maintaining the quality of final solutions. We also discuss a two-metric variant of the splitting algorithm that removes the computational overhead associated with the bound constraints on the density field without compromising convergence and quality of optimal solutions. In particular, we present and investigate a scheme based on the two-metric projection method of [10,11] that allows for the use of a more convenient metric for the projection step enforcing these bound constraints. This algorithm requires a simple and computationally inexpensive modification to the splitting scheme but features a min/max-type projection operation. We will see from the numerical examples that the two-metric variation retains the convergence characteristics of the forward–backward algorithm for various choices of algorithmic parameters. The details of the two types of algorithms are described for the finite-dimensional optimization problem obtained from the usual finite element approximation procedure, which we prove is convergent for the Tikhonov-regularized compliance minimization problem.

The remainder of this paper is organized as follows. In the next section, we describe the model topology optimization problem and its regularization. A general iterative scheme — one that encompasses the previous work [7] — for solving this problem based on forward–backward splitting is discussed in Section 3. Next, in Section 4, the connection is made with the sensitivity filtering method and the OC algorithm, and the appropriate choice of the approximate Hessian is identified. For the sake of concision and clarity, the discussion in these three sections is presented in the continuum setting. In Section 5, we begin by showing that the usual finite element approximations of the Tikhonov-regularized compliance minimization problem are convergent and derive the vector form of the discrete problem. The proposed algorithms along with some numerical investigation are presented in Sections 6 and 7. We conclude the work with some closing remarks and future research directions in Section 8.

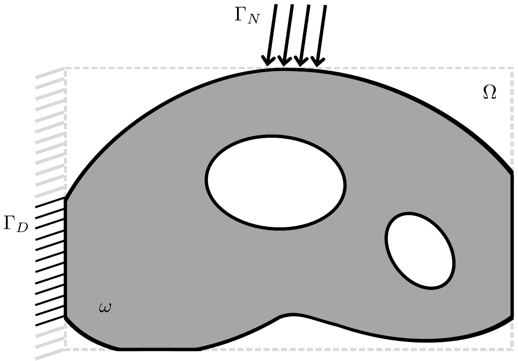

Fig. 1. Illustration of the prescribed boundary conditions defined on the design domain Ω. In a density formulation, each admissible shape $\omega \subseteq \Omega$ can be associated with some density function $\rho \in L^\infty(\Omega; [\delta_\rho, 1])$.

Before concluding the introduction, we briefly describe the notation adopted in this paper. As usual, $L^p(\Omega)$ and $H^k(\Omega)$ denote the standard Lebesgue and Sobolev spaces defined over a domain Ω with their vector-valued counterparts $L^p(\Omega; \mathbb{R}^d)$ and $H^k(\Omega; \mathbb{R}^d)$, and $L^p(\Omega; K) = \{f \in L^p(\Omega) : f(\mathbf{x}) \in K \text{ a.e.}\}$ for a given $K \subseteq \mathbb{R}$. Symbols $\wedge$ and $\vee$ denote the min and max operators, respectively, and when applied to functions are taken pointwise. Of particular interest are the inner product and norm associated with $L^2(\Omega)$, which are written as $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, respectively. Similarly, the inner product, norm and semi-norm associated with $H^k(\Omega)$ are denoted by $\langle \cdot, \cdot \rangle_k$, $\|\cdot\|_k$ and $|\cdot|_k$, respectively. Given a bounded and positive-definite linear operator $\mathcal{B}$, we write $\langle u, v \rangle_{\mathcal{B}} \equiv \langle u, \mathcal{B}v \rangle$ and the associated norm by $\|u\|_{\mathcal{B}} \equiv \langle u, u \rangle_{\mathcal{B}}^{1/2}$. Similarly, the standard Euclidean norm of a vector $\mathbf{v} \in \mathbb{R}^m$ is denoted by $\|\mathbf{v}\|$ and given a positive-definite matrix $\mathbf{B}$, we define $\|\mathbf{v}\|_{\mathbf{B}} = \left(\mathbf{v}^T \mathbf{B} \mathbf{v}\right)^{1/2}$. The ith components of vector $\mathbf{v}$ and the (i, j)th entry of matrix $\mathbf{B}$ are written as $[\mathbf{v}]_i$ and $[\mathbf{B}]_{ij}$, respectively.

2. Model problem and regularization

We begin with the description of the compliance minimization problem which is used as the model problem in this work. Let $\Omega \subseteq \mathbb{R}^d$, $d = 2, 3$ be the extended design domain with sufficiently smooth boundary. We consider boundary segments Γ_D and Γ_N that form a nontrivial partition of $\partial\Omega$, i.e., $\Gamma_D \cap \Gamma_N = \emptyset$, $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$ and Γ_D has non-zero surface measure (see Fig. 1). Each design over Ω is represented by a density function ρ, that is, a non-negative field bounded above by one, whose response is characterized by the solution $\mathbf{u}_\rho \in \mathcal{V}$ to the elasticity boundary value problem, given in the weak form by

$$a(\mathbf{u}, \mathbf{v}; \rho) = \ell(\mathbf{v}), \quad \forall \mathbf{v} \in \mathcal{V} \tag{1}$$

where $\mathcal{V} = \{\mathbf{u} \in H^1(\Omega; \mathbb{R}^d) : \mathbf{u}|_{\Gamma_D} = \mathbf{0}\}$ is the space of admissible displacements and

$$a(\mathbf{u}, \mathbf{v}; \rho) = \int_\Omega \rho^p \mathbf{C}\boldsymbol{\epsilon}(\mathbf{u}) : \boldsymbol{\epsilon}(\mathbf{v}) \mathrm{d}\mathbf{x}, \qquad \ell(\mathbf{v}) = \int_{\Gamma_N} \mathbf{t} \cdot \mathbf{v} \mathrm{d}s \tag{2}$$

are the usual energy bilinear and load linear forms. Moreover, $\boldsymbol{\epsilon}(\mathbf{u}) = (\nabla\mathbf{u} + \nabla\mathbf{u}^T)/2$ is the linearized strain tensor, $\mathbf{t} \in L^2(\Gamma_N; \mathbb{R}^d)$ is the prescribed tractions on Γ_N and $\mathbf{C}$ is the elasticity tensor for the constituent material. Observe that the classical Solid Isotropic Material with Penalization (SIMP) model is used to describe the dependence of the state equation on the density field, namely that the stiffness is related to the density through the power law relation ρ^p [12–14].[1] The bilinear form is continuous and also coercive provided that ρ is measurable and bounded below by some small positive constant $0 < \delta_\rho \ll 1$. In fact, there exist positive constants c and M such that for all $\rho \in L^\infty(\Omega; [\delta_\rho, 1])$,

$$|a(\mathbf{u}, \mathbf{v}; \rho)| \leq M \|\mathbf{u}\|_1 \|\mathbf{v}\|_1, \qquad a(\mathbf{u}, \mathbf{u}; \rho) \geq c \|\mathbf{u}\|_1^2, \quad \forall \mathbf{u}, \mathbf{v} \in \mathcal{V}. \tag{3}$$

[1] We use the classical SIMP parametrization with a positive lower bound on the densities. The reason is that later, we will consider Taylor expansions in $1/\rho$.

Together with continuity of the linear form ℓ (which follows from the assumed regularity of the applied tractions), these imply that (1) admits a unique solution $\mathbf{u}_\rho$ for all $\rho \in L^\infty(\Omega; [\delta_\rho, 1])$. Moreover, we have the uniform estimate $\|\mathbf{u}_\rho\|_1 \leq c^{-1}\|\ell\|$, where $\|\ell\|$ is the operator norm associated with ℓ. We also recall that by the principle of minimum potential, $\mathbf{u}_\rho$ is characterized by

$$\mathbf{u}_\rho = \underset{\mathbf{v}\in\mathcal{V}}{\operatorname{argmin}} \left[\frac{1}{2}a(\mathbf{v}, \mathbf{v}; \rho) - \ell(\mathbf{v})\right] \tag{4}$$

where the term in the bracket is the potential energy associated with deformation field $\mathbf{v}$. The following is a result that will be used later in the paper and readily follows from the stated assumptions (see, for example, [15]): Given a sequence $\{\rho_n\}$ and ρ in $L^\infty(\Omega; [\delta_\rho, 1])$ such that $\rho_n \to \rho$ strongly in $L^p(\Omega)$, $1 \leq p \leq \infty$, the associate displacement fields $\mathbf{u}_{\rho_n}$, up to a subsequence, converge in the strong topology of $H^1(\Omega; \mathbb{R}^d)$ to $\mathbf{u}_\rho$. This shows that if the cost functional depends continuously on $(\rho, \mathbf{u})$ in the strong topology of $L^p(\Omega) \times H^1(\Omega; \mathbb{R}^d)$, then compactness of the space of admissible densities in $L^p(\Omega)$ is a sufficient condition for existence of minimizers of the cost functional.

The cost functional for the compliance minimization problem is given by

$$J(\rho) = \ell(\mathbf{u}_\rho) + \lambda \int_\Omega \rho \mathrm{d}\mathbf{x}. \tag{5}$$

The first term in J is the compliance of the design while the second term represents a penalty on the volume of the material used. Minimizing this cost functional amounts to finding the stiffest arrangement while using the least amount of material with elasticity tensor $\mathbf{C}$. The parameter $\lambda > 0$ determines the trade-off between the stiffness provided by the material and the amount that is used (which presumably is proportional to the cost of the design). Since the SIMP model assigns smaller stiffness to the intermediate densities compared to the their contribution to the volume, it is expected that in the optimal regime, the density function are nearly binary (taking only values of δ_ρ and 1) provided that the penalty exponent p is sufficiently large. Later, in Section 7, we will discuss algorithms for the optimization problem featuring an explicit constraint on the volume of the structure.

As discussed in the introduction, the compliance minimization problem does not admit, in general, a solution in $L^\infty(\Omega; [\delta_\rho, 1])$, necessitating the introduction of additional restrictions on the regularity of density functions. This may be accomplished by the addition of a Tikhonov regularization term to the cost function [16,7,17] and considering:

$$\min_{\rho\in\mathcal{A}} \tilde{J}(\rho) = J(\rho) + \frac{\beta}{2}\langle\nabla\rho, \nabla\rho\rangle. \tag{6}$$

Here $\beta > 0$ is a positive constant determining the influence of this regularization. The use of larger values of β leads to smoother densities in the optimal regime and thus an appropriate value can be determined by means of trial-and-error in order to satisfy the given manufacturing considerations. The dimensional analysis used by Dede et al. [18], in the context of a phase field formulation, is also promising for a systematic selection of the regularization parameter. The minimization of $\tilde{J}$ is carried out over the set of admissible densities, defined as a subset of $H^1(\Omega)$, given by

$$\mathcal{A} = \left\{\rho \in H^1(\Omega) : \delta_\rho \leq \rho \leq 1 \text{ a.e.}\right\}. \tag{7}$$

The proof of existence of minimizers for (6) can be found in [7] (see also [19] for a weaker result) and essentially follows from compactness of the minimizing sequences of (6) in $L^p(\Omega)$, $1 \leq p < \infty$. We note that the norm of the density gradient also appears in phase field formulations of topology optimization (see, for example, [20–22,18]) as an interfacial energy term and is accompanied by a double-well potential penalizing intermediate densities. Taken together with appropriately chosen coefficients, the two terms can serve as approximation to the perimeter of the design.

Under additional assumptions of $\nabla\rho \cdot \mathbf{n} = 0$ on $\partial\Omega$ and $\rho \in H^2(\Omega)$, the Tikhonov regularization term can be written as

$$\frac{\beta}{2}\int_\Omega \nabla\rho \cdot \nabla\rho \mathrm{d}\mathbf{x} = \frac{\beta}{2}\left[-\int_\Omega \rho \Delta\rho \mathrm{d}\mathbf{x} + \int_{\partial\Omega} \rho\,(\nabla\rho \cdot \mathbf{n})\,\mathrm{d}s\right] = \frac{1}{2}\langle\rho, -\beta\Delta\rho\rangle. \tag{8}$$

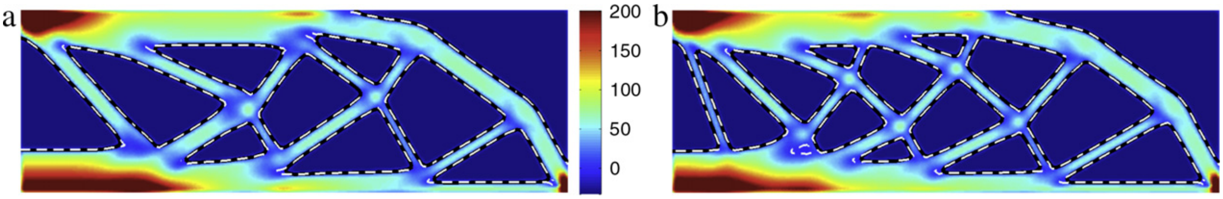

Fig. 2. Plot of $E(\rho)$–λ for two solutions to the MBB beam problem with $\beta = 0.01$. (a) corresponds to the solution shown in Fig. 7(b) and (b) corresponds to the solution shown in Fig. 7(c). The black line is the contour line for $\rho = 1/2$ and the dashed white line is the contour line where $E(\rho) = \lambda$. Note that only half the design domain is shown and the range of the colorbar is limited to $[-\lambda, 6\lambda]$ for better visualization.

Similarly, the more general regularization term $\frac{1}{2}\langle \nabla\rho, \kappa\nabla\rho\rangle$ in which $\kappa(\mathbf{x})$ is a bounded and positive-definite matrix, prescribing varying regularity of ρ in Ω and subsequently controlling feature size and orientation of the optimal design, can be written as $\frac{1}{2}\langle \rho, -\nabla\cdot(\kappa\nabla\rho)\rangle$. For brevity and emphasizing the quadratic form of this type of regularization, in the next two sections, we write the regularizer generically as

$$\frac{1}{2}\langle \rho, \mathcal{R}\rho\rangle \tag{9}$$

where $\mathcal{R}$ is a linear, self-adjoint and positive semi-definite operator on $\mathcal{A}$, though the additional assumption on densities are in fact not required (see also remarks in Section 3 of [7]).

Finally, we recall that the gradient of compliance, with respect to variations of density in the L^2-metric, is given by [19]

$$J'(\rho) = -E(\rho) + \lambda \tag{10}$$

where $E(\rho) = p\rho^{p-1}\mathbf{C}\boldsymbol{\epsilon}(\mathbf{u}_\rho) : \boldsymbol{\epsilon}(\mathbf{u}_\rho)$ is a strain energy density field. Note that $E(\rho)$ is non-negative for any admissible density and this is related to the monotonicity of the self-adjoint compliance problem: given densities ρ_1 and ρ_2 such that $\rho_1 \le \rho_2$ a.e., one can show $\ell(\mathbf{u}_{\rho_1}) \ge \ell(\mathbf{u}_{\rho_2})$. This property is the main reason why we restrict our attention in this paper to compliance minimization (though in Section 7, we will provide an example of compliant mechanism design which is not self-adjoint). Observe that $\hat{\rho}$ is a stationary point of J if

$$\begin{cases} E(\hat{\rho})(\mathbf{x}) < \lambda, & \text{if } \hat{\rho}(\mathbf{x}) = \delta_\rho \\ E(\hat{\rho})(\mathbf{x}) = \lambda, & \text{if } \delta_\rho < \hat{\rho}(\mathbf{x}) < 1 \\ E(\hat{\rho})(\mathbf{x}) > \lambda, & \text{if } \hat{\rho}(\mathbf{x}) = 1. \end{cases} \tag{11}$$

Thus, in regions where $E(\hat{\rho})$ exceeds the penalty parameter λ (regions that experience "large" deformation), density is at its maximum. Similarly, below this cutoff value the density is equal to the lower bound δ_ρ. Everywhere else, i.e., in the regions of intermediate density, the strain energy density is equal to the penalty parameter λ.

