SIAM J. NUMER. ANAL.
Vol. 37, No. 1, pp. 246–268

A DISCONTINUOUS FINITE ELEMENT METHOD FOR SOLVING A MULTIWELL PROBLEM*

MATTHIAS K. GOBBERT† AND ANDREAS PROHL‡

Abstract. Many physical materials of practical relevance can attain several variants of crystalline microstructure. The appropriate energy functional is necessarily nonconvex, and the minimization of the functional becomes a challenging problem. A new numerical method based on discontinuous finite elements and a scaled energy functional is proposed. It exhibits excellent convergence behavior for the energy (second order) as well as other crucial quantities of interest for general spatial meshes, contrary to standard (non-)conforming methods. Both theoretical analyses and numerical test calculations are presented and contrasted to other current finite element methods for this problem.

Key words. finite element method, nonconvex minimization, nonlinear conjugate gradients, multiwell problem, microstructure, multiscale, nonlinear elasticity, shape-memory alloys, materials science

AMS subject classifications. 49M07, 65K10, 65N30, 73C50, 73S10

PII. S0036142998333791

1. Introduction.

Many materials of interest in materials science and structural mechanics have been found to possess microscopic structure under certain ambient conditions. Since this microstructure is obtained by stress or temperature induced deformations from a reference state, the mathematical model can be posed as a minimization problem for an energy functional.

From the mathematical point of view, corresponding deformations are described as solutions of a minimization problem of the related energy functional. The existence of a minimizer of this functional cannot be ensured in general, since the energy density is not quasi convex and thus not weakly lower semicontinuous; see [12]. Therefore, minimizing sequences of deformations need to be considered that exhibit increasingly finer scale structures; see [1, 2, 24]. Apart from these analytical difficulties, the numerical modeling of this kind of problem is also a challenging task; see also [4, 17, 18]. For practical purposes, the numerical simulation of material properties is gaining interest in the engineering community, along with the rise of "rational" materials as seen for instance in shape-memory alloys and micromachines. For a more detailed review of the background and the state of the art of the numerical analysis for microstructure computations, we refer to the survey article by Luskin [22].

One area of particular interest is the simulation of austenite-martensite transformations for shape-memory alloys, and we refer to the work of Ball and James [1, 2]. Within this area, twinning is an important phenomenon, in which simple laminates are formed by a deformation gradient that oscillates on an infinitesimal scale in parallel layers between two stress-free states. These stress-free states are

*Received by the editors February 6, 1998; accepted for publication (in revised form) April 2, 1999; published electronically November 23, 1999.

http://www.siam.org/journals/sinum/37-1/33379.html

†Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD (gobbert@math.umbc.edu). This author was partially supported by a grant from the University of Maryland, Baltimore County, and from the National Science Foundation under grant DMS-9805547.

‡Mathematisches Seminar, Christian-Albrechts-Universität Kiel, Ludewig-Meyn-Str. 4, D–24098 Kiel, Germany (apr@numerik.uni-kiel.de).

given by two symmetry-related variants of the martensitic phase. It follows from the frame-indifference principle that the minimum value of the elastic energy is attained on multiple, rotationally invariant wells. In this paper, we study the approximation of deformations of martensitic crystals which can undergo an orthorhombic to monoclinic transformation, giving rise to a double well potential. We claim that the presented analysis can easily be extended to more complicated phase transitions like, e.g., the "triple well case" that describes the cubic to tetragonal phase transitions in atomic lattice structure; we refer to [19, 20] for a corresponding study of the rotated bi- or trilinear finite elements. For the present "double well case," deformations are energetically favored that have deformation gradients in the union of energy wells $\mathcal{U} = \bigcup_{i=1}^{2} \mathcal{U}_i$, which give zero energy contributions to the energy

$$\mathcal{E}(v) = \int_\Omega \phi(\nabla v(x))\, dx \tag{1.1}$$

for admissible deformations v to be defined below. Each well is of the form $\mathcal{U}_i = SO(3)U_i$ with $SO(3)$ being the group of proper rotations and the U_i, $i = 1, 2$ representing martensitic variants. In the double well case, we can assume (see [22]) that these wells are rank-one connected, i.e., there are $F_i \in \mathcal{U}_i$, $i = 1, 2$, such that the Hadamard condition is satisfied. This means that there are two nonvanishing vectors $a \in \mathbb{R}^3$ and $n \in \mathbb{R}^3$ such that

$$F_2 = F_1 + a \otimes n. \tag{1.2}$$

Without loss of generality, we assume $|n| = 1$.

Again in general, $\Omega \subset \mathbb{R}^3$ denotes the reference domain of the crystal, which is assumed to be polygonal. The mapping $v : \Omega \to \mathbb{R}^3$ represents then a continuous deformation of the reference configuration of the crystal with the deformation gradient $\nabla v : \Omega \to \mathbb{R}^{3\times 3}$.

The energy density is assumed to satisfy

$$\begin{aligned} \phi(A) &\geq 0 \qquad \forall\ A \in \mathbb{R}^{3\times 3}, \\ \phi(A) &= 0, \qquad \Longleftrightarrow \quad A \in \mathcal{U}. \end{aligned} \tag{1.3}$$

We shall also assume that the energy density ϕ grows quadratically away from the energy wells; that is,

$$\phi(F) \geq \kappa |||F - \pi(F)|||^2 \qquad \forall\ F \in \mathbb{R}^{3\times 3} \tag{1.4}$$

with $\kappa > 0$ constant and $\pi : \mathbb{R}^{3\times 3} \to \mathcal{U}$ a Borel measurable projection defined by

$$|||F - \pi(F)||| = \min_{G\in\mathcal{U}} |||F - G||| \qquad \forall\ F \in \mathbb{R}^{3\times 3}, \tag{1.5}$$

where $|||\cdot|||$ denotes the Frobenius norm of a matrix, i.e., $|||A||| = \sqrt{\sum_{i,j=1}^{3} A_{ij}^2}$. Note that the projection $\pi(F)$ exists for any $F \in \mathbb{R}^{3\times 3}$, since $\mathcal{U}$ is compact, although it may not be unique.

For the double well case, the laminate microstructure, which we are interested in, is uniquely described by the affine boundary condition [2]

$$v(x) = F_\lambda x \qquad \forall\ x \in \partial\Omega, \tag{1.6}$$

TABLE 1.1
Comparison of convergence results for the energy and other crucial quantities for different finite element methods. See the text for explanation of the notation.

Finite element method	$\mathcal{E}_h(u_h)$	$\|u_h - F_\lambda x\|$	$\|(\nabla u_h - F_\lambda)\, w\|$	$\left\|\frac{\mu(\omega_\rho^i(u_h))}{\mu(\omega)} - \lambda^i\right\|$
Classical conforming, see [21, 22]	$\mathcal{O}(h^{1/2})$	$\mathcal{O}(h^{1/8})$	$\mathcal{O}(h^{1/8})$	$\mathcal{O}(h^{1/16})$
Classical nonconforming, see [20]	$\mathcal{O}(h^{1/2})$	$\mathcal{O}(h^{1/8})$	$\mathcal{O}(h^{1/8})$	$\mathcal{O}(h^{1/16})$
Discontinuous, see Theorem 1.3	$\mathcal{O}(h^2)$	$\mathcal{O}(h^{1/4})$	$\mathcal{O}(h^{1/4})$	$\mathcal{O}(h^{1/8})$

where

$$F_\lambda = \lambda F_1 + (1-\lambda)F_2 \tag{1.7}$$

and $\lambda \in [0,1]$ represents the volume fraction of the two variants. The problem is now stated in the following way:

$$\inf_{v\in\mathcal{A}} \mathcal{E}(v) \tag{1.8}$$

for the set of admissible functions

$$\mathcal{A} = \left\{v \in C(\bar{\Omega};\mathbb{R}^3) : v(x)|_{\partial\Omega} = F_\lambda x\right\}. \tag{1.9}$$

The first finite element methods for problem (1.8) used classical conforming elements with piecewise (bi- or tri-)linear basis functions on each triangular or quadrilateral element, thus minimizing on a subset $\mathcal{A}_h \subset \mathcal{A}$ for conforming elements; see [7]. In the context of convex energy densities ϕ, classical conforming methods are well understood and yield optimal convergence results with order $\mathcal{O}(h^2)$. However, for the present problem of a nonconvex energy density, the results are rather sobering: In general, it can only be shown that a minimizing deformation $u_h \in \mathcal{A}_h$ satisfies

$$\mathcal{E}(u_h) \le Ch^{1/2}, \tag{1.10}$$

where C denotes a generic constant that may depend on the topology of the quasi-uniform triangulation $\mathcal{T}_h$ and the domain Ω but not on the mesh-size h; see [8, 21, 22] and [7] for a definition of quasi uniformity. For a complete list of results for important quantities, see Table 1.1. Moreover, it turns out that the quality of the approximation depends strongly on the degree of alignment of the numerical mesh with the physical laminates. This means that the laminated microstructure is well resolved on meshes, whose element edges run along the laminated direction. If this is not the case, the numerical results are often so polluted that the laminates are distorted beyond recognition; see Figure 3.1 for an example, which will be discussed in detail in section 3.