Fig. 2 shows the distribution of $E(\rho)$–λ for solutions to (6) obtained using the proposed algorithm (cf. Section 7 and Fig. 7(b) and (c)). Superimposed are the contour lines associated with $\rho = 1/2$ (plotted in black) representing the boundary of the optimal shape and $E(\rho) = \lambda$ (plotted in dashed white). The fact that these lines are nearly coincident shows that the solutions to the regularized problem, at least for sufficiently small regularization parameter β, are close to ideal in the sense that they nearly satisfy the stationarity condition for the structural objective J.

3. General splitting algorithm

In this section, we discuss a generalization of the forward–backward splitting algorithm that was explored in [7] for solving the regularized compliance minimization problem. The key idea behind this and other similar decomposition methods [23–25] is the separate treatment of constituent terms of the cost function.

A general algorithm for finding a minimizer of $\tilde{J}(\rho)$ consists of subproblems of the form:

$$\rho_{n+1} = \underset{\rho\in\mathcal{A}_n}{\operatorname{argmin}}\ J(\rho_n) + \left\langle \rho - \rho_n, J'(\rho_n)\right\rangle + \frac{1}{2\tau_n}\|\rho - \rho_n\|^2_{\mathcal{H}_n} + \frac{1}{2}\langle \rho, \mathcal{R}\rho\rangle \tag{12}$$

where $\mathcal{H}_n$ is a bounded and positive-definite linear operator. Compared to (6), we can see that while the regularization term has remained intact, J is replaced by a local quadratic model around ρ_n in which $\mathcal{H}_n$ may be viewed as an

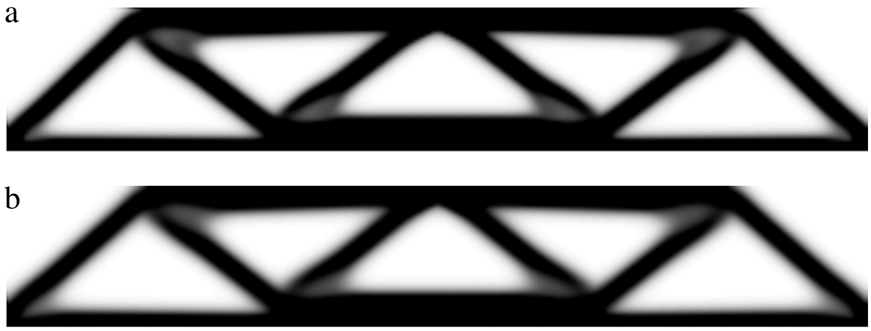

Fig. 3. (a) The solution to the MBB beam problem (see Section 6) using the sensitivity filtering method (consisting of (27) and (21)) (b) The solution using the update steps (28) and (21). In both cases, $\mathcal{F}$ is taken to be the "Helmholtz" filter and the move limit was set to $m_n = 0.25$.

approximation to the Hessian of J evaluated at ρ_n. The suitable choice of $\mathcal{H}_n$ is an important issue explored in this work. Note that constant terms such as $J(\rho_n)$ and $\langle \rho_n, J'(\rho_n) \rangle$ do not affect the optimization but are provided to emphasize the expansion of J. Moreover, $\tau_n > 0$ is a *step size* parameter that determines the curvature of this approximation. For sufficiently small τ_n (large curvature), the approximation is conservative in that it majorizes (lies above) J. This is crucial in guaranteeing descent in each iteration and overall convergence of the algorithm (see Section 6).

We have included another limiting measure in (12), a minor departure from the above-mentioned references, by replacing the constraint set $\mathcal{A}$ by a subset $\mathcal{A}_n$ in order to limit the point-wise change in the density to a specified *move limit* m_n. More specifically, we have defined

$$\mathcal{A}_n = \{\rho \in \mathcal{A} : |\rho - \rho_n| \le m_n \text{ a.e.}\} = \left\{\rho \in H^1(\Omega) : \rho_n^{\mathsf{L}} \le \rho \le \rho_n^{\mathsf{U}} \text{ a.e.}\right\} \tag{13}$$

where in the latter expression

$$\rho_n^{\mathsf{L}} = \delta_\rho \wedge (\rho_n - m_n), \qquad \rho_n^{\mathsf{U}} = 1 \vee (\rho_n + m_n). \tag{14}$$

The use of move limits, akin to a trust region strategy, is common in topology optimization literature as a means to stabilize the topology optimization algorithm, especially in the early iterations to prevent members from forming prematurely. As we will show with an example, this is only important when a smaller regularization parameter is used and the final topology is complex. Near the optimal solution, the move limit strategy is typically inoperative. Of course, by setting $m_n \equiv 1$, we can get $\mathcal{A} = \mathcal{A}_n$ and recover the usual form of (12).

We can show that (12) is equivalent to

$$\rho_{n+1} = \underset{\rho \in \mathcal{A}_n}{\operatorname{argmin}} \; \left\| \rho - \rho_{n+1}^* \right\|_{(\mathcal{H}_n + \tau_n \mathcal{R})}^2 \tag{15}$$

where the *interim density* ρ_{n+1}^* is given by

$$\rho_{n+1}^* = (\mathcal{H}_n + \tau_n \mathcal{R})^{-1} \left[\mathcal{H}_n \rho_n - \tau_n J'(\rho_n)\right]. \tag{16}$$

This can be seen from the following identity, obtained by a direct expansion,

$$\begin{aligned} \frac{1}{2\tau_n} \left\| \rho - \rho_{n+1}^* \right\|_{(\mathcal{H}_n + \tau_n \mathcal{R})}^2 &= J(\rho_n) + \langle \rho - \rho_n, J'(\rho_n) \rangle + \frac{1}{2\tau_n} \|\rho - \rho_n\|_{\mathcal{H}_n}^2 + \frac{1}{2} \langle \rho, \mathcal{R}\rho \rangle \\ &\quad - J(\rho_n) + \langle \rho_n, J'(\rho_n) \rangle - \frac{1}{2\tau_n} \|\rho_n\|_{\mathcal{H}_n}^2 + \frac{1}{2\tau_n} \left\| \rho_{n+1}^* \right\|_{(\mathcal{H}_n + \tau_n \mathcal{R})}^2 \end{aligned} \tag{17}$$

and noting that adding a constant term to the objective function or multiplying it by a scalar does not affect its minimizer.

Alternatively, the interim density can be written as an update where the gradient of $\tilde{J}$ is scaled by the inverse of its approximate Hessian, namely

$$\begin{aligned}\rho^*_{n+1} &= \rho_n - (\mathcal{H}_n + \tau_n\mathcal{R})^{-1}(\mathcal{H}_n\rho_n + \tau_n\mathcal{R}\rho_n) + (\mathcal{H}_n + \tau_n\mathcal{R})^{-1}\left[\mathcal{H}_n\rho_n - \tau_n J'(\rho_n)\right] \\ &= \rho_n + (\mathcal{H}_n + \tau_n\mathcal{R})^{-1}\left[-\tau_n\mathcal{R}\rho_n - \tau_n J'(\rho_n)\right] \\ &= \rho_n - \tau_n(\mathcal{H}_n + \tau_n\mathcal{R})^{-1}\left[J'(\rho_n) + \mathcal{R}\rho_n\right]. \end{aligned} \tag{18}$$

Returning to (15), we can see that next density ρ_{n+1} is the *projection* of the interim density, with respect to the norm defined by $\mathcal{H}_n + \tau_n\mathcal{R}$, onto the constraint space $\mathcal{A}_n$. From the assumptions on $\mathcal{H}_n$ and $\mathcal{R}$ and the fact that $\mathcal{A}_n$ is a closed convex subset of $H^1(\Omega)$, it follows that the projection is well-defined and there is a unique update ρ_{n+1}.

By setting $\mathcal{R} = -\beta\Delta$, which corresponds to the regularization term of (6) and choosing $\mathcal{H}_n$ to be the identity map $\mathcal{I}$, we recover the forward–backward algorithm investigated in [7]. In this case, the interim update satisfies the Helmholtz equation

$$(\mathcal{I} - \tau_n\beta\Delta)\,\rho^*_{n+1} = \rho_n - \tau_n J'(\rho_n) \tag{19}$$

with homogeneous Neumann boundary conditions. Note that the right-hand-side is the usual gradient descent step (with step size τ_n) associated with J (the forward step) and the interim density is obtained by applying the inverse of the Helmholtz operator (the backward step), which can be viewed as the filtering of right-hand-side with Green's function of the Helmholtz equation.[2] As mentioned in the introduction, this appearance of filtering is fundamentally different from density and sensitivity filtering methods. Moreover, the projection operation in this case is with respect to a scaled Sobolev metric, namely (see expressions (3.18) and (3.19) in [7])

$$\rho_{n+1} = \underset{\rho\in\mathcal{A}_n}{\operatorname{argmin}}\ \left\|\rho - \rho^*_{n+1}\right\|^2 + \beta\tau_n\left|\rho - \rho^*_{n+1}\right|_1^2 \tag{20}$$

which numerically requires the solution to a box-constrained convex quadratic program. In [7], we also explored an "inconsistent" variation of this algorithm where we neglected the second term in (20) and essentially used the L^2-metric for the projection step. Due to the particular geometry of the box constraints in $\mathcal{A}_n$, the L^2-projection has the explicit solution given by

$$\rho_{n+1} = \left(\rho^*_{n+1} \wedge \rho^{\mathsf{L}}_n\right) \vee \rho^{\mathsf{U}}_n. \tag{21}$$

The appeal of this min/max type operation is that it is trivial from the computational point of view. Moreover, it coincides with the last step in the OC update scheme [19]. However, this is an inconsistent step for the Tikhonov regularized problem since ρ_{n+1} need not lie in $H^1(\Omega)$. In fact, strictly speaking, (21) is valid only if $\mathcal{A}_n$ is enlarged from functions in $H^1(\Omega)$ to all functions in $L^2(\Omega)$ bounded below by ρ^{L}_n and above by ρ^{U}_n. In spite of this inconsistency, the algorithm composed of (19) and (21) was convergent and numerically shown to produce noteworthy solutions with minimal intermediate densities. This merits a separate investigation since as suggested in [7], this algorithm may in fact solve a smoothed version of the perimeter constraint problem where the regularization term is the total variation of the density field. We will return to the use of L^2-projection later in Section 6 but *this time in a consistent manner with the aid of the two-metric projection approach of* [10,11].

4. Optimality criteria and sensitivity filtering

Whenever applicable, the so-called optimality criteria (OC) method is preferred to other gradient descent algorithms of comparable simplicity in the structural optimization community. For example, we refer to [26] for a relationship and comparison between the OC method and the gradient projection algorithm. Our interest here in the OC method is that the density and sensitivity filtering methods are typically implemented in the OC framework. Moreover, as we shall see, this examination will lead to the choice of $\mathcal{H}_n$ in the algorithm (12).

[2] The designations "forward" and "backward" step come from the fact that (19) can be written as $\rho^*_{n+1} = (\mathcal{I} + \tau_n\mathcal{R})^{-1}(\mathcal{I} - \tau_n J')\,\rho_n$. Similarly, (16) has the equivalent expression $\rho^*_{n+1} = \left(\mathcal{I} + \tau_n\mathcal{H}_n^{-1}\mathcal{R}\right)^{-1}\left(\mathcal{I} - \tau_n\mathcal{H}_n^{-1}J'\right)\rho_n$.

The interim density in the OC method for the compliance minimization problem, in the absence of regularization, is obtained from the fixed point iteration

$$\rho_{n+1}^{*} = \rho_n \left[\frac{E(\rho_n)}{\lambda} \right]^{1/2} \equiv \rho_n \left[e_\lambda(\rho_n) \right]^{1/2}. \tag{22}$$

Note that the strain energy density $E(\rho_n)$ and subsequently its normalization $e_\lambda(\rho_n)$ are non-negative for any admissible density ρ_n and therefore ρ_{n+1}^{*} is well-defined. Recalling the necessary condition of optimality for an optimal density $\hat{\rho}$ stated in (11), it is evident that such $\hat{\rho}$ is a fixed point of the OC iteration. Intuitively, the current density ρ_n is increased (decreased) in regions where $E(\rho_n)$ is greater (less) than the penalty parameter λ by a factor of $[e_\lambda(\rho_n)]^{1/2}$. The next density ρ_{n+1} in the OC is given by the projection ρ_{n+1}^{*} in (21).

It is more useful here to adopt an alternative view of the OC scheme, namely that the OC update can be seen as the solution to an approximate subproblem where compliance is replaced by a Taylor expansion in the intermediate field ρ^{-1} [27]. The intuition behind such an expansion is that, locally, compliance is inversely proportional to ρ. In particular, ρ_{n+1}^{*} can be shown to be the stationary point of the "reciprocal approximation" around ρ_n defined by

$$J_{\mathsf{rec}}(\rho;\rho_n) \equiv \ell(\mathbf{u}_{\rho_n}) + \left\langle \frac{\rho_n}{\rho}(\rho - \rho_n), -E(\rho_n) \right\rangle + \lambda \int_\Omega \rho \mathrm{d}\mathbf{x}. \tag{23}$$

Note that the expansion in the inverse of density is carried out only for the compliance term, and the volume term, which is already linear, is not altered. The expression for $J_{\mathsf{rec}}(\rho;\rho_n)$ can be alternatively written as

$$J_{\mathsf{rec}}(\rho;\rho_n) = J(\rho_n) + \left\langle \rho - \rho_n, J'(\rho_n) \right\rangle + \frac{1}{2} \left\langle \rho - \rho_n, \frac{2E(\rho_n)}{\rho}(\rho - \rho_n) \right\rangle \tag{24}$$

which highlights the fact that the (nonlinear) curvature term in (24) makes it a more accurate approximation of compliance compared to the linear expansion. With regard to the OC update, one can show that the interim update satisfies $J'_{\mathsf{rec}}(\rho_{n+1}^{*};\rho_n) = 0$, and its L^2-projection is indeed the minimizer of $J_{\mathsf{rec}}(\rho;\rho_n)$ over $\left\{ \rho \in L^2(\Omega) : \rho_n^{\mathsf{L}} \le \rho \le \rho_n^{\mathsf{U}} \text{ a.e.} \right\}$.

We now turn to the *sensitivity filtering* method, which is described with the OC algorithm. Let $\mathcal{F}$ denote a linear filtering map, for example, the Helmholtz filter $\mathcal{F} = \left(\mathcal{I} - r^2\Delta\right)^{-1}$ discussed before or the convolution filter of radius r [28,15]

$$\mathcal{F}(\psi)(\mathbf{x}) \equiv \int_\Omega F_r(\mathbf{x} - \mathbf{y}) \psi(\mathbf{y}) \mathrm{d}\mathbf{y} \tag{25}$$

where the kernel is the linear hat function $F_r(\mathbf{x}) = \max(1 - |\mathbf{x}|/r, 0)$. The main idea in the sensitivity filtering method is that $e_\lambda(\rho_n)$ is heuristically replaced by the following smoothed version

$$\tilde{e}_\lambda(\rho_n) \equiv \frac{1}{\rho_n} \mathcal{F}[\rho_n e_\lambda(\rho_n)] \tag{26}$$

before entering the OC update. Notice that the filtering map is applied to the scaling of $e_\lambda(\rho_n)$ by the density field itself, which is not easy to justify. The interim density update is thus given by

$$\rho_{n+1}^{*} = \rho_n \left[\tilde{e}_\lambda(\rho_n) \right]^{1/2} = \rho_n \left\{ \frac{\mathcal{F}[\rho_n e_\lambda(\rho_n)]}{\rho_n} \right\}^{1/2} = \rho_n^{1/2} \mathcal{F}[\rho_n e_\lambda(\rho_n)]^{1/2}. \tag{27}$$

A key observation in this work is that if *we replace the gradient decent step in forward–backward algorithm (cf.* (19)*) with the OC step, we obtain a similar update scheme to that of the sensitivity filtering method.* More specifically, note that (19) can be written as $\rho_{n+1}^{*} = \mathcal{F}\left[\rho_n - \tau_n J'(\rho_n)\right]$. Substituting $\rho_n - \tau_n J'(\rho_n)$ with $\rho_n [e_\lambda(\rho_n)]^{1/2}$ gives

$$\rho_{n+1}^{*} = \mathcal{F}\left\{ \rho_n [e_\lambda(\rho_n)]^{1/2} \right\} \tag{28}$$

which resembles (27). In fact, as illustrated in Fig. 3, the two expressions produce very similar final results (in particular, observe the similarity between the patches of intermediate density in the corners that is characteristic of the sensitivity filtering method). Of course, the leap from the forward–backward algorithm to (28), just like the sensitivity filtering method, lacks mathematical justification. However, we will expand upon this observation and next derive an algorithm similar to this empirical modification of the forward–backward algorithm in a consistent manner.

Embedding the reciprocal approximation

Recalling the role of the reciprocal approximation of compliance in the OC method, the key idea is to embed such an approximation in the general subproblem of (12). We do so by choosing $\mathcal{H}_n$ to be the Hessian of $J_{\text{rec}}(\rho;\rho_n)$ evaluated at ρ_n, namely

$$\mathcal{H}_n = J''_{\text{rec}}(\rho_n;\rho_n) = \frac{2E(\rho_n)}{\rho_n}\mathcal{I}. \tag{29}$$

As noted in the introduction, the use of a quadratic approximation of the reciprocal and exponential expansions has been studied extensively in [8,9]. Observe that here $E(\rho)$ is a non-negative function for any admissible ρ but may vanish in some subset of Ω. This means that $\mathcal{H}_n$ is only positive semi-definite and does not satisfy the definiteness requirement for use in (12). We can remedy this by replacing $E(\rho_n)$ in (29) with $E(\rho_n) \wedge \delta_E$ where $0 < \delta_E \ll \lambda$ is a prescribed constant. However, in most compliance problems (e.g., the benchmark problem considered later in Section 7) the strain energy field is strictly positive for all admissible densities. In fact, the regions with zero strain energy density do not experience any deformation and in light of the conditions of optimality (11) should be assigned the minimum density. Therefore, to simplify the matters, *we assume in the remainder of this section that the loading and support conditions defined on Ω are such that $E(\rho) \geq \delta_E$ almost everywhere* for all $\rho \in L^\infty(\Omega; [\delta_\rho, 1])$.

Comparing the quadratic approximation of J with this choice of $\mathcal{H}_n$ and the reciprocal approximation itself (cf. (24)), we see that the difference is in their curvature terms (the linear terms of course match). The curvature of the quadratic model depends on and can be controlled by τ_n while the nonlinear curvature in J_{rec} is a function of ρ.

Substituting (29) into (16), the expression for the interim density becomes

$$\left[\frac{2E(\rho_n)}{\rho_n}\mathcal{I} + \tau_n\mathcal{R}\right]\rho^*_{n+1} = 2E(\rho_n) + \tau_n\left[E(\rho_n)–\lambda\right] = (2+\tau_n)\,E(\rho_n) - \tau_n\lambda. \tag{30}$$

Multiplying by $\rho_n/\left[2E(\rho_n)\right]$ and simplifying yields

$$\left[\mathcal{I} + \frac{\rho_n}{2E(\rho_n)}\tau_n\mathcal{R}\right]\rho^*_{n+1} = \rho_n\left[\left(1+\frac{\tau_n}{2}\right) - \frac{\tau_n}{2e_\lambda(\rho_n)}\right]. \tag{31}$$

To better understand the characteristics of this update, let us specialize this to the case of Tikhonov regularization and set $\tau_n = 1$ (so that the quadratic model and the reciprocal approximation have the same curvature at ρ_n). This gives

$$\left[\mathcal{I} - \frac{\rho_n}{2E(\rho_n)}\beta\Delta\right]\rho^*_{n+1} = \rho_n\left[\frac{3}{2} - \frac{1}{2e_\lambda(\rho_n)}\right]. \tag{32}$$

First note that in the absence of regularization (i.e., $\beta = 0$), the update relation has the same fixed-point iteration form as the OC update with the ratio $e_\lambda(\rho_n)$ determining the scaling of ρ_n. The scaling field here is $3/2 - 1/\left[2e_\lambda(\rho_n)\right]$ whereas in the OC method it is given by $\sqrt{e_\lambda(\rho_n)}$. As shown in Fig. 4, the scaling fields and their derivatives coincide in the regions where $e_\lambda(\rho_n) = 1$, which means that locally the two are similar. The reduction in density is more aggressive with this scaling when $e_\lambda(\rho_n) < 1$ whereas the OC update leads to larger increase for $e_\lambda(\rho_n) > 1$.

As with the forward–backward algorithm (cf. Eq. (19)), the presence of regularization again leads to the appearance of Helmholtz filtering (the inverse of the left-hand-side operator) but with two notable differences. First, the right-hand-side term now is an OC-like scaling of density instead of the gradient descent step (the same is true in (31) for an arbitrary step size τ_n). Furthermore, the filtering is not uniform across the domain and its degree of smoothening is scaled by $\rho_n/\left[2E(\rho_n)\right]$. The important result here is that, by embedding the reciprocal approximation of compliance in our quadratic model, we are able to obtain a relation for the ρ^*_{n+1} that features an OC-like right-hand-side and its filtering, very much similar in form to the heuristic update scheme of (28) that was compared to the sensitivity filtering.

We also remark that the right-hand-side of (31) for a general step size τ_n is related to the reciprocal expansions in interim variable $1/(\rho - L_n)$, used in the Method of Moving Asymptotes (MMA) [29], if the asymptote is set to

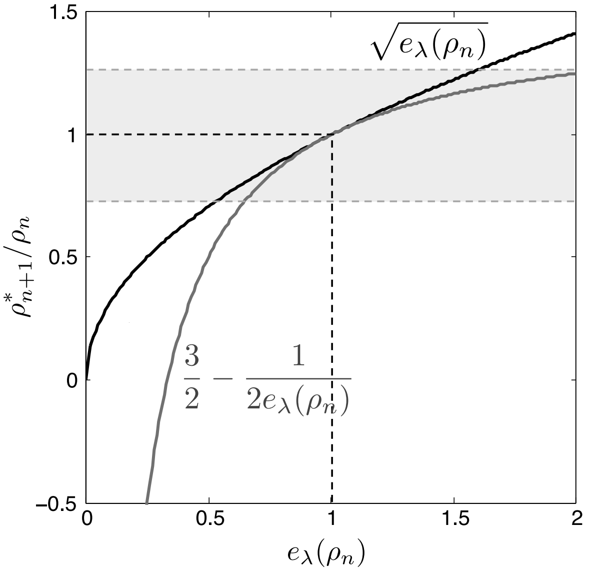

Fig. 4. Comparison between scaling terms appearing in the OC update (black line) and right hand side of (32) (gray line). The OC is more aggressive in regions $e_\lambda(\rho_n) > 1$ and less aggressive when $e_\lambda(\rho_n) < 1$.

$L_n = \rho_n(1 - \tau_n)$. The interim update for MMA, before accounting for the box constraints and in the absence of regularization, is given by

$$\rho^*_{n+1} = \rho_n\left[(1-\tau_n) + \tau_n\sqrt{e_\lambda(\rho_n)}\right]. \tag{33}$$

Similar to the case of $\tau_n = 1$, this scaling of ρ_n is similar to the right-hand-side of (31) in regions where $e_\lambda(\rho_n) \approx 1$.

Another key difference between the forward–backward algorithm and the OC-based filtering methods is that the projection of ρ^*_{n+1} defining the next iterate ρ_{n+1} in the forward–backward scheme is with respect to the metric induced by $\mathcal{H}_n + \tau_n\mathcal{R}$ in contrast to the L^2-projection given by (21). As discussed before, the L^2-projection is well-suited for the geometry of the constraint set $\mathcal{A}_n$ due to decomposition of box constraints. It may be tempting to inconsistently use the interim density (31) with the L^2-projection but this is not necessarily guaranteed to decrease the cost function. Arbitrary projections of unconstrained Newton steps are not mathematically warranted. Numerically one would observe that such an inconsistent algorithm excessively removes material and leads to final solutions with low volume fraction.

In Section 6, we explore a variant of the splitting algorithm that is related to the two-metric projection method of [10,11], and allows for the use of a more convenient metric for the projection step. This can be done provided that the operator whose associated norm defines the gradient[3] is modified appropriately in the regions where the constraints are active. More specifically, in the interim update step (cf. (18)), $\mathcal{H}_n + \tau_n\mathcal{R}$ is modified to produce a linear operator $\mathcal{D}_n$ with a particular structure that eliminates the coupling between regions of active and free constraints. The projection of the interim density given by

$$\rho^*_{n+1} = \rho_n - \tau_n\mathcal{D}_n^{-1}\left[J'(\rho_n) + \mathcal{R}\rho_n\right] \tag{34}$$

with respect to the L^2-norm is then guaranteed to decrease the cost function. Note that when there are no active constraints (e.g., in the beginning of the algorithm the density field takes mostly intermediate values), $\mathcal{D}_n = \mathcal{H}_n + \tau_n\mathcal{R}$ and (31) holds for the interim update and its L^2-projection produces the next iterate. In general, (31) holds locally for the regions where the box constraints are not active (i.e., regions of intermediate density) and so the analogy to the sensitivity filtering method holds in such regions.

To avoid some technical nuisances (that the L^2-projection on a closed convex subset of $H^1(\Omega)$ is not well-defined) and the cumbersome notation required to precisely define $\mathcal{D}_n$ in the continuum setting (that may obscure the simple

[3] Recall that $\mathcal{B}^{-1}f'(\rho)$ is the gradient of functional f with respect to the metric induced by $\mathcal{B}$.

procedure for its construction), we defer the details to Section 6 where we describe the algorithm for the finite-dimensional optimization problem obtained from the usual finite element approximation procedure. The intuition developed in the preceding discussion carries over to the discrete setting.

5. Finite element approximation

We begin with describing the approximate "finite element" optimization problem, based on a typical choice of discretization spaces, and establish the convergence of the corresponding optimal solutions to a solution of the continuum problem (6) in the limit of mesh refinement. Our result proves strong convergence of a subsequence of solutions, and thus rules out the possibility of numerical instabilities such as checkerboard patterns observed in density formulations. We remark that similar results are available for the density-based restriction formulations (see for example [30,31,28]) and the proof is along the same lines. Such convergence results are essential in justifying an approach where one first discretizes a well-posed continuum problem and then chooses an algorithm to solve the resulting finite dimensional problem (this is the procedure adopted in this work). Then, with the FE convergence result in hand, the only remaining task is to analyze the convergence of the proposed optimization algorithm, which is discussed in Section 6.

5.1. Convergence under mesh refinement

Consider a partitioning of Ω into pairwise disjoint finite elements $\mathcal{T}_h = \{\Omega_e\}_{e=1}^{l}$ with characteristic mesh size h. Let $\mathcal{A}_h$ be the FE discretization of $\mathcal{A}$ based on this partition:

$$\mathcal{A}_h = \left\{\rho \in C^0(\overline{\Omega}) : \rho|_{\Omega_e} \in \mathcal{P}(\Omega_e), \forall e = 1, \dots, l\right\} \cap \mathcal{A} \tag{35}$$

where $\mathcal{P}(\Omega_e)$ is an appropriate space of functions defined on Ω_e containing at least first-order polynomials (e.g., bilinear functions if Ω_e is a rectangular element). Similarly, we define:

$$\mathcal{V}_h = \left\{\mathbf{u} \in C^0(\overline{\Omega}; \mathbb{R}^d) : [\mathbf{u}]_i\,|_{\Omega_e} \in \mathcal{P}(\Omega_e), \forall e = 1, \dots, l, \forall i = 1, \dots, d\right\} \cap \mathcal{V}. \tag{36}$$

We also assume that the mesh $\mathcal{T}_h$ is chosen in such a way that the transition from Γ_D to Γ_N is properly aligned with the mesh. In practice, both density and displacement fields are discretized with linear elements (e.g., linear triangles, bilinear quads or linearly-complete convex polygons in two spatial dimensions). To avoid any ambiguity regarding the definition of the FE partitions, we assume a regular refinement such that the resulting finite element spaces are ordered, e.g., $\mathcal{A}_h \supseteq \mathcal{A}_{h'}$ whenever $h \le h'$. We consider the limit $h \to 0$ to establish convergence of solutions under mesh refinement. In what follows, C denotes a generic positive constant independent of h.

What is needed in the proof of convergence is the existence of an interpolation operator $\mathcal{I}_h : \mathcal{V} \to \mathcal{V}_h$ such that for all $\mathbf{u} \in \mathcal{V} \cap H^2(\Omega; \mathbb{R}^d)$

$$\|\mathbf{u} - \mathcal{I}_h\mathbf{u}\|_1 \le Ch\,|\mathbf{u}|_2\,. \tag{37}$$

Similarly, we need the mapping $i_h : \mathcal{A} \to \mathcal{A}_h$ for the design space such that $i_h\rho \to \rho$ as $h \to 0$. The construction of such interpolants is standard in finite element approximation theory and we refer the reader to [32].