The limitation of the convergence order for conforming elements has also been studied in [6]. For a different, nonconvex energy density and related deformations $v : \mathbb{R}^2 \supset \Omega \to \mathbb{R}$, [6] shows that suboptimal convergence rates are sharp in general. These observations demonstrate the severe drawbacks of classical conforming finite element methods in the context of highly oscillatory solutions, since the high number of continuity constraints locally limits the flexibility of the numerical method. Another newer result for a conforming method using a reduced integration scheme, which yields the same convergence order, is given in [11].

As a second numerical approach, classical nonconforming finite element methods (using piecewise rotated (bi-, tri-)linear basis functions) are presented in [15, 20, 22]. This method relaxes the continuity constraints between each two elements by only requiring continuity of the discrete deformations at the edge midpoints. Of course, the functional $\mathcal{E}(\cdot)$ is then defined in an appropriate elementwise setting by taking $\mathcal{E}_h(\cdot)$ instead. Computational tests of this scheme have been reported in [15, 16]. At first glance, this finite element method has increased flexibility to handle deformations with microstructure on general grid topologies due to the relaxation of the interelement continuity requirements. However, the theoretical analysis presented in [20] does not reflect this improved flexibility in comparison to the conforming method, and the result is still

$$\mathcal{E}_h(u_h) \leq Ch^{1/2}; \tag{1.11}$$

see Table 1.1.

Based on these sobering observations of classical finite element methods for the problem of twinning, we see an evident need for new finite element methods that yield more accurate approximations of crucial quantities such as the deformation, the structure of laminates, and the statistic properties of the microstructure (i.e., the Young measure) on *general meshes*. Moreover, we believe that the robust resolution of (laminate) microstructure on nonaligned meshes is an essential prerequisite to simulate force-driven deformations as well as phenomena occurring in evolutionary models both in this context and more complicated materials in general.

To this end, we present a new algorithm based on *discontinuous finite elements*. It will be shown that this algorithm allows much improved convergence rate estimates for the energy, namely, $\mathcal{O}(h^2)$ and other quantities of interest as they are given in Table 1.1. In particular, the resolution of laminate microstructure on general meshes is much better than by the classical (non-)conforming discussed above. This statement is justified through drastically improved convergence results and illustrated by numerical experiments on nonaligned meshes.

The underlying conceptual ideas of the new numerical method are the following:

1. The (averaged) boundary conditions will be treated in a more relaxed way to avoid the pollution impact from the boundary.
2. The cross-element continuity constraints are relaxed in a sense that small jumps are allowed.
3. The laminate structures are scaled differently from the transitions between laminates.

In order to explain these ideas, we will start with the proposal of a first numerical model, which incorporates the first two ideas above and which would seem appropriate for our purposes.

ALGORITHM 1.1. *Given a quasi-uniform triangulation $\mathcal{T}_h$ of the domain $\Omega \subset \mathbb{R}^3$, consider elementwise linear deformations $v_h \in \mathcal{A}_h \equiv \prod_{K\in\mathcal{T}_h} \mathcal{P}_1(K)$ with the scaled energy functional*

$$\begin{aligned}\mathcal{E}_h(v_h) = &\sum_{K\in\mathcal{T}_h} \int_K \phi(\nabla v_h(x))\, dx \\ &+ \alpha_{11}\left(\sum_{K\in\mathcal{T}_h} h\int_{\partial K} \left|[v_h](x)\right|\, d\sigma\right)^2 + \alpha_{12}\left(\sum_{K\in\mathcal{T}_h} h\int_{\partial K} \left|[v_h](x)\right|^2\, d\sigma\right)\end{aligned} \tag{1.12}$$

$$+ \alpha_2 \sum_{K\in\mathcal{T}_h} \int_{\partial K\cap\partial\Omega} |v_h(x) - F_\lambda x|^2 \, d\sigma$$

and perform the minimization

$$\min_{v_h\in\mathcal{A}_h} \mathcal{E}_h(v_h). \tag{1.13}$$

Here, $\mathcal{P}_1(K)$ denotes the space of affine polynomials defined on an element K. Since the finite elements are nonconforming here, $\mathcal{A}_h$ is not a subset of $\mathcal{A}$ anymore. The coefficients $\alpha_{11}, \alpha_{12}, \alpha_2$ are order one numbers that control the relative contributions from the interelement continuity constraint and the relaxation of the boundary condition. Here and throughout the paper, $[\cdot]$ denotes the jump (of a function) across element faces, i.e., $[v]\big|_{\mathcal{F}}(x) = v\big|_{K^+}(x) - v\big|_{K^-}(x)$ for two adjacent elements $K^+, K^- \in \mathcal{T}_h$ with face $\mathcal{F} = \partial K^- \cap \partial K^+$.

As will be seen from the subsequent analysis, no improved convergence results can be obtained for Algorithm 1.1 for quantities of interest. The algorithm offers too much freedom for minimizers to adopt spurious solutions that have no physical relevance at all, i.e., that exhibit a physically relevant microstructure.

It will turn out that an algorithm appropriate for representing laminated microstructure on general meshes has to incorporate the third feature already listed above. It should be able to distinguish between contributions to a minimizer from the laminate microstructure and from the transitions between laminates; the latter is where the underlying mesh demands its contributions. Therefore, we introduce a different scaling of laminates (which are of order $\mathcal{O}(h^{1-\beta})$) and of transitions between laminates (of order $\mathcal{O}(h)$) into our numerical model. This leads to the following algorithm.

ALGORITHM 1.2. *Given a quasi-uniform triangulation $\mathcal{T}_h$ of the domain $\Omega \subset \mathbb{R}^3$, consider elementwise linear deformations $v_h \in \mathcal{A}_h \equiv \prod_{K\in\mathcal{T}_h} \mathcal{P}_1(K)$ with the scaled energy functional*

$$\begin{aligned}\mathcal{E}_h^\beta(v_h) = &\sum_{K\in\mathcal{T}_h} \int_K \phi(\nabla v_h(x)) \, dx \\ &+ \alpha_{11} \left(\sum_{K\in\mathcal{T}_h} h^{1-\beta} \int_{\partial K} \big|[v_h](x)\big| \, d\sigma \right)^2 + \alpha_{12} \left(\sum_{K\in\mathcal{T}_h} h^{1-\beta} \int_{\partial K} \big|[v_h](x)\big|^2 \, d\sigma \right) \\ &+ \alpha_2 \sum_{K\in\mathcal{T}_h} h^{2\beta} \int_{\partial K\cap\partial\Omega} |v_h(x) - F_\lambda x|^2 \, d\sigma\end{aligned} \tag{1.14}$$

and perform the minimization

$$\min_{v_h\in\mathcal{A}_h} \mathcal{E}_h^\beta(v_h) \tag{1.15}$$

for a fixed constant $\beta \in (0,1)$.

We will demonstrate the superiority of Algorithm 1.2 over the classical conforming and nonconforming methods through a rigorous convergence analysis for the full three-dimensional case as well as numerical test calculations that have been carried out for a scalar prototype problem. In the latter, the deformations are scalar functions, but the energy density still exhibits the crucial mechanisms inherent to the Ericksen–James

energy density [4, 12]. We refer to section 3 for the definition of and the results for the prototype problem.

Throughout the remainder of the paper, we make use of the following standard notation; see [7]. For any integer $k \geq 0$ and $p \in [1, \infty]$, we define the space

$$W_h^{k,p}(\Omega) \equiv \big\{ v \in L^p(\Omega) : \ v\big|_K \in W^{k,p}(K) \ \forall \ K \in \mathcal{T}_h \big\},$$

and we equip $W_h^{k,p}(\Omega)$ with the standard norms $|\cdot|_{k,p} \equiv \big(\sum_{K\in\mathcal{T}_h} |\cdot|_{k,p,K}^p\big)^{1/p}$ for $1 \leq p < \infty$ and $\max_{K\in\mathcal{T}_h} |\cdot|_{k,\infty,K}$ for $p = \infty$. Correspondingly, $\|\cdot\|_{k,p}$ is defined, using norms instead of seminorms. Subsequently, we will omit the indices in situations where the meaning of the notation is clear from the context.