The approximate finite element problem, specialized to Tikhonov regularization, is defined by

$$\min_{\rho\in\mathcal{A}_h} \tilde{J}_h(\rho) = J_h(\rho) + \frac{\beta}{2}\,|\rho|_1^2\,. \tag{38}$$

In this expression, the approximation to $J(\rho)$ is defined by

$$J_h(\rho) = \ell(\mathbf{u}_{\rho,h}) + \lambda\int_\Omega \rho\mathrm{d}\mathbf{x}$$

where $\mathbf{u}_{\rho,h} \in \mathcal{V}_h$ is the solution to the Galerkin approximation of (1) satisfying

$$a(\mathbf{u}_{\rho,h}, \mathbf{v}_h; \pi_h\rho) = \ell(\mathbf{v}_h), \qquad \forall\mathbf{v}_h \in \mathcal{V}_h. \tag{39}$$

Here π_h is a projection onto the space of piecewise constant fields on $\mathcal{T}_h$, an example of which is given in the next section. While it is possible to define the finite element approximation of the state equation without this projection, we have included this in the present convergence analysis since the use of piecewise constant densities is common practice in topology optimization and can simplify the finite element implementation (cf. [4]). In particular, this shows that even though the choice of continuous densities is natural for $\mathcal{A}_h$, the discretization of the boundary value problem can still be based on element-wise constant densities. Since π_h fixes piecewise constant fields, we have the following estimate for the error in the projection of $\rho \in H^1(\Omega)$ (cf. [32])

$$\|\rho - \pi_h \rho\|_0 \leq Ch\,|\rho|_1\,. \tag{40}$$

By the principle of minimum potential, we can write

$$\ell(\mathbf{u}_{\rho,h}) = -2 \min_{\mathbf{v}_h \in \mathcal{V}_h} \left[\frac{1}{2} a(\mathbf{v}_h, \mathbf{v}_h; \pi_h \rho) - \ell(\mathbf{v}_h)\right] = \max_{\mathbf{v}_h \in \mathcal{V}_h} \left[2\ell(\mathbf{v}_h) - a(\mathbf{v}_h, \mathbf{v}_h; \pi_h \rho)\right]. \tag{41}$$

From the above relation, it is easy to see that $\mathcal{V}_h \subseteq \mathcal{V}$ implies $\ell(\mathbf{u}_{\rho,h}) \leq \ell(\mathbf{u}_{\pi_h \rho})$ for any given ρ. Since compliance is uniformly bounded over $L^\infty(\Omega; [\delta_\rho, 1])$, it follows that J_h is also uniformly bounded over $L^\infty(\Omega; [\delta_\rho, 1])$ and all h.

Consider a sequence of FE partitions $\mathcal{T}_h$ with $h \to 0$ and let ρ_h be the optimal solution to the associated finite element approximation (38), i.e., the minimizer of $\tilde{J}_h$ in $\mathcal{A}_h$. We first show that the sequence ρ_h is bounded in $H^1(\Omega)$. To see this, fix h_0 in this sequence. If $\hat{\rho}_h$ is the minimizer of $\tilde{J}$ in $\mathcal{A}_h$ (there is no approximation of the displacement field involved here), then for $h \leq h_0$,

$$\tilde{J}(\hat{\rho}_h) \leq \tilde{J}(\rho_{h_0}) \tag{42}$$

since $\rho_{h_0} \in \mathcal{A}_{h_0} \subseteq \mathcal{A}_h$. Now, from the definition of ρ_h, we have

$$\tilde{J}_h(\rho_h) \leq \tilde{J}_h(\hat{\rho}_h) = \tilde{J}(\hat{\rho}_h) + \left[J_h(\hat{\rho}_h) - J(\hat{\rho}_h)\right] \leq \tilde{J}(\rho_{h_0}) + \left[J_h(\hat{\rho}_h) - J(\hat{\rho}_h)\right] \tag{43}$$

which, in turn, implies

$$\frac{\beta}{2}\,|\rho_h|_1^2 \leq \tilde{J}(\rho_{h_0}) + \left[J_h(\hat{\rho}_h) - J(\hat{\rho}_h) - J(\rho_h)\right]. \tag{44}$$

Since the term in the bracket is bounded in h, it follows that $\limsup_h |\rho_h|_1^2 < \infty$. Thus, the sequence ρ_h is bounded in $H^1(\Omega)$ and by Rellich's theorem [33], we have convergence of a subsequence, which we shall again denoted by $\{\rho_h\}$, strongly in $L^2(\Omega)$ and weakly in $H^1(\Omega)$ to some $\rho^* \in H^1(\Omega)$. To see that ρ^* satisfies the bound constraints and thus belongs to $\mathcal{A}$, we can consider another subsequence for which the convergence is pointwise. We can show that $\pi_h \rho_h$ also converges to ρ^* strongly in $L^2(\Omega)$ since by the triangle inequality and (40), we have

$$\left\|\pi_h \rho_h - \rho^*\right\|_0 = \left\|\rho_h - \rho^*\right\|_0 + \|\pi_h \rho_h - \rho_h\|_0 \leq \left\|\rho_h - \rho^*\right\|_0 + Ch\,|\rho_h|_1\,. \tag{45}$$

Recall that $|\rho_h|_1$ is bounded. Subsequently, by the remark made in Section 2, we can also conclude that $\mathbf{u}_{\pi_h \rho_h} \to \mathbf{u}_{\rho^*}$ in $H^1(\Omega; \mathbb{R}^d)$ as $h \to 0$.

We next show that ρ^* is a solution to the continuum problem, thereby establishing the convergence of the FE approximate problems. First, note that by lower semi-continuity of the norm under weak convergence,

$$\left|\rho^*\right|_1^2 \leq \liminf_h |\rho_h|_1^2\,. \tag{46}$$

Furthermore, we can establish the convergence of $\mathbf{u}_{\rho_h,h}$ to $\mathbf{u}_{\rho^*}$ in $H^1(\Omega; \mathbb{R}^d)$ by noting that

$$\begin{aligned}
\left\|\mathbf{u}_{\rho_h,h} - \mathbf{u}_{\rho^*}\right\|_1 &\leq \left\|\mathbf{u}_{\pi_h \rho_h} - \mathbf{u}_{\rho^*}\right\|_1 + \left\|\mathbf{u}_{\pi_h \rho_h} - \mathbf{u}_{\rho_h,h}\right\|_1 \\
&\leq \left\|\mathbf{u}_{\pi_h \rho_h} - \mathbf{u}_{\rho^*}\right\|_1 + \frac{M}{c}\left\|\mathbf{u}_{\pi_h \rho_h} - \mathcal{I}_h \mathbf{u}_{\pi_h \rho_h}\right\|_1 \\
&\leq \left\|\mathbf{u}_{\pi_h \rho_h} - \mathbf{u}_{\rho^*}\right\|_1 + Ch\left|\mathbf{u}_{\pi_h \rho_h}\right|_2
\end{aligned} \tag{47}$$

where the second inequality follows from Cea's lemma [32] and last inequality follows from estimate (37). Hence $\mathbf{u}_{\rho_h,h} \to \mathbf{u}_{\rho^*}$ in $H^1(\Omega;\mathbb{R}^d)$ and so $J_h(\rho_h) \to J(\rho^*)$. Together with the inequality (46), we have

$$\tilde{J}(\rho^*) \leq \liminf_h \tilde{J}_h(\rho_h). \tag{48}$$

To establish optimality of ρ^*, take any $\rho \in \mathcal{A}_h$. The definition of ρ_h as the optimal solution to (38) implies

$$\tilde{J}_h(\rho_h) \leq \tilde{J}_h\left(i_h\rho\right). \tag{49}$$

Using a similar argument as above, we can pass (49) to the limit to show $\tilde{J}(\rho^*) \leq \tilde{J}(\rho)$.

5.2. The discrete problem

We proceed to obtain explicit expressions for the discrete problem (38) for a given finite element partition $\mathcal{T}_h$. For each $\rho_h \in \mathcal{A}_h$, we have the expansion $\rho_h(\mathbf{x}) = \sum_{k=1}^m [\mathbf{z}]_k\, \varphi_k(\mathbf{x})$ where $\mathbf{z}$ is the vector of nodal densities characterizing ρ_h and $\{\varphi_k\}_{k=1}^m$ the set of finite element basis functions for $\mathcal{A}_h$. We assume that the basis functions are such that for any $\mathbf{z} \in \left[\delta_\rho, 1\right]^m$, the associated density field lies in $\left[\delta_\rho, 1\right]$ everywhere, and conversely any density field in $\mathcal{A}_h$ can be identified with a vector of nodal densities in the closed cube $\left[\delta_\rho, 1\right]^m$. This is satisfied, for example, if $0 \leq \varphi_k \leq 1$ for all k. Moreover, the vector form for the Tikhonov regularization term is

$$\frac{\beta}{2}\left|\rho_h\right|_1^2 = \frac{1}{2}\mathbf{z}^T\mathbf{G}\mathbf{z} \tag{50}$$

where $\mathbf{G}$ is the usual finite element matrix defined by $[\mathbf{G}]_{k\ell} = \beta\int_\Omega \nabla\varphi_k \cdot \nabla\varphi_\ell \mathrm{d}\mathbf{x}$, which is positive semi-definite. Similarly, the volume term $\int_\Omega \rho \mathrm{d}\mathbf{x}$ can be written as $\mathbf{v}^T\mathbf{z}$ where $[\mathbf{v}]_k = \int_\Omega \varphi_k \mathrm{d}\mathbf{x}$.

With regards to the projection map π_h used in the discretization of the state equation (cf. (39)), there are several possible choices. One can simply take π_h to be the usual L^2-projection onto the space of piecewise constant fields. Here, we define $\pi_h\rho$ to be the piecewise constant field whose value over an element is equal to the value of ρ at the centroid of that element. More specifically, if $\mathbf{x}_e$ denotes the location of the centroid of element Ω_e, we have

$$\pi_h\rho = \sum_{e=1}^{l} \rho(\mathbf{x}_e)\chi_{\Omega_e} \tag{51}$$

where χ_{Ω_e} is the characteristic function associated with Ω_e (i.e., a function that takes value of 1 for $\mathbf{x} \in \Omega_e$ and zero otherwise).

If $\{\mathbf{N}_i\}_{i=1}^q$ denotes the basis functions for the displacement field such that $\mathbf{u}_h(\mathbf{x}) = \sum_{i=1}^q [\mathbf{U}]_i\mathbf{N}_i(\mathbf{x})$, the vector form of (39) is given by

$$\mathbf{KU} = \mathbf{F} \tag{52}$$

where the load vector $[\mathbf{F}]_i = \int_{\Gamma_N} \mathbf{t}\cdot\mathbf{N}_i \mathrm{d}s$ and the stiffness matrix is

$$[\mathbf{K}]_{ij} = \int_\Omega (\pi_h\rho_h)^p\, \mathbf{C}\boldsymbol{\epsilon}(\mathbf{N}_i) : \boldsymbol{\epsilon}(\mathbf{N}_j)\mathrm{d}\mathbf{x} = \sum_{e=1}^{l} [\rho_h(\mathbf{x}_e)]^p \int_{\Omega_e} \mathbf{C}\boldsymbol{\epsilon}(\mathbf{N}_i) : \boldsymbol{\epsilon}(\mathbf{N}_j)\mathrm{d}\mathbf{x}. \tag{53}$$

Let us define the matrix $\mathbf{P}$ whose (e,k)-entry is given by $[\mathbf{P}]_{ek} = \varphi_k(\mathbf{x}_e)$. Then

$$\rho_h(\mathbf{x}_e) = \sum_{k=1}^{m} \varphi_k(\mathbf{x}_e)\,[\mathbf{z}]_k = \sum_{k=1}^{m} [\mathbf{P}]_{ek}\,[\mathbf{z}]_k = [\mathbf{Pz}]_e\,. \tag{54}$$

The vector $\mathbf{Pz}$ thus gives the vector of elemental density values. Returning to (53) and denoting the *element stiffness* matrix by $[\mathbf{k}_e]_{ij} = \int_{\Omega_e} \mathbf{C}\boldsymbol{\epsilon}(\mathbf{N}_i) : \boldsymbol{\epsilon}(\mathbf{N}_j)\mathrm{d}\mathbf{x}$, we have the simplified expression for the global stiffness matrix

$$\mathbf{K}(\mathbf{z}) = \sum_{e=1}^{l} \left([\mathbf{Pz}]_e\right)^p \mathbf{k}_e. \tag{55}$$

The summation represents the assembly routine in practice. We note that the continuity and ellipticity of the bilinear form (cf. (3)) and non-degeneracy of the finite element partition imply that the eigenvalues of $\mathbf{K}(\mathbf{z})$ are bounded below by c_h and above by M_h (see chapter 9 of [32]) for all admissible density vectors $\mathbf{z} \in \left[\delta_\rho, 1\right]^m$.

The discrete optimization problem (38) can now be equivalently written as (with a slight abuse of notation for J and $\tilde{J}$)

$$\min_{\mathbf{z}\in[\delta_\rho,1]^m} \tilde{J}(\mathbf{z}) := J(\mathbf{z}) + \frac{1}{2}\mathbf{z}^T\mathbf{G}\mathbf{z} \tag{56}$$

where

$$J(\mathbf{z}) = \mathbf{F}^T\mathbf{U}(\mathbf{z}) + \lambda\mathbf{v}^T\mathbf{z} \tag{57}$$

and $\mathbf{U}(\mathbf{z})$ is the solution to $\mathbf{K}(\mathbf{z})\mathbf{U} = \mathbf{F}$. Observe that matrices $\mathbf{P}$ and $\mathbf{G}$, the vector $\mathbf{v}$, as well as the element stiffness matrices $\mathbf{k}_e$ and load vector $\mathbf{F}$ are all fixed and do not change in the course of optimization. Thus they can be computed once in the beginning and stored.

The gradient of J with respect to the nodal densities $\mathbf{z}$ can readily computed as

$$\partial_k J(\mathbf{z}) = -\mathbf{U}(\mathbf{z})^T\left(\partial_k\mathbf{K}\right)\mathbf{U}(\mathbf{z}) + \lambda[\mathbf{v}]_k. \tag{58}$$

The expression for $\partial_k\mathbf{K}$ can be obtained from (55). Defining the vector of strain energy densities $[\mathbf{E}(\mathbf{z})]_e = p\,[\mathbf{P}\mathbf{z}]_e^{p-1}\,\mathbf{U}(\mathbf{z})^T\mathbf{k}_e\mathbf{U}(\mathbf{z})$, we have

$$\nabla J(\mathbf{z}) = -\mathbf{P}^T\mathbf{E}(\mathbf{z}) + \lambda\mathbf{v}. \tag{59}$$

With the first order gradient information in hand, we can find the reciprocal approximation[4] of compliance, expanded about the point $\mathbf{y}$, as

$$J_{\mathsf{rec}}(\mathbf{z};\mathbf{y}) \equiv J(\mathbf{y}) + \lambda\left(\mathbf{z}-\mathbf{y}\right)^T\mathbf{v} + \sum_{k=1}^{m}\left(\frac{[\mathbf{y}]_k}{[\mathbf{z}]_k}\right)\left([\mathbf{z}]_k - [\mathbf{y}]_k\right)\left[-\mathbf{P}^T\mathbf{E}(\mathbf{y})\right]_k. \tag{60}$$

The Hessian of $J_{\mathsf{rec}}(\mathbf{z};\mathbf{y})$, evaluated at $\mathbf{z}=\mathbf{y}$, is a diagonal matrix with entries

$$h_k(\mathbf{y}) = \partial_{kk}J_{\mathsf{rec}}(\mathbf{y};\mathbf{y}) = \frac{2}{[\mathbf{y}]_k}\left[\mathbf{P}^T\mathbf{E}(\mathbf{y})\right]_k, \quad k = 1,\ldots,m. \tag{61}$$

The entries of the vector $\mathbf{E}(\mathbf{y})$ are non-negative for all admissible nodal densities but can be zero and therefore Hessian of $J_{\mathsf{rec}}(\mathbf{z};\mathbf{y})$ is only positive semi-definite.

6. Algorithms for the discrete problem

We begin with the generalization of the forward–backward algorithm for solving the discrete problem (56) before discussing the two-metric projection variation. As in Section 3, we consider a splitting algorithm with iterations of the form

$$\mathbf{z}_{n+1} = \underset{\mathbf{z}_n^{\mathsf{L}}\le\mathbf{z}\le\mathbf{z}_n^{\mathsf{U}}}{\operatorname{argmin}}\; Q_J(\mathbf{z};\mathbf{z}_n,\tau_n) + \frac{1}{2}\mathbf{z}^T\mathbf{G}\mathbf{z} \tag{62}$$

where, compared to (56), the regularization term is unchanged while J is replaced by the following local quadratic model around the current iterate $\mathbf{z}_n$

$$Q_J(\mathbf{z};\mathbf{z}_n,\tau_n) = J(\mathbf{z}_n) + \left(\mathbf{z}-\mathbf{z}_n\right)^T\nabla J(\mathbf{z}_n) + \frac{1}{2\tau_n}\left\|\mathbf{z}-\mathbf{z}_n\right\|_{\mathbf{H}_n}^2. \tag{63}$$

[4] The reciprocal approximation to the function $f(\mathbf{x})$ at a point $\mathbf{y}$ is given by $f_{\mathsf{rec}}(\mathbf{x}) = f(\mathbf{y}) + \sum_{k=1}^m\left[x_k^{-1}y_k\left(x_k-y_k\right)\partial_k f(\mathbf{y})\right]$. One can directly verify that $f_{\mathsf{rec}}(\mathbf{y}) = f(\mathbf{y})$ and $\nabla f_{\mathsf{rec}}(\mathbf{y}) = \nabla f(\mathbf{y})$.

The move limit constraint is accounted for through the bounds

$$\left[\mathbf{z}_n^{\mathsf{L}}\right]_k = \delta_\rho \wedge ([\mathbf{z}_n]_k - m_n)\,, \qquad \left[\mathbf{z}_n^{\mathsf{U}}\right]_k = 1 \vee ([\mathbf{z}_n]_k + m_n)\,, \quad k = 1, \ldots, m. \tag{64}$$

In order to embed the curvature information from the reciprocal approximation (60) in the quadratic model, we choose

$$\mathbf{H}_n = \operatorname{diag}(\hat{h}_1(\mathbf{z}_n), \ldots, \hat{h}_m(\mathbf{z}_n)) \tag{65}$$

where $\hat{h}_k(\mathbf{z}_n) \equiv h_k(\mathbf{z}_n) \wedge \delta_E$ and, as defined before, $0 < \delta_E \ll \lambda$ is a small positive constant. This modification not only ensures that $\mathbf{H}_n$ is positive definite but also that the eigenvalues of $\mathbf{H}_n$ are uniformly bounded above and below, a condition that is useful for the proof of convergence of the algorithm [34]. Observe that for all $\mathbf{z} \in \left[\delta_\rho, 1\right]^m$,

$$0 \leq h_k(\mathbf{z}) \leq 2\delta_\rho^{-1} \left\|\mathbf{E}(\mathbf{z})\right\|_\infty \leq 2p\delta_\rho^{-p-1} M_h \left\|\mathbf{U}(\mathbf{z})\right\|^2 \leq 2p\delta_\rho^{-p-1} M_h c_h^{-2} \left\|\mathbf{F}\right\|^2 \tag{66}$$

where we used the fact that $\mathbf{U}^T\mathbf{k}_e\mathbf{U} \leq \delta_\rho^{-p}\mathbf{U}^T\mathbf{K}(\mathbf{z})\mathbf{U} \leq \delta_\rho^{-p} M_h \left\|\mathbf{U}(\mathbf{z})\right\|^2$ and that the eigenvalues of $\mathbf{K}^{-1}$ are bounded above by c_h^{-1}.

The step size parameter τ_n in (62) must be sufficiently small so that the quadratic model is a conservative approximation and majorizes J. If $\tau_n > 0$ is chosen so that the update $\mathbf{z}_{n+1}$ satisfies

$$J(\mathbf{z}_{n+1}) \leq Q_J(\mathbf{z}_{n+1}; \mathbf{z}_n, \tau_n) \tag{67}$$

then one can show [34]

$$\tilde{J}(\mathbf{z}_n) - \tilde{J}(\mathbf{z}_{n+1}) \geq \frac{1}{2\tau_n} \left\|\mathbf{z}_n - \mathbf{z}_{n+1}\right\|^2_{\mathbf{H}_n}. \tag{68}$$

If $\mathbf{z}_n$ is a stationary point of $\tilde{J}$, that is if $(\mathbf{z} - \mathbf{z}_n)^T \nabla\tilde{J}(\mathbf{z}_n) \geq 0$ for all $\mathbf{z} \in \left[\delta_\rho, 1\right]^m$, then $\mathbf{z}_{n+1} = \mathbf{z}_n$ for all $\tau_n > 0$. To see this, we write (62) equivalently as

$$\min_{\mathbf{z}_n^{\mathsf{L}} \leq \mathbf{z} \leq \mathbf{z}_n^{\mathsf{U}}} (\mathbf{z} - \mathbf{z}_n)^T \nabla\tilde{J}(\mathbf{z}_n) + \frac{1}{2\tau_n} \left\|\mathbf{z} - \mathbf{z}_n\right\|^2_{\mathbf{H}_n + \tau_n\mathbf{G}}. \tag{69}$$

Since $\mathbf{H}_n + \tau_n\mathbf{G}$ is positive definite and $\mathbf{z}_n$ is a stationary point, the objective function is strictly positive for all $\mathbf{z} \in \left[\mathbf{z}_n^{\mathsf{L}}, \mathbf{z}_n^{\mathsf{U}}\right]$ with $\mathbf{z} \neq \mathbf{z}_n$ while it vanishes at $\mathbf{z}_n$, thereby establishing optimality of $\mathbf{z}_n$ for subproblem (62). Otherwise, if $\mathbf{z}_n$ is not a stationary point of $\tilde{J}$, then $\mathbf{z}_{n+1} \neq \mathbf{z}_n$ for sufficiently small τ_n, and (68) shows that there is a decrease in the objective function. This latter fact shows that *the algorithm is monotonically decreasing*.

A step size parameter satisfying (67) is guaranteed to exist if J has a Lipschitz gradient, that is, for some positive constant L,

$$\left\|\nabla J(\mathbf{z}) - \nabla J(\mathbf{y})\right\| \leq L \left\|\mathbf{z} - \mathbf{y}\right\|, \quad \forall \mathbf{z}, \mathbf{y} \in \operatorname{dom}(J). \tag{70}$$

One can show $J(\mathbf{z}) \leq Q_J(\mathbf{z}; \mathbf{z}_n, \tau_n)$ for all $\mathbf{z} \in \left[\delta_\rho, 1\right]^m$ if the step size satisfies

$$\tau_n^{-1}\mathbf{H}_n > L\mathbf{I} \tag{71}$$

in the sense of quadratic forms, i.e., $\tau_n^{-1}\mathbf{H}_n - L\mathbf{I}$ is positive definite [34]. This condition is in fact stronger than (67).

The step size τ_n can be selected with *a priori* knowledge of the Lipschitz constant L but this may be too conservative and may slow down the convergence of the algorithm. Instead, at each iteration, one can gradually decrease the step size via a backtracking routine until $\mathbf{z}_{n+1}$ satisfies (67). An alternative, possibly weaker, descent condition is the Armijo rule which requires that for some constant $0 < \nu < 1$, the update satisfies

$$\tilde{J}(\mathbf{z}_n) - \tilde{J}(\mathbf{z}_{n+1}) \geq \nu\,(\mathbf{z}_n - \mathbf{z}_{n+1})^T\, \nabla\tilde{J}(\mathbf{z}_n). \tag{72}$$

Though the implementation of such step size selection routine is straightforward, due to the high cost of function evaluations for the compliance problem (which requires solving the state equation to compute the value of J), the number of trials in satisfying the descent condition must be limited. Therefore, there is a tradeoff between attempting to choose a large step size to speed up convergence and the cost associated with the selection routine. As shown in the next section, we have found that fixing $\tau_n = 1$, generally eliminates the cost of backtracking routine and leads to a stable and convergent algorithm. In some cases, however, the overall cost can be reduced by using larger step sizes.

As with the derivation of (15) and (18), ignoring constant terms in $\mathbf{z}_n$ and rearranging, we can write (62) equivalently as

$$\mathbf{z}_{n+1} = \underset{\mathbf{z}_n^{\mathsf{L}} \leq \mathbf{z} \leq \mathbf{z}_n^{\mathsf{U}}}{\operatorname{argmin}} \ \left\| \mathbf{z} - \mathbf{z}_{n+1}^{*} \right\|_{\mathbf{H}_n + \tau_n \mathbf{G}}^{2} \tag{73}$$

where the interim update $\mathbf{z}_{n+1}^{*}$ is given by

$$\mathbf{z}_{n+1}^{*} = \mathbf{z}_n - \tau_n \left(\mathbf{H}_n + \tau_n \mathbf{G} \right)^{-1} \left[\nabla \tilde{J}(\mathbf{z}_n) \right]. \tag{74}$$

With the appropriate choice of step size (satisfying any one of the conditions (67), (71), or (72)) and boundedness of $\mathbf{H}_n$, it can be shown that every limit point of the sequence $\mathbf{z}_n$ generated by the algorithm is a critical point of $\tilde{J}$. For the particular case of quadratic regularization, it is evident from (74) that the algorithm reduces to the so-called scaled gradient projection algorithm, and the convergence proof can be found in [34]. A more general proof can be found in the review paper [35] on proximal splitting method though the metric associated with the proximal term, i.e., $\|\mathbf{z} - \mathbf{z}_n\|_{\mathbf{H}_n + \tau_n \mathbf{G}}^{2}$ in (62), is fixed there.

As seen from (62) or (73), the forward–backward algorithm requires the solution to a sparse, strictly convex quadratic program subject to simple bound constraints which can be efficiently solved using a variety of methods, e.g., the active set method. Alternatively, the projection of $\mathbf{z}_{n+1}^{*}$ can be recast as a bound constrained sparse least squares problem and solved using algorithms in [36].

Two-metric projection variation

Next we discuss a variation of the splitting algorithm that simplifies the projection step (73) by augmenting the interim density (74). More specifically, we adopt a variant of the two-metric projection method [10,11], in which the norm in (73) is replaced by the usual Euclidean norm, and the scaling matrix $\mathbf{H}_n + \tau_n \mathbf{G}$ in the interim step (74) is made diagonal with respect to the active components of $\mathbf{z}_n$.

Let $I_n = I_n^{\mathsf{L}} \cup I_n^{\mathsf{U}}$ denote the set of active constraints where

$$I_n^{\mathsf{L}} = \left\{ k : [\mathbf{z}_n]_k \leq \delta_\rho + \epsilon \text{ and } \left[\nabla \tilde{J}(\mathbf{z}_n) \right]_k > 0 \right\} \tag{75}$$

$$I_n^{\mathsf{U}} = \left\{ k : [\mathbf{z}_n]_k \geq 1 - \epsilon \text{ and } \left[\nabla \tilde{J}(\mathbf{z}_n) \right]_k < 0 \right\}. \tag{76}$$

Here ϵ is an algorithmic parameter (we fix it at 10^{-3} for the numerical results) that enlarges the set of active constraints in order to avoid the discontinuities that may otherwise arise [10]. Then

$$[\mathbf{D}_n]_{ij} \equiv \begin{cases} 0 & \text{if } i \neq j \text{ and } i \in I_n \text{ or } j \in I_n \\ [\mathbf{H}_n + \tau_n \mathbf{G}]_{ij} & \text{otherwise} \end{cases} \tag{77}$$

is a scaling matrix formed from $\mathbf{H}_n + \tau_n \mathbf{G}$ that is diagonal with respect to I_n and therefore removes the coupling between the active and free constraints. The operation in (77) essentially consists of zeroing out all the off-diagonal entries of $\mathbf{H}_n + \tau_n \mathbf{G}$ for the active components. Note that any other positive matrix with the same structure as $\mathbf{D}_n$ can be used. The new interim density is then defined as

$$\mathbf{z}_{n+1}^{*} = \mathbf{z}_n - \tau_n \mathbf{D}_n^{-1} \left[\nabla \tilde{J}(\mathbf{z}_n) \right] \tag{78}$$

and the next iterate is given by the Euclidean projection of this interim density onto the constraint set

$$\mathbf{z}_{n+1} = \underset{\mathbf{z}_n^{\mathsf{L}} \leq \mathbf{z} \leq \mathbf{z}_n^{\mathsf{U}}}{\operatorname{argmin}} \ \left\| \mathbf{z} - \mathbf{z}_{n+1}^{*} \right\|^{2} \tag{79}$$

which has the following explicit solution

$$\left[\mathbf{z}_{n+1} \right]_k = \left(\left[\mathbf{z}_n^{\mathsf{L}} \right]_k \wedge \left[\mathbf{z}_{n+1}^{*} \right]_k \right) \vee \left[\mathbf{z}_n^{\mathsf{U}} \right]_k, \quad k = 1, \ldots, m. \tag{80}$$

Since $\mathbf{D}_n^{-1}\nabla\tilde{J}(\mathbf{z}_n)$ can be viewed as the gradient of $\tilde{J}$ with respect to the metric induced by $\mathbf{D}_n$, we can see that the present algorithm consisting of (78) and (79) utilizes two separate metrics for differentiation and projection operations. The significant computational advantage of carrying out the projection step with respect to the Euclidean norm is due to the particular structure of the constraint set. Compared to the forward–backward algorithm discussed before, at the cost of modifying the scaling matrix, the overhead associated with solving the quadratic program (cf. (73)) is eliminated.

As in the previous algorithm, one can show that $\mathbf{z}_n$ is a critical point of $\tilde{J}$ if and only if $\mathbf{z}_{n+1} = \mathbf{z}_n$ for all $\tau_n > 0$. Similarly, if $\mathbf{z}_n$ is not a stationary point, then for a sufficiently small step size, the next iterate decreases the value of the cost function, i.e., $\tilde{J}(\mathbf{z}_{n+1}) < \tilde{J}(\mathbf{z}_n)$. The choice of τ_n can be again obtained from an Armijo-type condition along the projection arc (cf. [10]), namely,

$$\tilde{J}(\mathbf{z}_n) - \tilde{J}(\mathbf{z}_{n+1}) \geq \nu \mathbf{d}_n^T \nabla\tilde{J}(\mathbf{z}_n) \tag{81}$$

where the direction vector $\mathbf{d}_n$ is given by

$$[\mathbf{d}_n]_k = \begin{cases} [\mathbf{z}_n]_k - \left[\mathbf{z}_{n+1}\right]_k & k \in I_n \\ \left[\tau_n \mathbf{D}_n^{-1}\nabla\tilde{J}(\mathbf{z}_n)\right]_k & k \notin I_n. \end{cases} \tag{82}$$

In the next section, we will compare the performance of the forward–backward algorithm consisting of (73) and (74) with the two-metric projection consisting of (78) and (80).