For the summary of the main result, we fix the following additional notation. The orientation of the simply laminated microstructure is uniquely determined by its normal vector $n \in \mathbb{R}^3$. In the following, we make also use of vectors $w \in \mathbb{R}^3$ along the laminates that satisfy $w \cdot n = 0$. Furthermore, the accuracy of representing the volume fractions $\lambda^1 := \lambda$ and $\lambda^2 := (1 - \lambda)$ will be given in terms of volume fractions that represent the two different variants. To this end, we introduce the following sets for any subset $\omega \subset \Omega$, $\rho > 0$, and $v_h \in \mathcal{A}_h$:

$$\omega_\rho^i(v_h) = \bigcup_{K\in\mathcal{T}_h} \big\{ x \in \omega \cap K : \Pi(\nabla v_h)(x) = F_i \ \text{ and } \ \|\nabla v_h(x) - F_i\| < \rho \big\}$$

for $i \in \{1, 2\}$. Here, we made use of the operator $\Pi : \mathbb{R}^{3\times3} \to \{F_1, F_2\}$ which is related to the operator π in the following way:

$$\pi(F) = \Theta(F)\Pi(F) \qquad \text{with} \qquad \Theta : \mathbb{R}^{3\times3} \to SO(3) \ \ \forall \ F \in \mathbb{R}^{3\times3}.$$

Finally, $\mu(\omega)$ denotes the Lebesgue measure of the region ω. We refer to subsection 2.4 for further details. The verification of the following theorem is the subject of section 2.

THEOREM 1.3. *Consider problem* (1.15) *with* $\beta = 1/2$ *as an approximation of problem* (1.8)–(1.9) *with* $\Omega \subset \mathbb{R}^3$ *a bounded set, and suppose* $u \in \mathcal{A}$ *is the weak limit of a minimizing sequence of* (1.8)–(1.9). *Then problem* (1.15) *has at least one solution* $u_h \in \mathcal{A}_h \equiv \prod_{K\in\mathcal{T}_h} \mathcal{P}_1(K)$, *and* u_h *satisfies the following convergence estimates for all* $\omega \subset \Omega$ *and* $h < \rho < 1$ *and for all* $\rho > 0$, *for positive constants* $\alpha_{11}, \alpha_{12}, \alpha_2 = \mathcal{O}(1)$

(a) $\mathcal{E}_h^{1/2}(u_h) \leq Ch^2$,
(b) $\|u_h - F_\lambda x\|_{L^2(\Omega)} \leq Ch^{1/4}$,
(c) $\|(\nabla u_h - F_\lambda)w\|_{L^2(\Omega)} \leq Ch^{1/4}$,
(d) $\big|\frac{\mu(\omega_\rho^i(u_h))}{\mu(\omega)} - \lambda^i\big| \leq Ch^{1/8}$ *for* $i \in \{1, 2\}$.

The generic constant C *may depend on the parameters of the continuous problem* (1.8) *and the values* $\alpha_{11}, \alpha_{12}, \alpha_2$ *but not on the mesh parameter* h. *In the case* (d), *it additionally depends on the choice of the value of* ρ.

Again, we stress the fact that these convergence results are much better than those derived for the conforming (using (bi-, tri-)linear ansatz functions; see [21]) or classical nonconforming (using piecewise rotated (bi-,tri-)linear ansatz functions; see [20, 21]) finite element methods; see also Table 1.1. This reflects the increased accuracy of the ansatz for nonaligned meshes. The misaligned triangulation does *not* lead to a dramatic pollution of the computed solution anymore. This can be clearly seen in the subsequent theoretical analysis in section 2 as well as in the numerical investigation of our new method in section 3.

Finally, we stress the fact that the analysis presented below is heavily motivated by the work of Luskin [21, 22] and Li and Luskin [19, 20]. More extensive information on the numerical test problem is given in [13], and extensions of this work using concepts of adaptivity are analyzed in [23].

2. Analysis for the discontinuous element. Let us recall that there are three contributions in the scaled energy functional $\mathcal{E}_h^\beta(\cdot)$: the first is the bulk energy term as it is given in the continuous version $\mathcal{E}(\cdot)$. The two subsequent terms are responsible to "ensure" certain continuity constraints. The last term in $\mathcal{E}_h^\beta(\cdot)$ allows slight fluctuations of the boundary data to improve the flexibility of the finite element method to model laminate microstructure. The latter energy terms in $\mathcal{E}_h^\beta(\cdot)$ are introduced to allow flexibility of the finite element method with respect to general meshes, allowing small cross-element jumps of the computed solution.

2.1. Discontinuous finite elements. The Lagrange interpolation operator

$$\mathcal{I}_{\mathcal{T}_h} : \prod_{K\in\mathcal{T}_h} C(K) \to \prod_{K\in\mathcal{T}_h} \mathrm{Aff}(K)$$

with $\mathrm{Aff}(K)$ the set of affine-linear functions on the triangle $K \in \mathcal{T}_h$, is defined in a standard way as a point interpolate. From this, inverse inequalities are valid since they hold on each triangle; compare [7].

LEMMA 2.1. *Let k and ℓ be two integers such that $0 \le k \le \ell \le 2$. The following inverse inequalities are valid for any $K \in \mathcal{T}_h$ and any $v_h \in \mathrm{Aff}(K)$:*

1. $|v_h|_{\ell,K} \le Ch^{k-\ell}|v_h|_{k,K}$,
2. $|v_h|_{\ell,\infty,K} \le Ch^{k-\ell-3/2}|v_h|_{k,K}$.

2.2. Properties of minimizers of the functional $\mathcal{E}_h^\beta(\cdot)$. It is possible to construct a deformation $\tilde{v}_h \in \prod_{K\in\mathcal{T}_h} \mathrm{Aff}(K)$ that satisfies $\mathcal{E}_h^\beta(\tilde{v}_h) \le Ch^2$ for $\beta \in [0,1]$. The construction is accomplished in the proof of the following lemma.

LEMMA 2.2. *Let $\mathcal{T}_h$ be a quasi-uniform triangulation covering $\Omega \subset \mathbb{R}^3$. Then, there exists a minimizer $u_h \in \mathcal{A}_h$ of the functional $\mathcal{E}_h^\beta : v_h \mapsto \mathcal{E}_h^\beta(v_h)$ for $0 \le \beta \le 1$, creating an energy that is bounded by*

$$\mathcal{E}_h^\beta(u_h) \le Ch^2.$$

Proof. We define a deformation $C(\Omega) \ni w(x) : \Omega \to \mathbb{R}^3$ by

$$w(x) = \gamma h^{1-\beta}\tilde{w}\left(\frac{x}{\gamma h^{1-\beta}}\right) \tag{2.1}$$

with

$$\tilde{w}(x) = F_1 x + \Big[\int_0^{x\cdot n} \xi(s)\; ds\Big]a,$$

where $\xi(\tilde{s}) : \mathbb{R} \to \mathbb{R}$ is a characteristic function with period 1 given by

$$\xi(\tilde{s}) = \begin{cases} 1 & \forall\ 0 \le \tilde{s} \le \lambda, \\ 0 & \forall\ \lambda < \tilde{s} < 1 \end{cases} \tag{2.2}$$

and an arbitrary choice of the constant $\gamma = \mathcal{O}(1)$. It is evident that the following inequality holds true:

$$|w(x) - F_\lambda x| \le Ch^{1-\beta} \qquad \forall\ x \in \Omega. \tag{2.3}$$

We now have

$$\nabla w(x) = F_1 + \xi \left(\frac{x \cdot n}{\gamma h^{1-\beta}} \right) a \otimes n. \tag{2.4}$$

We are given a triangulation $\mathcal{T}_h = \{K_i\}_{i \in I}$ such that in general $w \neq \mathcal{I}_{\mathcal{T}_h}(w)$. Because of w being piecewise affine it is clear that there exists a *refinement* $\tilde{\mathcal{T}}_h = \{\tilde{K}_{ij}\}_{i \in I, j \in J_i}$ of $\mathcal{T}_h$ with

$$\mathcal{T}_h \ni K_i = \bigcup_{j \in J_i} \tilde{K}_{ij}, \qquad \tilde{K}_{ij} \in \tilde{\mathcal{T}}_h,$$

such that the following holds: $w = \mathcal{I}_{\tilde{\mathcal{T}}_h}(w)$. In our notation, $\mathrm{card} J_i = 1$ stands for no refinement, whereas $\mathrm{card} J_i > 1$ denotes a refinement of $K_i \in \mathcal{T}_h$.