7. Numerical investigations

For all the results presented in this section, the constituent material $\mathbf{C}$ is assumed to be isotropic with unit Young's modulus and Poisson ratio of 0.3. The lower bound on the density is set to $\delta_\rho = 10^{-3}$ and, unless otherwise stated, the SIMP penalty exponent is fixed at $p = 3$. The following backtracking algorithm is used to determine the value of the step size parameter: given constants $\tau_0 > 0$ and $0 < \sigma < 1$, the step size parameter in the nth iteration is given by

$$\tau_n = \sigma^{k_n}\tau_0 \tag{83}$$

where k_n is the smallest non-negative integer such that τ_n satisfies (72) or (81). In practice, this means that we begin with the initial step size τ_0 and reduce it by a factor of σ until descent conditions are satisfied. The descent parameter is set to $\nu = 10^{-3}$ and the backtracking parameter is $\sigma = 0.6$. Note that larger ν leads to a more severe descent requirement and subsequently smaller τ_n. Similarly, smaller σ reduces the step size parameter by a larger factor which can decrease the number of backtracking steps. Note, however, that using small step sizes may lead to slow convergence of the algorithm.

Since each backtracking step involves evaluating the cost functional and therefore solving the state equation, as a measure of computational cost, we keep track of the total number of backtracking steps (i.e., $\sum_n k_n$) in addition to the total number of iterations. The convergence criteria adopted here is based on the relative decrease in the objective function

$$E_1 = \frac{\left|\tilde{J}(\mathbf{z}_{n+1}) - \tilde{J}(\mathbf{z}_n)\right|}{\left|\tilde{J}(\mathbf{z}_n)\right|} \leq \epsilon_1 \tag{84}$$

and the satisfaction of the first order conditions of optimality according to

$$E_2 = \frac{\left\|\Pi[\mathbf{z}_{n+1} - \nabla\tilde{J}(\mathbf{z}_{n+1})] - \mathbf{z}_{n+1}\right\|}{\left\|\mathbf{z}_{n+1}\right\|} \leq \epsilon_2. \tag{85}$$

Here Π is the Euclidean projection onto the constraint set $\left[\delta_\rho, 1\right]^m$ defined by $[\Pi(\mathbf{y})]_i = \left(\delta_\rho \wedge [\mathbf{y}]_i\right) \vee 1$. Unless otherwise stated, we have selected $\epsilon_1 = 10^{-5}$ and $\epsilon_2 = 10^{-4}$.

MBB beam problem

The model compliance minimization problem adopted here is the benchmark MBB beam problem, whose domain geometry and prescribed loading and boundary conditions are shown in Fig. 5. Using appropriate boundary conditions,

Table 1
Summary of influence of various factors in the algorithm for the MBB problem with $\beta = 0.06$. The acronyms FBS, TMP, MMA, and GP designate the forward–backward, two-metric projection, Method of Moving Asymptotes, and the gradient projection algorithms, respectively. Fourth and fifth columns show the total number of iterations and backtracking steps. The remaining columns show the final value of compliance $\ell(\mathbf{u}_\rho)$, regularization term $R(\rho) = \beta\,|\rho|_1^2/2$, volume fraction $V(\rho) = |\Omega|^{-1}\int_\Omega \rho \mathrm{d}\mathbf{x}$, the regularized objective $\tilde{J}(\rho)$, the relative change in cost function value E_1 and the error in satisfaction of the first order conditions of optimality E_2. The asterisk indicates that the maximum allowed iteration count of 1000 was reached before the convergence criteria was met.

Algorithm	$\mathbf{H}_n$	τ_0	# it.	# bt.	$\ell(\mathbf{u}_\rho)$	$R(\rho)$	$V(\rho)$	$\tilde{J}(\rho)$	E_1	E_2
FBS	Identity	1.0	316	0	100.019	8.553	0.512	210.965	9.962e−6	9.943e−5
FBS	Identity	2.0	215	154	100.093	8.537	0.511	210.914	9.178e−6	5.812e−5
FBS	Reciprocal	1.0	186	0	99.937	8.594	0.513	211.032	9.769e−6	9.363e−5
FBS	Reciprocal	2.0	91	39	100.095	8.568	0.512	211.008	4.926e−6	9.746e−5
TMP	Identity	1.0	330	0	100.076	8.533	0.512	210.951	9.958e−6	9.973e−5
TMP	Identity	2.0	151	78	100.060	8.556	0.512	210.938	9.639e−6	5.900e−5
TMP	Reciprocal	1.0	179	0	99.943	8.592	0.513	211.031	9.878e−6	9.453e−5
TMP	Reciprocal	2.0	85	34	100.078	8.578	0.512	210.999	9.043e−6	8.074e−5
GP	Identity	0.25	1000*	0	100.108	8.563	0.510	210.736	4.094e−6	1.557e−4
GP	Identity	0.50	568	79	100.241	8.560	0.509	210.685	4.384e−6	9.691e−5
MMA	–	–	1000*	–	98.871	9.889	0.523	213.388	5.795e−7	1.764e−4

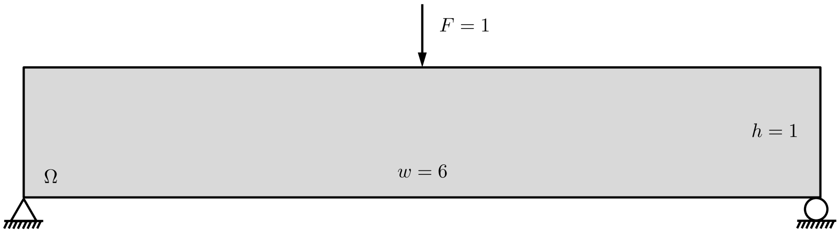

Fig. 5. The design domain and boundary conditions for the MBB beam problem.

the symmetry of the problem is exploited to pose and solve the state equation only on half of the extended domain. The volume penalty parameter is $\lambda = 200/\,|\Omega|$ where $|\Omega|$ is the area of the extended design domain.

We begin with the investigation of the behavior of the two splitting algorithms with different choice of parameters discussed in the previous section. In particular, we compare the forward–backward algorithm with the two-metric projection method and investigate the influence of the Hessian approximation. In addition to the choice of $\mathbf{H}_n$ defined by (65), we also consider a fixed scaling of the identity matrix

$$\mathbf{H}_n \equiv \alpha \mathbf{I}, \quad n = 1, 2, \ldots \tag{86}$$

for which the algorithm becomes the basic forward–backward algorithm of [7] with the same proximal term in every iteration. The scaling coefficient α is set to $4\lambda A$ where A is the area of an element. This choice is made so that the step size parameter τ_n is on the same order of magnitude as with “reciprocal” Hessian. The other parameter investigated here is the initial step size parameter τ_0 and we consider two choices $\tau_0 = 1$ and $\tau_0 = 2$. In all cases, the move limit is fixed at $m_n = 1$ for all n and thus $\mathcal{A}_n = \mathcal{A}$.

The domain for the MBB beam is discretized with a grid of 300 by 50 bilinear quadrilateral elements and the Tikhonov regularization parameter is set to $\beta = 0.06$. The initial guess in all cases is taken to be uniform density field $\rho_h \equiv 1/2$. All the possible combinations of the above choices produce the same final topology, similar to the representative solution shown in Fig. 6. *This shows the framework exhibits stable convergence to the same final solution and is relatively insensitive to various choices of algorithmic parameters for this level of regularization.* What is different, however, is the speed of convergence and the required computational effort as measured by the number of the backtracking steps, total number of iterations, and cost per iteration. The results are summarized in Table 1.

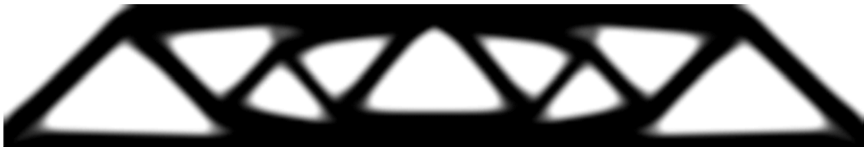

Fig. 6. Final density field for the MBB problem and $\beta = 0.06$ plotted in grayscale. This result was generated using the TMP algorithm with $\tau_0 = 2$ and $m_n = 1$.

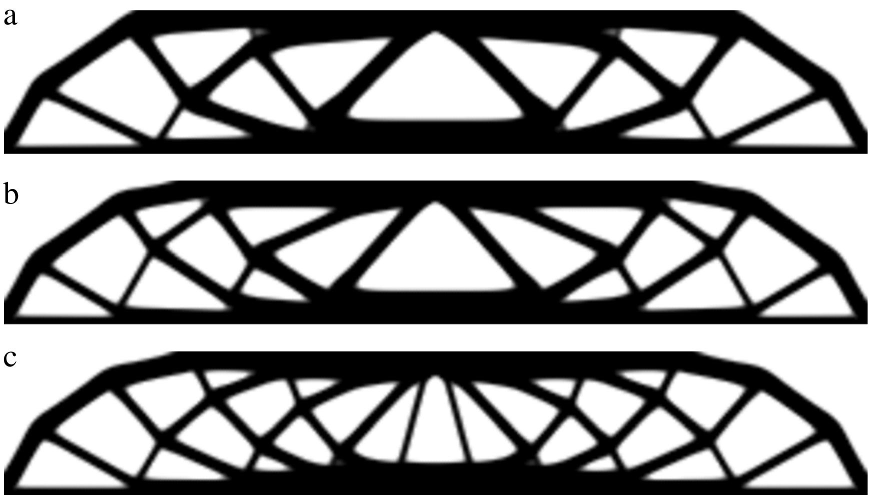

Fig. 7. Final densities plotted in grayscale for the MBB problem and $\beta = 0.01$. The results are generated using the TMP algorithm with (a) $\tau_0 = 2$, $m_n = 1$, (b) $\tau_0 = 1, m_n = 1$ and (c) $\tau_0 = 1, m_n = 0.03$.

First we note that the initial step size $\tau_0 = 1$ does not lead to any backtracking steps which means that at each iteration the step size parameter is $\tau_n = 1$. By contrast, using the larger initial step size parameter $\tau_0 = 2$ sometimes requires backtracking steps to satisfy the descent condition but substantially reduces the total number of iterations. Moreover, in all cases, the constant Hessian (86) requires nearly twice as many iterations and backtracking steps compared to the "reciprocal" Hessian. *This highlights the fact that embedding the reciprocal approximation of compliance does indeed lead to faster convergence.* Overall, the best performance is obtained using the reciprocal approximation and larger initial step size parameter.

For this problem, the forward–backward algorithm and the two-metric projection method have roughly the same number of iterations and backtracking steps. However, the cost per iteration for the two-metric projection is significantly lower since the projection step is computationally trivial. Therefore, the two-metric projection is more efficient.

Since the splitting algorithm presented here is a first-order method, it is also appropriate to compare its performance to the gradient projection algorithm, which is among the most basic first-order methods for solving constrained optimization problems. The next iterate in the gradient projection method is simply the projection of the unconstrained gradient descent step onto the admissible space. In the absence of move limits and in the discrete setting, we have the following update expression

$$\mathbf{z}_{n+1} = \underset{\mathbf{z}\in[\delta_\rho,1]^m}{\operatorname{argmin}} \left\| \mathbf{z} - \left[\mathbf{z}_n - \frac{\tau_n}{\alpha} \nabla \tilde{J}(\mathbf{z}_n) \right] \right\|^2 \tag{87}$$

where the scaling parameter $\alpha = 4\lambda A$ is defined as before in order to allow for a direct comparison with the forward–backward splitting in the case of $\mathbf{H}_n = \alpha \mathbf{I}$. We determine the step size parameter τ_n at each iteration using the backtracking procedure (83) based on the Armijo-type descent condition (72). Note that due to the simple structure of the constraint set, computing the gradient $\nabla \tilde{J}$ constitutes the main computational cost of the gradient projection algorithm at each iteration. Table 1 shows the results for the same problem for two different choices of initial step size parameter τ_0. First observe that the step sizes are smaller compared to the forward–backward algorithm, a fact that can be seen from the equivalent expression for (87) given by

$$\mathbf{z}_{n+1} = \underset{\mathbf{z}\in[\delta_\rho,1]^m}{\operatorname{argmin}} \ \tilde{J}(\mathbf{z}_n) + (\mathbf{z}-\mathbf{z}_n)^T \nabla \tilde{J}(\mathbf{z}_n) + \frac{1}{2\tau_n} \|\mathbf{z}-\mathbf{z}_n\|^2_{\alpha\mathbf{I}} . \tag{88}$$

This shows that at each iteration, we construct a quadratic model for the composite objective $\tilde{J}$. By contrast, the quadratic model in (62) is only used for J and the regularization term appears exactly. Since $\nabla\tilde{J}$ has a larger Lipschitz constant compared to ∇J, it is therefore expected that τ_n must be smaller to ensure descent. It is also instructive to recall the informal derivation of the forward–backward algorithm in [7] where the main difference with the gradient projection algorithm was the use of a semi-implicit (in place of an explicit) temporal discretization of the gradient flow equation. Note that the gradient projection algorithm converged to the same solution as before (cf. Fig. 6) though in the case of $\tau_0 = 0.25$, the convergence was too slow and we terminated the algorithm after 1000 iterations.

We also tested the performance of MMA [29] since it is perhaps the most widely used algorithm in the topology optimization literature. We followed the common practice and used the algorithm as a black-box optimization routine. In particular, we provided the algorithm with the gradient of composite objective $\tilde{J}$ and did not make any changes to the open source code provided by Svanberg.[5] MMA internally generates a separable convex approximation to $\tilde{J}$ using reciprocal-type expansions with appropriately defined and updated asymptotes. Though such approximations are suitable for the structural term, they may be inaccurate for the Tikhonov regularizer and thus $\tilde{J}$. As shown in Table 1, MMA did not converge (according to the convergence criteria described earlier) in 1000 iterations before it was terminated. Furthermore, not only was the final value of the objective function larger than that obtained by gradient projection or either splitting algorithm, the final density was also topologically different from the solution shown in Fig. 6.

Next we investigate the performance of the algorithm for a smaller value of the regularization parameter which is expected to produce more complex topologies. For the next set of results, we set $\beta = 0.01$. In all cases considered, the forward–backward and the two-metric projection algorithms both give identical final topologies with roughly the same number of iterations and so we only report the results for the two-metric projection algorithm. Also, as demonstrated by the first study, the use of reciprocal approximation leads to better and faster convergence of the algorithm so we limit the remaining results to the "reciprocal" $\mathbf{H}_n$. The tolerance level $\epsilon_2 = 10^{-4}$ for satisfaction of the optimality condition is relatively stringent in this case due to the complexity of final designs (compared to $\beta = 0.06$) and leads to a large number of iterations with little change in density near the optimum. We therefore increase the tolerance to $\epsilon_2 = 2 \times 10^{-4}$ which gives nearly identical final topologies but with fewer iterations.

We examine the influence of the step size parameter and move limit, which unlike the previous case of large regularization parameter, can lead to different final solutions. We consider two possible initial step size parameters $\tau_0 = 1$ and $\tau_0 = 2$, as well as two choices for the move limit $m_n \equiv 1$ and $m_n \equiv 0.03$. Here we are using a fixed move limit m_n for all iterations n. It may be possible to devise a strategy to increase m_n in the latter stages of optimization to improve convergence. The results are summarized in Table 2 and the final solutions are shown in Fig. 7.

First note that with no move limit constraints, i.e., $m_n = 1$, the final solution with the more aggressive choice of the initial step size parameter ($\tau_0 = 2$) is less complex and has fewer members compared to $\tau_0 = 1$, which as before does not require any backtracking steps. Note, however, that the more aggressive scheme in fact requires more iterations to converge. In the presence of move limits, there is no backtracking step with either choice of step size but the larger step size does reduce the total number of iterations. The final topologies are identical and have more members compared to the solutions obtained without the move limits. It is interesting to note that the overall iteration count is lowest for $\tau_0 = 2$ and $m_n = 0.03$ despite the limit on the change in density at each iteration. As noted earlier, the use of move limits can stabilize the convergence of the topology optimization problem.

The overall trend that the more aggressive choice of parameters produce less complex final solutions is due to the fact that member formation occurs early on in the algorithm. The most aggressive algorithm ($\tau_0 = 2$, $m_n = 1$) still produces the best solution as measured by $\tilde{J}$ while the solution obtained enforcing the move limit $m_n = 0.03$ has the lowest value of compliance J.

[5] We remark that MMA is used with some modifications by Borrvall in [16] where the behavior of various regularizations schemes, including Tikhonov regularization, are compared.

Table 2
Summary of the results for the MBB problem with $\beta = 0.01$.

Algorithm	τ_0	m_n	# it.	# bt.	$\ell(\mathbf{u}_\rho)$	$R(\rho)$	$V(\rho)$	$\tilde{J}(\rho)$	E_1	E_2
TMP	1	1	138	0	102.306	4.669	0.474	201.779	6.989e−6	1.978e−5
TMP	2	1	169	62	102.716	4.075	0.472	201.189	9.780e−6	1.679e−5
TMP	1	0.03	153	0	100.738	5.185	0.486	203.014	7.217e−6	1.998e−4
TMP	2	0.03	98	0	100.568	5.173	0.486	202.970	9.795e−6	1.566e−4

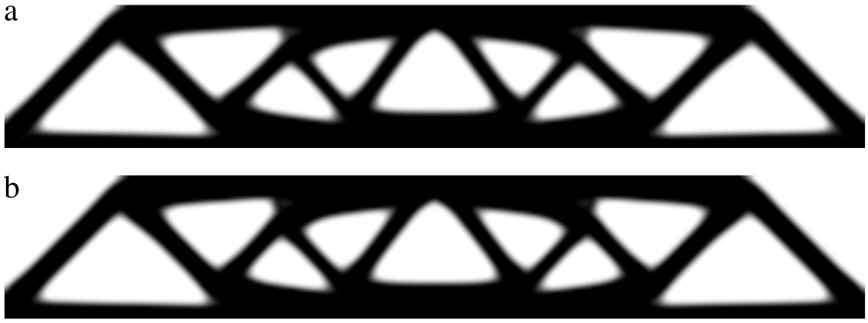

Fig. 8. Final densities plotted in grayscale for the MBB problem with $\beta = 0.06$ and SIMP penalty exponent (a), $p = 4$ (b) $p = 5$.

We note that aside from the higher degree of complexity, the optimal densities for $\beta = 0.01$ contain fewer intermediate values compared to the solution for $\beta = 0.06$. One measure of discreteness, used in [5], is given by

$$M(\rho) = \frac{1}{|\Omega|}\int_{\Omega} 4\left(\rho - \delta_\rho\right)(1-\rho)\,\mathrm{d}\mathbf{x} \tag{89}$$

which is equal to zero if ρ takes only values of δ_ρ and 1. For the solutions shown in Fig. 7, $M(\rho)$ is equal to 6.98%, 7.64% and 8.90% from top to bottom, respectively. By contrast, the optimal density for $\beta = 0.06$ (cf. Fig. 6) has a discreteness measure of 15.0%. By increasing the value of the SIMP exponent p, the optimal densities can be made more discrete. The results for $\beta = 0.06$ using $p = 4$ and $p = 5$ are shown in Fig. 8. While the optimal topologies are nearly identical to the solution for $p = 3$, the discrete measure is lowered to 13.1% and 12.1%, respectively. Observe, however, that the layer of intermediate densities around the boundary cannot be completely eliminated even when p is increased to a very large value since the Tikhonov regularizer is unbounded in the discontinuous limit of density.

As shown in the previous section, the optimal solutions to the discrete problem converge to an optimal solution of the continuum problem as the finite element mesh is refined. We next demonstrate numerically that solutions produced by the present optimization algorithms appear to be stable with respect to mesh refinement. We do so for $\beta = 0.01$ using the two-metric projection algorithm with $\tau_n \equiv 1$ where the final topology is relatively complex and the algorithm is expected to be more sensitive. As shown in Fig. 9, we solve the problem using finer grids consisting of 600 × 100 and 1200 × 200 bilinear quadrilateral elements, which required 104 and 106 iterations, respectively. The final density distribution is nearly identical indicating convergence of optimal densities in the L^2-norm.

Explicit enforcement of volume constraint

We proceed to discuss the forward–backward splitting algorithm in the presence of an explicit constraint on the volume of the design, as topology optimization problems are often formulated. Compared to the compliance minimization problem considered so far with a volume penalty term appearing in the cost function (cf. Section 5.2), here we have a given upper bound volume fraction $0 \le \Theta \le 1$ limiting the volume of the design to $\Theta\,|\Omega|$. We are thus asked to solve the following problem

$$\min_{\mathbf{z}\in[\delta_\rho,1]^m,\, V(\mathbf{z})\le 0} \bar{J}(\mathbf{z}) + \frac{1}{2}\mathbf{z}^T\mathbf{G}\mathbf{z} \tag{90}$$

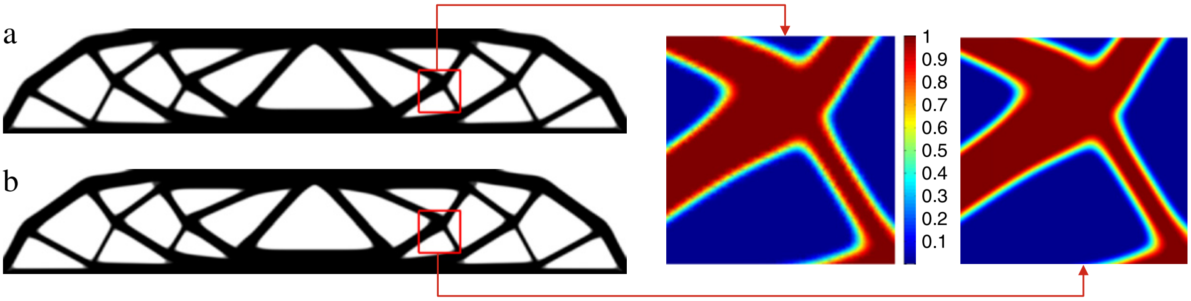

Fig. 9. Results of the mesh refinement study with (a) 600 × 100 (b) 1200 × 200 elements.

where $\overline{J}(\mathbf{z}) = \mathbf{F}^T\mathbf{U}(\mathbf{z})$ is the compliance and $V(\mathbf{z}) = \mathbf{v}^T\mathbf{z} - \Theta\,|\Omega|$ is the volume of the design associated with $\mathbf{z}$. As in Section 6, the forward–backward splitting algorithm for this problem involves iterations where $\overline{J}$ is replaced by a convex quadratic approximation at the current design point with the regularization and constraints left intact. Thus, the nth iteration is simply defined by

$$\mathbf{z}_{n+1} = \underset{\mathbf{z}_n^{\mathsf{L}} \le \mathbf{z} \le \mathbf{z}_n^{\mathsf{U}},\, V(\mathbf{z}) \le 0}{\operatorname{argmin}} \quad Q_{\overline{J}}(\mathbf{z}; \mathbf{z}_n, \tau_n) + \frac{1}{2}\mathbf{z}^T\mathbf{G}\mathbf{z} \tag{91}$$

where, as before, $Q_{\overline{J}}(\mathbf{z}; \mathbf{z}_n, \tau_n) = \overline{J}(\mathbf{z}_n) + (\mathbf{z} - \mathbf{z}_n)^T\,\nabla\overline{J}(\mathbf{z}_n) + (2\tau_n)^{-1}\,\|\mathbf{z} - \mathbf{z}_n\|^2_{\mathbf{H}_n}$, with step size determined by the Armijo condition, and the bounds are given in (64), accounting for the move limits introduced. Note that there is no need for approximation of the volume constraint, since in the context of SIMP formulation, it is a linear constraint. The subproblem (91) is a sparse convex quadratic program subject to linear constraints, and can thus be solved using an appropriate large-scale quadratic programming algorithm. In our numerical studies, we have employed the interior-point algorithm that is built in Matlab. We should also note that the subproblems need not be solved very accurately in the early iterations, assuming that the approximation solutions still respect the bound constraints. The proof of convergence of this algorithm is similar to that of the bound-constrained problem discussed in Section 6.

We will also consider here a heuristic version of the two-metric projection algorithm for solving (90). To this end, we note that the dual problem for (91) is given by

$$\max_{\lambda \ge 0}\left[\min_{\mathbf{z}_n^{\mathsf{L}} \le \mathbf{z} \le \mathbf{z}_n^{\mathsf{U}}} Q_{\overline{J}}(\mathbf{z}; \mathbf{z}_n, \tau_n) + \frac{1}{2}\mathbf{z}^T\mathbf{G}\mathbf{z} + \lambda V(\mathbf{z})\right] = \max_{\lambda \ge 0}\left[\min_{\mathbf{z}_n^{\mathsf{L}} \le \mathbf{z} \le \mathbf{z}_n^{\mathsf{U}}} Q_{J_\lambda}(\mathbf{z}; \mathbf{z}_n, \tau_n) + \frac{1}{2}\mathbf{z}^T\mathbf{G}\mathbf{z}\right] \tag{92}$$

where $J_\lambda(\mathbf{z}) = \overline{J}(\mathbf{z}) + \lambda V(\mathbf{z})$. The minimization is simply the bound-constrained problem of Section 6 with λ as the volume penalty parameter, and the outer maximization is a one-dimensional problem which can be solved, for instance, using a simple bisection algorithm. Inspired by this, we consider the iteration defined by the two-metric projection method of Section 6 as a surrogate for solution of the inner minimization, with λ itself determined through the bi-section algorithm. In other words, similar to the OC method for the volume-constrained problem (cf. [37]), each iteration utilizes the bi-section method for determining the volume penalty parameter and, for each λ, the candidate iteration is defined by (80).

The results for the MBB problem, with prescribed volume fraction $\Theta = 0.5$, and different initial step size τ_0 for both algorithms are shown in Table 3. For these results, $\mathbf{H}_n$ is the "reciprocal" Hessian defined in (65) and $m_n = 1$. We can see that a large initial step size leads to faster convergence but requires more backtracking steps. Also, the forward–backward splitting and the heuristic two-metric algorithm exhibited very similar trajectories, as evidenced by a similar number of iterations and backtracking steps and nearly identical final solutions (a representative solution is shown in Fig. 10). By comparison, the solution produced by MMA, though similar, has a larger value of compliance and regularizer.

Compliant mechanism design

The discussion so far has been limited to the problem of compliance minimization which, as noted earlier, is self-adjoint and its gradient has the same sign. We conclude this section with the design of a compliant force inverter for

Table 3
Summary of the results for the MBB problem with $\beta = 0.06$ and prescribed volume fraction of $\Theta = 0.5$. Here $\tilde{J}(\rho) = J(\rho) + R(\rho)$. Convergence was determined only using the relative decrease in the objective function $\tilde{J}$, as measured by E_1.

Algorithm	τ_0	# it.	# bt.	$\ell(\mathbf{u}_\rho)$	$R(\rho)$	$V(\rho)$	$\tilde{J}(\rho)$	E_1
FBS	1.0	315	0	102.478	7.433	0.500	109.912	9.664e−6
FBS	2.0	146	62	102.557	7.490	0.500	110.047	2.741e−6
TMP	1.0	325	0	102.478	7.434	0.500	109.912	9.806e−6
TMP	2.0	154	62	102.601	7.520	0.500	110.116	3.990e−6
MMA	–	279	–	103.985	9.342	0.495	113.332	6.177e−6

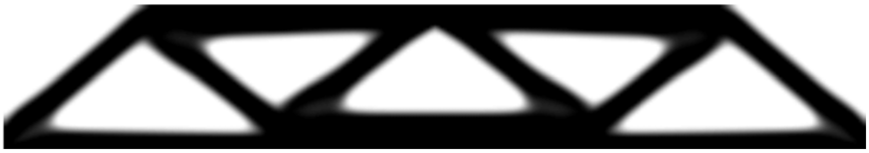

Fig. 10. Final density field for the MBB problem with $\beta = 0.06$ and subject to volume constraint with $\Theta = 0.5$. This result was generated using the FBS algorithm with $\tau_0 = 2$.

which the cost functional is no longer self-adjoint and therefore, unlike compliance, the gradient field may take both negative and positive values in the domain.

The objective of the mechanism design is to identify a structure that maximizes the force exerted on a workpiece under the action of an external actuator. As illustrated in Fig. 11, the force inverter transfers the input force of the actuator to a force at the prescribed output location in the opposite direction. We assume in this setting that both the workpiece and the actuator are elastic and their stiffness are represented by vector fields $\mathbf{k}_1 \in L^\infty(\Gamma_{S_1})$ and $\mathbf{k}_2 \in L^\infty(\Gamma_{S_2})$, respectively. Here $\Gamma_{S_1}, \Gamma_{S_2}$ are segments of the traction boundary $\Gamma_N \subseteq \partial\Omega$ where the structure is interacting with these elastic bodies. The tractions experienced by the structure through this interaction for a displacement field $\mathbf{u}$ can be written as

$$\mathbf{t}_{S_r}(\mathbf{u}) = -(\mathbf{k}_r \cdot \mathbf{u})\frac{\mathbf{k}_r}{\|\mathbf{k}_r\|}, \quad \text{on } \Gamma_{S_r} \text{ for } r = 1, 2. \tag{93}$$

Accordingly, the displacement $\mathbf{u}_\rho$ for a given distribution of material ρ in Ω is the solution to the following boundary problem

$$a(\mathbf{u}_\rho, \mathbf{v}; \rho) + a_s(\mathbf{u}_\rho, \mathbf{v}) = \ell(\mathbf{v}) \quad \forall \mathbf{v} \in \mathcal{V} \tag{94}$$

where

$$a_s(\mathbf{u}, \mathbf{v}) = \sum_{r=1,2} \int_{\Gamma_{S_r}} \frac{(\mathbf{k}_r \cdot \mathbf{u})\,(\mathbf{k}_r \cdot \mathbf{v})}{\|\mathbf{k}_r\|} \mathrm{d}s. \tag{95}$$

The cost functional for the mechanism design problem is defined as

$$\overline{J}(\rho) = -\int_{\Gamma_{S_1}} \mathbf{k}_1 \cdot \mathbf{u}_\rho \mathrm{d}s \tag{96}$$

which is a measure of the (negative of) force applied to the workpiece in the direction of $\mathbf{k}_1$, as seen from the following relation:

$$\int_{\Gamma_{S_1}} \mathbf{k}_1 \cdot \mathbf{u}_\rho \mathrm{d}s = \int_{\Gamma_{S_1}} \left[-\mathbf{t}_{S_1}(\mathbf{u}_\rho)\right] \cdot \frac{\mathbf{k}_1}{\|\mathbf{k}_1\|} \mathrm{d}s. \tag{97}$$

Viewed another way, the minimization of (96) amounts to maximizing the displacement of the structure at the location of the workpiece in the direction of $\mathbf{k}_1$.