Using the triangulation $\mathcal{T}_h = \{K_i\}_{i \in I}$, we will now construct a deformation $v_h \in \mathcal{A}_h$ from the function w defined in (2.1) by

$$v_h(x) := \begin{cases} w(x) & \forall K_i \in \mathcal{T}_h \text{ with card } J_i = 1, \\ \mathrm{Ext}_{K_i}(w)(x) & \forall K_i \in \mathcal{T}_h \text{ with card } J_i > 1. \end{cases} \tag{2.5}$$

For our purposes, we define the extension operator

$$\mathrm{Ext}_{K_i} : \prod_{j \in J_i} \mathrm{Aff}(\tilde{K}_{ij}) \to \mathrm{Aff}(K_i)$$

with $\mathrm{Ext}_{K_i}(w)(x)$ a linear extension of $w\big|_{\tilde{K}_{ij_0}}$ with $j_0 \in J_i$, onto $\bigcup_{j=J_i} \tilde{K}_{ij} = K_i \in \mathcal{T}_h$, satisfying

$$\nabla \mathrm{Ext}_{K_i}(w)(x) = \nabla w(x)\big|_{\tilde{K}_{ij_0}} \qquad \text{with } \mu(\tilde{K}_{ij_0}) \geq \mu(\tilde{K}_{ij}) \ \ \forall\, j \in J_i;$$

in case of ambiguity due to equally sized elements $\tilde{K}_{ij_0}$ in this definition, choose the smallest j_0. This gives a deformation candidate $v_h = \mathcal{I}_{\mathcal{T}_h}(v_h)$ that satisfies the given energy bound. Since the jumps of order $\mathcal{O}(h)$ in v_h are along $\mathcal{O}(h^{\beta-1})$ lines that are touched by $\mathcal{O}(h^{-2})$ elements, the second term in (1.14) can be bounded as

$$\left(\sum_{K \in \mathcal{T}_h} h^{1-\beta} \int_{\partial K} \big|[v_h](x)\big| \, d\sigma \right)^2 \leq C \left(h^{\beta-1} h^{-2} h^{1-\beta} h^2 h \right)^2 \leq C h^2.$$

The third term can be controlled in an analogous fashion. Deviations from the prescribed boundary data are penalized by the fourth term, which amounts to $\mathcal{O}(h^2)$ for the deformation v_h, using (2.3),

$$\sum_{K \in \mathcal{T}_h} h^{2\beta} \int_{\partial K \cap \partial \Omega} |v_h(x) - F_\lambda x|^2 \, d\sigma \leq C h^{-2} h^{2\beta} h^2 h^{2(1-\beta)} \leq C h^2.$$

The existence of a minimizer $u_h \in \mathcal{A}_h$ now follows from compactness arguments for the continuous functional $\mathcal{E}_h^\beta(\cdot)$ enjoying property (1.4), since the underlying set of admissible functions $\mathcal{A}_h$ is finite-dimensional. We refer to [14] for a corresponding elaboration of this argument for nonconforming finite elements. ∎

Based on this result, we can prove the following theorem that is crucial for the further analysis of Algorithm 1.2. In particular, this explains why Algorithm 1.1

(corresponding to the case $\beta = 0$) fails and the additional concept of scaling is essential in Algorithm 1.2.

THEOREM 2.3. *A minimizer $u_h \in \mathcal{A}_h$ of the energy functional $\mathcal{E}_h^\beta(\cdot)$ given in (1.14) satisfies the following estimate:*

$$\Big|\Big|\Big| \sum_{K\in\mathcal{T}_h} \int_K \{\nabla u_h(x) - F_\lambda\}\, dx \Big|\Big|\Big| \le C\Big\{h^\beta + h^{1-\beta}\Big\} \qquad \textit{for } \beta \in [0,1].$$

Proof. Set $z_h(x) = u_h(x) - F_\lambda x$. Then we have

$$\begin{aligned}
\sum_{K\in\mathcal{T}_h} \int_K \nabla z_h(x)\, dx &= \sum_{K\in\mathcal{T}_h} \int_{\partial K} z_h\Big|_K(x) \otimes \nu\, d\sigma \\
&= \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\not\subset\partial\Omega} \int_{\mathcal{F}} z_h\Big|_K(x) \otimes \nu\, d\sigma + \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\subset\partial\Omega} \int_{\mathcal{F}} z_h\Big|_K(x) \otimes \nu\, d\sigma \\
&=: I_1 + I_2.
\end{aligned} \tag{2.6}$$

Since two neighboring elements K^+ and K^- such that $\overline{K^+} \cap \overline{K^-} \neq \emptyset$ share one face with their related normal vectors changing their sign (i.e., $\nu|_{K^-} = -\nu|_{K^+}$), we can continue with the first term as follows, setting $z_h^\pm = z_h\big|_{K^\pm}$,

$$\begin{aligned}
|||I_1||| &= \frac{1}{2}\Big|\Big|\Big| \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\not\subset\partial\Omega} \int_{\mathcal{F}} (z_h^+ - z_h^-)(x) \otimes \nu\big|_{K^+}\, d\sigma \Big|\Big|\Big| \\
&\le \frac{1}{2} \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\not\subset\partial\Omega} \int_{\mathcal{F}} \big|[u_h](x)\big|\, d\sigma \le Ch^\beta.
\end{aligned} \tag{2.7}$$

The last bound is a consequence of Lemma 2.2. Another application of it further leads to an upper bound for the term I_2:

$$\begin{aligned}
|||I_2||| &\le \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\subset\partial\Omega} \int_{\mathcal{F}} |z_h(x)|\, d\sigma \le C \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\subset\partial\Omega} h\Big(\int_{\mathcal{F}} |z_h(x)|^2\, d\sigma\Big)^{1/2} \\
&\le C \left(\sum_{\mathcal{F}\subset\partial K, \mathcal{F}\subset\partial\Omega} \int_{\mathcal{F}} |z_h(x)|^2\, d\sigma \right)^{1/2} \le Ch^{1-\beta}.
\end{aligned} \tag{2.8}$$

This concludes the proof. ∎

Remark 2.4. We stress the fact that in the case of the classical conforming or classical nonconforming elements—as described above—we have the even sharper result

$$\sum_{K\in\mathcal{T}_h} \int_K \{\nabla u_h(x) - F_\lambda\}\, dx = 0.$$

This ensures that the gradient of the computed deformation u_h is identical to the macroscopically observable F_λ in an averaged sense. In fact, this "stability requirement" is not necessary to obtain an efficient scheme, as we will see in the following.

We can now continue our analysis with the verification of the following theorem.

THEOREM 2.5. *A minimizer $u_h \in \mathcal{A}_h$ of the functional $\mathcal{E}_h^\beta(\cdot)$ that is given in* (1.14) *satisfies the following estimate for all normalized vectors $w \in \mathbb{R}^3$:*

$$\left(\sum_{K\in\mathcal{T}_h} \int_K |u_h(x) - F_\lambda x|^2 \, dx \right)^{1/2} \leq C \left\{ \left(\sum_{K\in\mathcal{T}_h} \int_K \left| \{\nabla u_h(x) - F_\lambda\} w \right|^2 \, dx \right)^{1/2} + h^{1-\beta} + h^{\beta/2} \right\}.$$

Proof. We will again use the abbreviative notation $z_h(x) = u_h(x) - F_\lambda x$ for $x \in \Omega$. Using integration by parts, we obtain

$$\begin{aligned} \int_\Omega |z_h(x)|^2 \, dx &= \sum_{K\in\mathcal{T}_h} \int_{\partial K} |z_h(x)|^2 (w\cdot x)(w\cdot\nu) \, d\sigma \\ &\quad - \sum_{K\in\mathcal{T}_h} \int_K (\nabla |z_h(x)|^2 \cdot w)(w\cdot x) \, dx =: I_1 + I_2 \end{aligned} \tag{2.9}$$

with an arbitrary vector $w \in \mathbb{R}^3$, $|w| = 1$.

The second term can be controlled as follows:

$$\begin{aligned} |I_2| &= \Big| \sum_{K\in\mathcal{T}_h} \int_K (\nabla |z_h(x)|^2 \cdot w)(w\cdot x) \, dx \Big| \\ &\leq C \max_{x\in\overline{\Omega}} |w\cdot x| \left(\sum_{K\in\mathcal{T}_h} \int_K |\nabla z_h(x) w|^2 \, dx \right)^{1/2} \left(\int_\Omega |z_h(x)|^2 \, dx \right)^{1/2} \\ &\leq \frac{1}{4} \int_\Omega |z_h(x)|^2 \, dx + C \sum_{K\in\mathcal{T}_h} \int_K |\nabla z_h(x) w|^2 \, dx. \end{aligned} \tag{2.10}$$

The constant $1/4$ in front of the first term is obtained by Young's inequality applied to the previous formula; the generic constant C in front of the second term also absorbs $\max_{x\in\overline{\Omega}} |w\cdot x|$ as well as the (square of the) generic constant from the previous step. In order to handle the term I_1, we distinguish between the edges in the interior of the domain and those on the boundary $\partial\Omega$:

$$\begin{aligned} I_1 &= \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\subset\partial\Omega} \int_{\mathcal{F}} |z_h(x)|^2 (w\cdot x)(w\cdot\nu) \, d\sigma \\ &\quad + \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\not\subset\partial\Omega} \int_{\mathcal{F}} |z_h(x)|^2 (w\cdot x)(w\cdot\nu) \, d\sigma =: I_{11} + I_{12}. \end{aligned} \tag{2.11}$$

Because of Lemma 2.2, we can bound the first term by

$$|I_{11}| \leq C \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\subset\partial\Omega} \int_{\mathcal{F}} |z_h(x)|^2 \, d\sigma \leq C h^{2(1-\beta)}. \tag{2.12}$$

In order to bound I_{12}, reorder the summation by element faces and let $z_h^+(x)$ and $z_h^-(x)$ denote the value of $z_h(x)$ along the face $\mathcal{F} = \overline{K^+} \cap \overline{K^-}$ with normal vectors ν^+

and ν^- corresponding to the elements K^+ and K^-, respectively. Then $\nu^+ = -\nu^-$, and it holds $|z_h^+(x) - z_h^-(x)| = |[u_h](x)|$. In the following, the factor $1/2$ reflects the fact that each element face is counted twice after reordering; this factor is later absorbed into the generic constant C:

$$\begin{aligned}
I_{12} &= \frac{1}{2} \sum_{\mathcal{F}\subset\partial K,\mathcal{F}\not\subset\partial\Omega} \int_{\mathcal{F}} \big\{ |z_h^+(x)|^2 (w\cdot x)(w\cdot\nu^+) + |z_h^-(x)|^2 (w\cdot x)(w\cdot\nu^-) \big\}\, d\sigma \\
&= \frac{1}{2} \sum_{\mathcal{F}\subset\partial K,\mathcal{F}\not\subset\partial\Omega} \int_{\mathcal{F}} \big\{ |z_h^+(x)|^2 - |z_h^-(x)|^2 \big\} (w\cdot x)(w\cdot\nu^+)\, d\sigma \\
&= \frac{1}{2} \sum_{\mathcal{F}\subset\partial K,\mathcal{F}\not\subset\partial\Omega} \int_{\mathcal{F}} \big(\{ z_h^+ - z_h^- \}(x) \cdot \{ z_h^+ + z_h^- \}(x) \big) (w\cdot x)(w\cdot\nu^+)\, d\sigma \\
&= \frac{1}{2} \sum_{\mathcal{F}\subset\partial K,\mathcal{F}\not\subset\partial\Omega} \int_{\mathcal{F}} \big(\{ u_h^+ - u_h^- \}(x) \cdot \{ z_h^+ + z_h^- \}(x) \big) (w\cdot x)(w\cdot\nu^+)\, d\sigma \\
&\le C \sum_{\mathcal{F}\subset\partial K,\mathcal{F}\not\subset\partial\Omega} \int_{\mathcal{F}} \big|[u_h](x)\big| \big\{ |z_h^-(x)| + |z_h^+(x)| \big\}\, d\sigma \\
&\le C \sum_{\mathcal{F}\subset\partial K,\mathcal{F}\not\subset\partial\Omega} \Big(\int_{\mathcal{F}} \big|[u_h](x)\big|^2 d\sigma \Big)^{1/2} \Big(\int_{\mathcal{F}} \big\{ |z_h^-(x)|^2 + |z_h^+(x)|^2 \big\}\, d\sigma \Big)^{1/2}.
\end{aligned}$$

Now, apply Lemma 2.1 to obtain

$$\int_{\mathcal{F}} |z_h^-(x)|^2\, d\sigma \le \mu(\mathcal{F}) |z_h^-(x)|_{0,\infty,K^-}^2 \le C h^2 h^{-3} |z_h^-(x)|_{0,K^-}^2 = C h^{-1} \int_{K^-} |z_h^-(x)|^2\, dx$$

and analogously for $z_h^+(x)$. Then we can continue from above:

$$I_{12} \le C \sum_{\mathcal{F}\subset\partial K,\mathcal{F}\not\subset\partial\Omega} \Big(\int_{\mathcal{F}} \big|[u_h](x)\big|^2\, d\sigma \Big)^{1/2} \Big(h^{-1} \int_{K^-} |z_h^-(x)|^2\, dx + h^{-1} \int_{K^+} |z_h^+(x)|^2\, dx \Big)^{1/2} \tag{2.13}$$

and by Young's inequality

$$\begin{aligned}
I_{12} &\le \sum_{\mathcal{F}\subset\partial K,\mathcal{F}\not\subset\partial\Omega} \left\{ \frac{C}{h} \int_{\mathcal{F}} \big|[u_h](x)\big|^2\, d\sigma + \frac{1}{8} \int_{K^-} |z_h^-(x)|^2\, dx + \frac{1}{8} \int_{K^+} |z_h^+(x)|^2\, dx \right\} \\
&\le \frac{C}{h} \sum_{\mathcal{F}\subset\partial K,\mathcal{F}\not\subset\partial\Omega} \int_{\mathcal{F}} \big|[u_h](x)\big|^2\, d\sigma + \frac{1}{4} \sum_{K\in\mathcal{T}_h} \int_K |z_h(x)|^2\, dx,
\end{aligned}$$

where the last two sums were combined into one by reordering and C denotes a generic constant again. Because of the result

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} \big|[u_h](x)\big|^2\, d\sigma \le C h^{1+\beta}, \tag{2.14}$$

which is an immediate consequence of Lemma 2.2, we can now insert (2.10) through (2.14) in (2.9), and the proof is finished. ∎

The following theorem is a local trace inequality, bounding errors on interior edges in terms of the error on elements. For this purpose, we will introduce subdomains $\omega_h \subset \bigcup_{i\in L} K_i$, referred to as "pseudoparallelepipeds." By this, we mean a

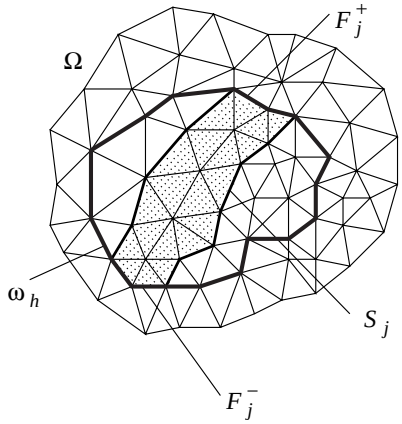


FIG. 2.1. *Sketch for the construction used in the proof of Theorem* 2.6.

perturbed parallelepiped assembled from elements $K_i \in \mathcal{T}_h$ with faces being piecewise affine curves to be considered as $\mathcal{O}(h)$ perturbations of planes that constitute the parallelepiped $\omega \supset \omega_h$. See Figure 2.1 for a sketch of the construction in the two-dimensional case. Owing to the quasi-uniform character of the triangulation $\mathcal{T}_h$, it is always possible to construct at least one pseudoparallelepiped ω_h for a given parallelepiped ω such that $\mathrm{dist}(\partial\omega_h, \partial\omega) < \mathcal{O}(h)$. Moreover, we will employ "pseudotubes" $\mathcal{S}_j$ of thickness $\mathcal{O}(h)$; by this, we mean collections of triangles with faces $\mathcal{F}_j^{\pm} \subset \partial\omega_h$ such that $\omega = \bigcup_{j\in M} \mathcal{S}_j$ with $M \subset L$. A selection criterion for a triangle $K_i \subset \mathcal{T}_h$ to belong to a fixed pseudotube $\mathcal{S}_j$ can for instance be formulated through the introduction of a new "Cartesian reference triangulation" $\mathcal{K}_h$ of the parallelepiped $\omega = \mathrm{span}\{o_1, o_2, o_3\} \subset \Omega$ with $o_i \in \mathbb{R}^3$ and $|o_3| = \Lambda(\omega)$, where the construction process of tubes $\hat{\mathcal{S}}_j = \mathrm{span}\{ho_1, ho_2, o_3\}$ and $\bigcup_{j\in L} \hat{\mathcal{S}}_j = \omega$ is straightforward. We can now make use of these tubes in a way that triangles $K_i \in \mathcal{T}_h$ with $K_i \subset \omega_h$ belong to $\mathcal{S}_j$, provided

$$\mu(K_i \cap \hat{\mathcal{S}}_j) \geq \frac{1}{2}\mu(K_i) \qquad \forall\ i \in I \ \ \forall\ j \in L,$$

and the faces $\mathcal{F} \subset \partial K_i$ for $\partial K_i \cap \partial\omega_h \neq \emptyset$ are faces of the tube $\mathcal{S}_j$, provided the Euclidean orthogonal projection of $\hat{\mathcal{S}}_j$ along $\pm o_3$ onto $\partial\omega_h$, denoted by $\mathcal{B}_j$, satisfies

$$\mu(\mathcal{F} \cap \mathcal{B}_j) \geq \frac{1}{2}\mu(\mathcal{F}) \qquad \forall\ \mathcal{F} \subset \partial K_i \ \text{ such that } \ \mathcal{F} \subset \partial\omega_h$$

with $\mu(\mathcal{F})$ denoting the (two-dimensional) area of $\mathcal{F}$. Of course, this selection process can be "ambiguous" but this does not cause severe problems, thanks to the quasi uniformity of the triangulation $\mathcal{T}_h$. In the following, we refer to distinct sets $\mathcal{F}_j^- := \{\mathcal{F}_{jl}^-\}_{l=1}^r$ and $\mathcal{F}_j^+ := \{\mathcal{F}_{jl}^+\}_{l=1}^q$ with numbers $r = r(j)$, $q = q(j)$ that constitute the bottom and top of the tube $\hat{\mathcal{S}}_j$.