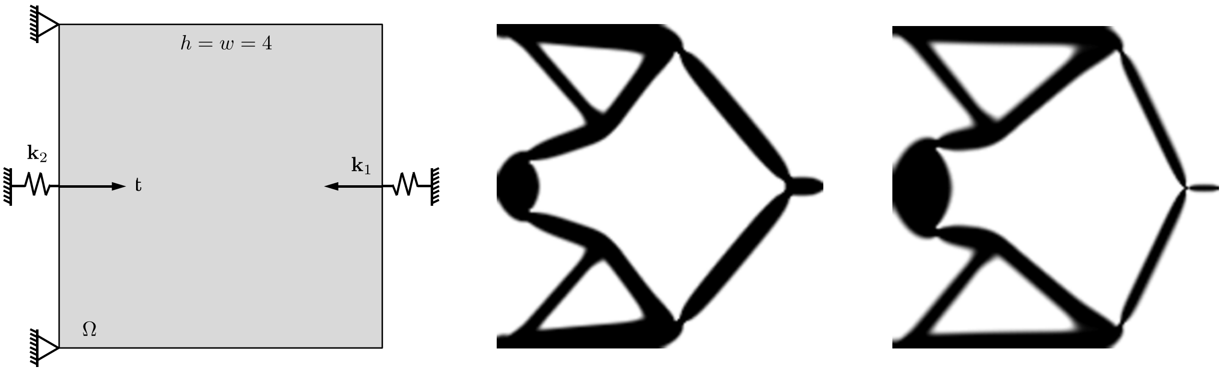

Fig. 11. The design domain and boundary conditions for the force inverter problem (left), the optimal topology for $\|\mathbf{k}_1\| = 1$, $\|\mathbf{k}_2\| = 0.02$ and $\beta = 10^{-4}$ (middle), and optimal topology for $\|\mathbf{k}_1\| = 1$, $\|\mathbf{k}_2\| = 10^{-3}$ and $\beta = 10^{-3}$. For both problems, the prescribed volume fraction is $\Theta = 0.25$.

The cost functional, in the discrete setting, is given by

$$\overline{J}(\mathbf{z}) = -\mathbf{L}^T\mathbf{U}(\mathbf{z}) \tag{98}$$

where $[\mathbf{L}]_i = \int_{\Gamma_{S_1}} \mathbf{k}_1 \cdot \mathbf{N}_i \mathrm{d}s$ and $\mathbf{U}(\mathbf{z})$ solves and, as before, $\mathbf{U}(\mathbf{z})$ is the solution to $[\mathbf{K}(\mathbf{z}) + \mathbf{K}_s]\,\mathbf{U} = \mathbf{F}$. Here $\mathbf{K}_s$ is the stiffness matrix associated with bilinear form $a_s(\cdot,\cdot)$ and is independent of the design. The gradient of J can be readily computed as $\nabla J(\mathbf{z}) = -\mathbf{P}^T\overline{\mathbf{E}}(\mathbf{z})$ where

$$\left[\overline{\mathbf{E}}(\mathbf{z})\right]_e = p\,[\mathbf{Pz}]_e^{p-1}\,\overline{\mathbf{U}}(\mathbf{z})^T\mathbf{k}_e\mathbf{U}(\mathbf{z}) \tag{99}$$

and $\overline{\mathbf{U}}(\mathbf{z})$ is the solution to the *adjoint* problem

$$[\mathbf{K}(\mathbf{z}) + \mathbf{K}_s]\,\overline{\mathbf{U}} = \mathbf{L}. \tag{100}$$

We refer the reader to [38,19] for more details on the formulation of the compliant mechanism design. It is evident that $\nabla\overline{J}$ can take both positive and negative values. The main implication of this for the proposed algorithm is that the reciprocal approximation of the cost functional is not convex and so we cannot use its Hessian directly in the proximal term of the quadratic model. A simple alternative that we tested is to use (65) with the diagonal entries modified as

$$h_k(\mathbf{y}) = \left|\frac{2}{[\mathbf{y}]_k}\left[\mathbf{P}^T\overline{\mathbf{E}}(\mathbf{y})\right]_k\right|. \tag{101}$$

Such an approximation has been previously explored in [8,9]. Other possibilities based on exponential expansions or MMA-type reciprocal expansions featuring asymptotes can also be considered.

In the numerical studies for the compliant mechanism problem, we consider two cases with different ratio of stiffness of workpiece and actuator, one of which is a proposed benchmark in [5]. Representative results along with the value of parameters used are provided in Fig. 11. The mesh used for this problem consists of 160 × 160 bilinear quadrilateral elements. Moreover, a volume constraint, with $\Theta = 0.25$, is explicitly enforced. Table 4 contains a summary of performance of the two-metric projection algorithm, discussed in the previous subsection, for different initial step sizes, as well as MMA. For both examples, the results using the splitting algorithm are nearly identical and outperform the solution produced by MMA. Another noteworthy fact is that one-node hinges, typically encountered in the solution of compliant mechanism problems, appear to be sufficiently penalized by the Tikhonov regularizer.

8. Concluding remarks

The rather restricted and narrow comparison with the gradient projection algorithm and MMA in the previous section is meant to motivate the virtue of developing tailored algorithms for each problem at hand. In the splitting algorithm proposed here, we use additional knowledge about the behavior of J to construct accurate approximations using only first order information and minimal storage requirements. Furthermore, the two-metric approach allows

Table 4
Summary of the results for the force inverter problem. Here $\tilde{J}(\rho) = \overline{J}(\rho) + R(\rho)$. Convergence was determined here only using the relative decrease in the objective function $\tilde{J}$, as measured by E_1.

Problem	Algorithm	τ_0	# it.	# bt.	$\overline{J}(\rho)$	$R(\rho)$	$V(\rho)$	$\tilde{J}(\rho)$	E_1
$\Vert\mathbf{k}_1\Vert / \Vert\mathbf{k}_2\Vert = 50$	TMP	0.5	315	0	−0.3063	0.0131	0.250	−0.2932	2.159e−6
	TMP	1.0	193	0	−0.3080	0.0133	0.250	−0.2947	9.821e−6
	TMP	1.5	158	0	−0.3090	0.0132	0.250	−0.2958	9.593e−6
	MMA	–	344	–	−0.2959	0.0134	0.249	−0.2826	1.649e−6
$\Vert\mathbf{k}_1\Vert / \Vert\mathbf{k}_2\Vert = 1000$	TMP	0.5	368	0	−1.5180	0.1081	0.250	−1.4099	8.806e−6
	TMP	1.0	283	21	−1.5186	0.1096	0.250	−1.4090	8.789e−6
	TMP	1.5	189	64	−1.5149	0.1072	0.250	−1.4075	8.184e−6
	MMA	–	300	–	−1.4628	0.0999	0.250	−1.3629	6.422e−6

for a computationally efficient treatment of the constraint set. In fact, the proposed approach is aligned with the renewed interest in first-order convex optimization algorithms for solving large-scale inverse problems in signal recovery, statistical estimation, and machine learning [39–42]. We note that, aside from efficiency, robustness is also a major issue for solving topology optimization problems (see, for example, comments in [16] on total variation regularization). Although the high sensitivity to parameters is, to a large extent, intrinsic to the size, nonconvexity and sometimes nonsmoothness of these problems, we emphasize that it should be minimized as much as possible. Developing an appropriately-designed optimization algorithm that fits the structure of the problem can be key to achieving this.

In the extensions of this work, we intend to consider nonsmooth regularizers such as the total variation of density within the present variable metric scheme. This would require the extension of available denoising algorithms (e.g. [43, 41]) for solving the resulting subproblems at each iteration. Also of interest is the use of accelerated first order methods such as those proposed in [44,45] that can improve the convergence speed of the algorithms. Developing a two-metric variation of such algorithms for the constrained minimization problems of topology optimization is promising.

Acknowledgments

The authors acknowledge the support by the Department of Energy Computational Science Graduate Fellowship Program of the Office of Science and National Nuclear Security Administration in the Department of Energy under contract DE-FG02-97ER25308, and the National Science Foundation (NSF) through grants #1234243 and #1321661 (Civil, Mechanical and Manufacturing Innovation Division). The information presented in this paper is the sole opinion of the authors and does not necessarily reflect the views of the sponsoring agencies.

References

[1] G. Allaire, Shape Optimization by the Homogenization Method, Springer, New York, 2001.
[2] O. Sigmund, J. Petersson, Numerical instabilities in topology optimization: a survey on procedures dealing with checkerboards, mesh-dependencies and local minima, Struct. Optim. 16 (1998) 68–75.
[3] T. Bruns, D.A. Tortorelli, Topology optimization of non-linear elastic structures and compliant mechanisms, Comput. Methods Appl. Mech. Engrg. 190 (2001) 3443–3459.
[4] C. Talischi, G.H. Paulino, A. Pereira, I.F.M. Menezes, PolyTop: a Matlab implementation of a general topology optimization framework using unstructured polygonal finite element meshes, Struct. Multidiscip. Optim. 45 (2012) 329–357.
[5] O. Sigmund, Morphology-based black and white filters for topology optimization, Struct. Multidiscip. Optim. 33 (2007) 401–424.
[6] O. Sigmund, K. Maute, Sensitivity filtering from a continuum mechanics perspective, Struct. Multidiscip. Optim. 46 (2012) 471–475.
[7] C. Talischi, G.H. Paulino, An operator splitting algorithm for Tikhonov-regularized topology optimization, Comput. Methods Appl. Mech. Engrg. 253 (2013) 599–608.
[8] A.A. Groenwold, L.F.P. Etman, A quadratic approximation for structural topology optimization, Int. J. Numer. Meth. Eng. 82 (2010) 505–524.
[9] A.A. Groenwold, L.F.P. Etman, D.W. Wood, Approximated approximations for SAO, Struct. Multidiscip. Optim. 41 (2010) 39–56.
[10] D.P. Bertsekas, Projected Newton methods for optimization problems with simple constraints, SIAM J. Control Optim. 20 (1982) 221–246.
[11] E.M. Gafni, D. Bertsekas, Two-metric projection methods for constrained optimization, SIAM J. Control Optim. 20 (1984) 936–964.
[12] M.P. Bendsøe, Optimal design as material distribution probelm, Struct. Optim. 1 (1989) 193–202.
[13] G.I.N. Rozvany, M. Zhou, T. Birker, Generalized shape optimization without homogenization, Struct. Optim. 4 (1992) 250–252.
[14] G.I.N. Rozvany, A critical review of established methods of structural topology optimization, Struct. Multidiscip. Optim. 37 (2009) 217–237.

[15] T. Borrvall, J. Petersson, Topology optimization using regularized intermediate density control, Comput. Methods Appl. Mech. Engrg. 190 (2001) 4911–4928.

[16] T. Borrvall, Topology optimization of elastic continua using restriction, Arch. Comput. Methods Eng. 8 (2001) 251–285.

[17] C. Talischi, Restriction Methods for Shape and Topology Optimization, University of Illinois at Urbana-Champaign Thesis, 2012.

[18] L. Dede, M.J. Borden, T.J.R. Hughes, Isogeometric analysis for topology optimization with a phase field model, Arch. Comput. Methods Eng. 19 (2012) 427–465.

[19] M.P. Bendsøe, O. Sigmund, Topology Optimization: Theory, Methods and Applications, Springer, 2003.

[20] B. Bourdin, A. Chambolle, Design-dependent loads in topology optimization, ESAIM Control Optim. Calc. Var. 9 (2003) 19–48.

[21] M. Burger, R. Stainko, Phase-field relaxation of topology optimization with local stress constraints, SIAM J. Control Optim. 45 (2006) 1447–1466.

[22] A. Takezawa, S. Nishiwaki, M. Kitamura, Shape and topology optimization based on the phase field method and sensitivity analysis, J. Comput. Phys. 229 (2010) 2697–2718.

[23] G. Cohen, Optimization by decomposition and coordination: a unified approach, IEEE Trans. Automat. Control 23 (1978) 222–232.

[24] G.H.G. Chen, R.T. Rockafellar, Convergence rates in forward–backward splitting, SIAM J. Optim. 7 (1997) 421–444.

[25] M. Patriksson, Cost approximation: a unified framework of descent algorithms for nonlinear programs, SIAM J. Optim. 8 (1998) 561–582.

[26] J.S. Arora, Analysis of optimality criteria and gradient projection methods for optimal structural design, Comput. Methods Appl. Mech. Engrg. 23 (1980) 185–213.

[27] A.A. Groenwold, L.F.P. Etman, On the equivalence of optimality criterion and sequential approximate optimization methods in the classical topology layout problem, Int. J. Numer. Meth. Eng. 73 (2008) 297–316.

[28] B. Bourdin, Filters in topology optimization, Int. J. Numer. Meth. Eng. 50 (2001) 2143–2158.

[29] K. Svanberg, The Method of moving asymptotes–a new method for structural optimization, Int. J. Numer. Meth. Eng. 24 (1987) 359–373.

[30] J. Petersson, O. Sigmund, Slope constrained topology optimization, Int. J. Numer. Meth. Engng. 41 (1998) 1417–1434.

[31] J. Petersson, Some convergence results in perimeter-controlled topology optimization, Comput. Methods Appl. Mech. Engrg. 171 (1999) 123–140.

[32] S.C. Brenner, L.R. Scott, The Mathematical Theory of Finite Element Methods, second ed., Springer, 2002.

[33] L.C. Evans, Partial differential equations, in: Graduate Studies in Mathematics, American Mathematical Society, Rhode Island, 1998.

[34] D.P. Bertsekas, Nonlinear Programming, second ed., Athena Scientific, 1999.

[35] A. Beck, M. Teboulle, Gradient-based algorithms with applications to signal recovery problems, in: Convex Optimization in Signal Processing and Communications, Cambridge University Press, 2010.

[36] M. Adlers, Sparse Least Squares Problems with Box Constraints, Department of Mathematics, Linkoping University, Thesis, 1998.

[37] O. Sigmund, A 99 line topology optimization code written in Matlab, Struct. Multidiscip. Optim. 21 (2001) 120–127.

[38] O. Sigmund, On the design of compliant mechanisms using topology optimization, Mech. Based Des. Struct. Mach. 25 (1997) 493–524.

[39] S.J. Wright, Optimization in machine learning, in: Neural Information Processing Systems (NIPS) Workshop, 2008.

[40] P.L. Combettes, V.R. Wajs, Signal recovery by proximal forward–backward splitting, Multiscale Model. Simul. 4 (2006) 1168–1200.

[41] K. Bredies, A forward–backward splitting algorithm for the minimization of non-smooth convex functionals in Banach space, Inverse Problems 25 (2009) p. 015005.

[42] J. Duchi, Y. Singer, Efficient online and batch learning using forward backward splitting, J. Mach. Learn. Res. 10 (2009) 2899–2934.

[43] A. Chambolle, An algorithm for total variation minimization and applications, J. Math. Imaging Vision 20 (2004) 89–97.

[44] Y. Nesterov, Gradient methods for minimizing composite objective function, 2007, available at http://www.ecore.be/DPs/dp1191313936.pdf.

[45] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (2008) 183–202.