THEOREM 2.6. *Suppose $u_h \in \mathcal{A}_h$ to be a minimizer of $\mathcal{E}_h^\beta(\cdot)$ as it is given in* (1.14). *Then, there exists a constant $C = C(\omega) > 0$ such that for any parallelepiped ω and any associated pseudoparallelepiped $\omega_h = \{K_i\}_{i\in L} \subset \overline{\Omega}$ with $L \subset I$, the following*

is valid:

$$\left(\sum_{\mathcal{F}\subset\partial K,\mathcal{F}\cap\partial\omega_h\neq\emptyset,K\subset\omega_h}\int_{\mathcal{F}}|u_h^\bullet(x)-F_\lambda x|^2\,d\sigma\right)^{1/2}$$
$$\leq\frac{C}{\Lambda(\omega)}\left(\int_{\omega_h}|u_h(x)-F_\lambda x|^2\,dx\right)^{1/2}+Ch^{\beta/2}$$
$$+C\left(\int_{\omega_h}|u_h(x)-F_\lambda x|^2\,dx\right)^{1/4}\left(\sum_{K\in\omega_h}\int_K|||\nabla u_h(x)-F_\lambda|||^2\,dx\right)^{1/4},$$

where $\Lambda(\omega)$ *is the length of the shortest edge of a corresponding parallelepiped* ω *and where* $u_h^\bullet\big|_{\partial K\cap\partial\omega_h\neq\emptyset}$ *is the trace of* $u_h\big|_{K\subset\omega_h}$ *on* $\partial K\cap\partial\omega_h$.

Proof. Using the terminology introduced above, we shall make use of the following construction:

1. Given $\omega\subset\Omega$, determine a pseudoparallelepiped $\omega_h=\bigcup_{i\in L}K_i$.
2. Construct pseudotubes $\mathcal{S}_j$ of thickness $\mathcal{O}(h)$ as collections of triangles with faces $\mathcal{F}_j^\pm\subset\partial\omega_h$ and the property $\bigcup_{j\in M}\mathcal{S}_j=\omega_h$ with $M\subset L$.

See Figure 2.1 for a sketch of the two-dimensional case.

Let us fix one $j\in L$. We are given a direction from the collection of elements $\mathcal{F}_j^-$ to $\mathcal{F}_j^+$ in a canonical way. This allows for a construction to distinguish between z_h^+ and z_h^- on an edge ∂K_i with $K_i\subset\mathcal{S}_j$. Further, let us define a function $\alpha_{jl}^{(i)}:\mathcal{F}_{jl}^{(i)}\mapsto\operatorname{dist}(c_{\mathcal{F}_{jl}^{(i)}},\mathcal{F}_j^-)$ for $1\leq l\leq\operatorname{card}\mathcal{F}_j^{(i)}$, and $\mathcal{F}_j^{(i)}:=\{\mathcal{F}_{jl}^{(i)}\}_{l=1}^{\operatorname{card}\mathcal{F}_j^{(i)}}$, on the edges of the elements of $\mathcal{S}_j$ that represents the distance from the center of $\mathcal{F}_{jl}^{(i)}$, denoted by $c_{\mathcal{F}_{jl}^{(i)}}$, from $\mathcal{F}_j^-$.

For the following, we consider chains $\{\alpha_j^{(i)}\}_{i=0}^k$ that satisfy the following condition with $k=k(j)$ a number that describes the "end" of the chain at $\mathcal{F}_j^+$:

$$\operatorname{dist}(\mathcal{F}_j^{(i)},\mathcal{F}_j^{(i+1)})\leq Ch\qquad\forall\,0\leq i\leq k-1\ \ \forall\,j\in L.$$

Further, each face $\mathcal{F}_{jl}^-=\mathcal{F}_{jl}^{(0)}$ of the collection of bottom faces $\{\mathcal{F}_{jl}^-\}_{l=1}^r$ can be written as the convex hull of three distinct nodal points, $\mathcal{F}_{jl}^{(0)}=\operatorname{conv}(a_{jl1},a_{jl2},a_{jl3})$. Therefore, each $x_l^{(0)}\in\mathcal{F}_{jl}^{(0)}$ can be uniquely represented by a triple $\{\lambda_{l1},\lambda_{l2},\lambda_{l3}\}$ such that $x_l^{(0)}=\sum_{m=1}^3\lambda_{lm}a_{jlm}^{(0)}$ with $\lambda_{lm}\geq0$ and $\sum_{lm=1}^3\lambda_{lm}=1$.

For the following construction, we choose corresponding points in the k levels $\{\mathcal{F}_j^{(i)}\}_{i=1}^k$, defined by

$$\mathcal{F}_{jl}^{(i)}\ni x_l^{(i)}=\sum_{m=1}^3\lambda_{lm}a_{jlm}^{(i)}$$

with $\operatorname{conv}\big(\{a_{jlm}^{(i)}\}_{m=1}^3\big)=\mathcal{F}_{jl}^{(i)}$. Therefore, for each $x_l^{(0)}\in\mathcal{F}_{jl}^{(0)}$, we can define chains $\{x_l^{(i)}\}_{i=0}^k$ that connect $x_l^{(0)}\in\mathcal{F}_{jl}^-$, for $1\leq l\leq r$ with $x_l^{(k)}\in\mathcal{F}_{jl}^+$, $1\leq l\leq q$.

Subsequently, we employ the shorthand notation $z_h(x)=u_h(x)-F_\lambda x$. We are

now prepared to carry out the following consideration $\forall\ x_l^{(0)} \in \mathcal{F}_{jl}^{(0)}$,

$$
\begin{aligned}
&-\operatorname{dist}(c_{\mathcal{F}_{jl}^+}, c_{\mathcal{F}_{jl}^-})\big|z_h^\bullet\big|_{\mathcal{F}_{jl}^{(0)}}(x_l^{(0)})\big|^2 \\
&\qquad = \sum_{i=0}^{k-1}\Big[(\alpha_{jl}^{(k)} - \alpha_{jl}^{(i+1)})\big|z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)})\big|^2 - (\alpha_{jl}^{(k)} - \alpha_{jl}^{(i)})\big|z_h^-\big|_{\mathcal{F}_{jl}^{(i)}}(x_l^{(i)})\big|^2\Big].
\end{aligned}
\tag{2.15}
$$

We can make the following reformulations for each term of the sum:

$$
\begin{aligned}
&\Big|(\alpha_{jl}^{(k)} - \alpha_{jl}^{(i+1)})\big|z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)})\big|^2 - (\alpha_{jl}^{(k)} - \alpha_{jl}^{(i)})\big|z_h^-\big|_{\mathcal{F}_{jl}^{(i)}}(x_l^{(i)})\big|^2\Big| \\
&\quad = \Big|(\alpha_{jl}^{(k)} - \alpha_{jl}^{(i)})\big[|z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)})|^2 - |z_h^-\big|_{\mathcal{F}_{jl}^{(i)}}(x_l^{(i)})|^2\big] \\
&\quad +(\alpha_{jl}^{(i)} - \alpha_{jl}^{(i+1)})|z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)})|^2\Big| \\
&\quad \le C\Lambda(\omega)\Big|\big[z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)}) - z_h^-\big|_{\mathcal{F}_{jl}^{(i)}}(x_l^{(i)})\big] \\
&\quad \times\big[z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)}) + z_h^-\big|_{\mathcal{F}_{jl}^{(i)}}(x_l^{(i)})\big]\Big| + Ch\big|z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)})\big|^2 \\
&\quad \le C\Lambda(\omega)\Big|\big[z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)}) - z_h^+\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)}) \\
&\quad +z_h^+\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)}) - z_h^-\big|_{\mathcal{F}_{jl}^{(i)}}(x_l^{(i)})\big]\cdot\big[z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)}) + z_h^-\big|_{\mathcal{F}_{jl}^{(i)}}(x_l^{(i)})\big]\Big| \\
&\quad +Ch\big|z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)})\big|^2.
\end{aligned}
\tag{2.16}
$$

Because of

$$
z_h^+\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)}) - z_h^-\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)}) = \big[u_h\big]\big|_{\mathcal{F}_{jl}^{(i+1)}}(x_l^{(i+1)}),
\tag{2.17}
$$

we can therefore write, thanks to the quasi uniformity of the triangulation $\mathcal{T}_h$,

$$
\begin{aligned}
\sum_{\mathcal{F}\subset\partial K,\partial K\cap\partial\omega_h\neq\emptyset,K\subset\omega_h}\int_{\mathcal{F}}|z_h^\bullet(x)|^2\ d\sigma &\le \sum_{j\in L}\sum_{K_i\subset\mathcal{S}_j}\Big(h^3\|\nabla z_h\|_{0,\infty,K_i}\|z_h\|_{0,\infty,K_i} \\
&+\int_{\partial K_i}\big|[u_h](x)\big|\big\{|z_h^-(x)| + |z_h^+(x)|\big\}\ d\sigma + \frac{h^3}{\Lambda(\omega)}\|z_h\|_{0,\infty,K_i}^2\Big).
\end{aligned}
\tag{2.18}
$$

We can benefit from the inverse inequality in Lemma 2.1. Further, the second term in the sum can be treated in a way analogous to (2.13), (2.14), and we finally obtain

$$
\le C\left\{\sum_{K\subset\omega_h}\Big(\|\nabla z_h\|_{0,K}\|z_h\|_{0,K} + \frac{1}{\Lambda(\omega)}\|z_h\|_{0,K}^2\Big) + h^\beta + \|z_h\|_{0,\omega_h}^2\right\}.
$$

This concludes the proof. ∎

2.3. Approximation of limiting macroscopic deformations. We start with an approximation result for discrete deformations, given in terms of the corresponding energy.

LEMMA 2.7. *The following inequality is valid:*

$$
\sum_{K\in\mathcal{T}_h}\int_K |||\nabla v_h(x) - \pi(\nabla v_h(x))|||^2\ dx \le C\mathcal{E}_h^\beta(v_h) \qquad \forall\ v_h \in \mathcal{A}_h.
$$

The justification of this lemma can immediately be given, exploiting the quadratic growth of the bulk energy density close to the union of wells $\mathcal{U}$; see (1.4).

Another lemma will be useful for the subsequent studies.

LEMMA 2.8. *For any $w \in \mathbb{R}^3$ satisfying $w \cdot n = 0$, there exists a constant $C > 0$ such that*

$$\left(\sum_{K \in \mathcal{T}_h} \int_K \left| \{\pi(\nabla u_h(x)) - F_\lambda\} w \right|^2 \, dx \right)^{1/2} \leq C \Big\{ h^{1/2} + h^{\beta/2} + h^{(1-\beta)/2} \Big\}$$

with $u_h \in \mathcal{A}_h$ being a minimizer of problem (1.15).

Proof. For the orthorhombic to monoclinic transformation, we have

$$\pi(F) \in SO(3)F_1 \cup SO(3)F_2 \qquad \forall\ F \in \mathbb{R}^{3\times 3}.$$

Because of the rank-one connection and the identity

$$F_\lambda = \lambda F_1 + (1-\lambda)F_2 = F_1 + (1-\lambda) a \otimes n = F_2 - \lambda a \otimes n,$$

this implies

$$|\pi(F)w| = |F_1 w| = |F_2 w| = |F_\lambda w| \qquad \forall\ F \in \mathbb{R}^{3\times 3}$$

for an arbitrary $w \in \mathbb{R}^3$ such that $w \cdot n = 0$. Thanks to this, we find

$$\begin{aligned}
&\sum_{K \in \mathcal{T}_h} \int_K \left| \{\pi(\nabla u_h(x)) - F_\lambda\} w \right|^2 \, dx \\
&\quad = \sum_{K \in \mathcal{T}_h} \int_K \big\{ |\pi(\nabla u_h(x))w|^2 + |F_\lambda w|^2 - 2\pi(\nabla u_h(x))w \cdot F_\lambda w \big\} \, dx \\
&\quad = 2F_\lambda w \cdot \sum_{K \in \mathcal{T}_h} \int_K \big\{ F_\lambda - \pi(\nabla u_h(x)) \big\} w \, dx \\
&\quad = 2F_\lambda w \cdot \sum_{K \in \mathcal{T}_h} \int_K \big\{ \nabla u_h(x) - \nabla u_h(x) + F_\lambda - \pi(\nabla u_h(x)) \big\} w \, dx \\
&\quad = 2F_\lambda w \cdot \sum_{K \in \mathcal{T}_h} \left\{ \int_K \big\{ \nabla u_h(x) - \pi(\nabla u_h(x)) \big\} w \, dx + \int_K \big\{ F_\lambda - \nabla u_h(x) \big\} w \, dx \right\} \\
&\quad \leq C \left(\sum_{K \in \mathcal{T}_h} \int_K \| \nabla u_h(x) - \pi(\nabla u_h(x)) \|^2 \, dx \right)^{1/2} \\
&\quad + 2F_\lambda w \cdot \left(\sum_{K \in \mathcal{T}_h} \int_K \big\{ F_\lambda - \nabla u_h(x) \big\} \, dx \right) w.
\end{aligned} \tag{2.19}$$

We can now make use of Lemma 2.7 and Theorem 2.3 to complete the proof. ∎

The next result is a consequence of the last two results.

THEOREM 2.9. *For any $w \in \mathbb{R}^3$ satisfying $w \cdot n = 0$, there exists a constant $C > 0$ such that*

$$\left(\sum_{K \in \mathcal{T}_h} \int_K \left| \{\nabla u_h(x) - F_\lambda\} w \right|^2 \, dx \right)^{1/2} \leq C \Big\{ h^2 + h^{1/2} + h^{\beta/2} + h^{(1-\beta)/2} \Big\}.$$

The subsequent result is preliminary for proving a result pertaining to the capability of the present finite element method to approximate laminate microstructures.

THEOREM 2.10. *Given a parallelepiped ω and an associated pseudoparallelepiped $\omega_h \subset \bar{\Omega}$, there exists a constant $C = C(\omega) > 0$ such that the following result holds for a minimizer $u_h \in \mathcal{A}_h$:*

$$\Big|\Big|\Big| \sum_{K\cap\omega\neq\emptyset, K\in\mathcal{T}_h} \int_K \{\nabla u_h(x) - F_\lambda\}\, dx \Big|\Big|\Big| \leq C\Big\{h + h^{3/2} + h^\beta + h^{1/4} + h^{(1-\beta)/4} + h^{\beta/4}\Big\}.$$

Proof. In the following, we will employ the notation of ω and ω_h. Then we can split the following integral into contributions from ω_h and from $\omega - \omega_h$; thus,

$$\begin{aligned}\sum_{K\cap\omega\neq\emptyset, K\in\mathcal{T}_h} \int_K \{\nabla u_h(x) - F_\lambda\}\, dx &= \sum_{K\subset\omega_h, K\in\mathcal{T}_h} \int_K \{\nabla u_h(x) - F_\lambda\}\, dx \\ &+ \sum_{K\cap(\omega-\omega_h)\neq\emptyset, K\in\mathcal{T}_h} \int_K \big\{\nabla u_h(x) - F_\lambda\big\}\, dx =: I_1 + I_2.\end{aligned} \tag{2.20}$$

Because of $\mu(\omega - \omega_h) = \mathcal{O}(h)$ and using the Cauchy–Schwarz inequality, we have

$$\begin{aligned}|||I_2||| &\leq \Big|\Big|\Big| \sum_{K\cap(\omega-\omega_h)\neq\emptyset, K\in\mathcal{T}_h} \int_K \big\{\nabla u_h(x) - \pi(\nabla u_h(x))\big\}\, dx \Big|\Big|\Big| \\ &+ \Big|\Big|\Big| \sum_{K\cap(\omega-\omega_h)\neq\emptyset, K\in\mathcal{T}_h} \int_K \big\{\pi(\nabla u_h(x)) - F_\lambda\big\}\, dx \Big|\Big|\Big| \\ &\leq C\left\{ h^{1/2}\left(\sum_{K\in\mathcal{T}_h} \int_K \big|||\nabla u_h(x) - \pi(\nabla u_h(x))|||^2 dx \right)^{1/2} + h \right\} \\ &\leq C\big\{h^{1/2}\big(\mathcal{E}_h^\beta(u_h)\big)^{1/2} + h\big\}.\end{aligned} \tag{2.21}$$

In order to bound the term I_1 in (2.20), we have

$$\begin{aligned}|||I_1||| &= \Big|\Big|\Big| \sum_{K\subset\omega_h, K\in\mathcal{T}_h} \int_K \{\nabla u_h(x) - F_\lambda\}\, dx \Big|\Big|\Big| \\ &= \Big|\Big|\Big| \sum_{K\subset\omega_h, K\in\mathcal{T}_h} \int_{\partial K} \{u_h(x) - F_\lambda x\} \otimes \nu\, d\sigma \Big|\Big|\Big| \\ &\leq C\Big|\Big|\Big| \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\not\subset\partial\omega_h, K\in\mathcal{T}_h} \int_{\mathcal{F}} \{u_h(x) - F_\lambda x\} \otimes \nu\, d\sigma \Big|\Big|\Big| \\ &+ C\Big|\Big|\Big| \sum_{\mathcal{F}\subset\partial K, \mathcal{F}\subset\partial\omega_h, K\in\mathcal{T}_h} \int_{\mathcal{F}} \{u_h(x) - F_\lambda x\} \otimes \nu\, d\sigma \Big|\Big|\Big| =: I_{11} + I_{12}.\end{aligned} \tag{2.22}$$

The term I_{11} can be bounded by Ch^β, according to an identical argument already presented in the proof of Theorem 2.3; see (2.7). In order to bound I_{12}, we can make use of Theorem 2.6, in combination with Theorem 2.5 and Theorem 2.9. If we

introduce the abbreviative notation $z_h(x) = u_h(x) - F_\lambda x$, we find

$$\begin{aligned}\|I_{12}\| &\le C\big(\mu(\partial\omega_h)\big)^{1/2}\left(\sum_{\mathcal{F}\subset\partial K,\mathcal{F}\cap\partial\omega_h\neq\emptyset,K\subset\omega_h}\int_{\mathcal{F}}|z_h^\bullet(x)|^2\,d\sigma\right)^{1/2}\\ &\le \frac{C}{\Lambda(\omega_h)}\Big(\int_{\omega_h}|z_h(x)|^2\,dx\Big)^{1/2}\\ &+ C\Big(\int_{\omega_h}|z_h(x)|^2\,dx\Big)^{1/4}\Big(\sum_{K\subset\omega_h}\int_K\big|\big|\big|\nabla z_h(x)\big|\big|\big|^2\,dx\Big)^{1/4}+Ch^{\beta/2}\\ &\le \frac{C}{\Lambda(\omega_h)}\Big\{\big(\mathcal{E}_h^\beta(u_h)\big)^{1/4}+h^{\beta/2}+h^{(1-\beta)/2}+h^{1-\beta}\Big\}^{1/2}.\end{aligned}$$

In this context the following auxiliary argument ensures the existence of an appropriate constant C:

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\int_K\big|\big|\big|\nabla z_h(x)\big|\big|\big|^2\,dx &\le C\sum_{K\in\mathcal{T}_h}\int_K\Big|\Big|\Big|\nabla u_h(x)-\pi(\nabla u_h(x))\Big|\Big|\Big|^2\,dx\\ &+ C\sum_{K\in\mathcal{T}_h}\int_K\Big|\Big|\Big|\pi(\nabla u_h(x))-F_\lambda\Big|\Big|\Big|^2\,dx\le C.\end{aligned}\tag{2.23}$$

This concludes the proof. ∎

2.4. Approximation of simply laminated microstructure. We start with an approximation result that states how well $\{F_1, F_2\}$ are approximated by a minimizer $u_h \in \mathcal{A}$.

THEOREM 2.11. *There exists a constant $C > 0$ such that the following inequality is valid for a minimizer $u_h \in \mathcal{A}_h$:*

$$\left(\sum_{K\in\mathcal{T}_h}\int_K\Big|\Big|\Big|\nabla u_h(x)-\Pi(\nabla u_h(x))\Big|\Big|\Big|^2\,dx\right)^{1/2}\le C\Big\{h+h^{1/2}+h^{\beta/2}+h^{(1-\beta)/2}\Big\}.$$

Proof. We have $\forall w\in\mathbb{R}^3$, $w\cdot n=0$:

$$\Pi(F)w = F_1w = F_2w = F_\lambda w \qquad \forall\ F\in\mathbb{R}^{3\times3}.$$

The following identity is valid:

$$F-\Pi(F) = F-\pi(F)+\{\Theta(F)-\mathrm{I}\}\Pi(F).\tag{2.24}$$

We can then proceed in the following way:

$$\{\Theta(F)-\mathrm{I}\}\Pi(F)w = \{\pi(F)-F_\lambda\}w \qquad \forall\ F\in\mathbb{R}^{3\times3}.$$

We can now benefit from Lemma 2.8. By choosing a vector $w\in\mathbb{R}^3$ such that $w\cdot n=0$, we have

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\int_K\big|\{\Theta(\nabla u_h(x))-\mathrm{I}\}F_1w\big|^2\,dx &= \sum_{K\in\mathcal{T}_h}\int_K\big|\{\pi(\nabla u_h(x))-F_\lambda\}w\big|^2\,dx\\ &\le C\Big\{\big(\mathcal{E}_h^\beta(u_h)\big)^{1/2}+h^\beta+h^{1-\beta}\Big\}.\end{aligned}\tag{2.25}$$

Now, we choose $w_1, w_2 \in \mathbb{R}^3$, such that it holds $w_1 \cdot n = w_2 \cdot n = 0$ with w_1, w_2 being linearly independent. Set $m = F_1 w_1 \times F_1 w_2$. Since

$$Qm = QF_1 w_1 \times QF_1 w_2 \qquad \forall\, Q \in SO(3),$$

we have $\forall F \in \mathbb{R}^{3\times 3}$:

$$\begin{aligned}\{\Theta(F) - \mathrm{I}\}m &= \big\{\Theta(F)F_1 w_1 \times \Theta(F)F_1 w_2\big\} - \big\{F_1 w_1 \times F_1 w_2\big\} \\ &= \big\{\{\Theta(F) - \mathrm{I}\}F_1 w_1 \times \Theta(F)F_1 w_2\big\} - \big\{F_1 w_1 \times \{\mathrm{I} - \Theta(F)\}F_1 w_2\big\}.\end{aligned}$$

This, together with (2.25), implies

$$\sum_{K\in\mathcal{T}_h} \int_K \big|\big\{\Theta(\nabla u_h(x)) - \mathrm{I}\big\}m\big|^2 \, dx \le C\big\{\big(\mathcal{E}_h^\beta(u_h)\big)^{1/2} + h^\beta + h^{1-\beta}\big\}. \tag{2.26}$$

Now, the triple $\{F_1 w_1, F_1 w_2, m\}$ is a basis for $\mathbb{R}^3$, and (2.25), (2.26) lead to

$$\sum_{K\in\mathcal{T}_h} \int_K \big|\big|\big|\Theta(\nabla u_h(x)) - \mathrm{I}\big|\big|\big|^2 \, dx \le C\big\{\big(\mathcal{E}_h^\beta(u_h)\big)^{1/2} + h^\beta + h^{1-\beta}\big\}.$$

The result now follows from (2.24) and Lemma 2.7. ∎

We are now in a position to verify the following result which states approximation of the volume fraction λ.

THEOREM 2.12. *Suppose $u_h \in \mathcal{A}_h$ to be a minimizer of $\mathcal{E}_h^\beta(\cdot)$. Then, for any rectangular parallelepiped $\omega \subset \Omega$ and any $\rho > 0$, there exists a constant $C = C(\omega, \rho) > 0$ such that the following statement is valid:*

$$\Big|\frac{\mu(\omega_\rho^1(u_h))}{\mu(\omega)} - \lambda^1\Big| + \Big|\frac{\mu(\omega_\rho^2(u_h))}{\mu(\omega)} - \lambda^2\Big| \le C\Big\{h^{1/4} + h^{\beta/4} + h^{(1-\beta)/4}\Big\}.$$

Proof. We start with the following identity, which is a consequence of the definition of the $\omega_\rho^i(u_h)$ given in the beginning of this section:

$$\begin{aligned}&\big\{\,\mu(\omega_\rho^1(u_h)) - \lambda^1\mu(\omega)\big\}F_1 + \big\{\mu(\omega_\rho^2(u_h)) - \lambda^2\mu(\omega)\big\}F_2 \\ &\quad= \sum_{K\in\mathcal{T}_h} \Big\{\int_{K\cap\omega} \big\{\Pi(\nabla u_h(x)) - F_\lambda\big\}\, dx - \int_{K\cap(\omega - \{\omega_\rho^1 \cup \omega_\rho^2\})} \Pi(\nabla u_h(x))\, dx\Big\}.\end{aligned} \tag{2.27}$$

In order to treat the first term, we make use of Theorem 2.11 and Theorem 2.10,

leading to

$$
\begin{aligned}
&\left|\!\left|\!\left| \sum_{K\in\mathcal{T}_h} \int_{K\cap\omega} \big\{\Pi(\nabla u_h(x)) - F_\lambda\big\}\, dx \right|\!\right|\!\right| \\
&\qquad \le \left|\!\left|\!\left| \sum_{K\in\mathcal{T}_h} \int_{K\cap\omega} \big\{\Pi(\nabla u_h(x)) - \nabla u_h(x)\big\}\, dx \right|\!\right|\!\right| \\
&\qquad + \left|\!\left|\!\left| \sum_{K\in\mathcal{T}_h} \int_{K\cap\omega} \big\{\nabla u_h(x) - F_\lambda\big\}\, dx \right|\!\right|\!\right| \\
&\qquad \le \mu^{1/2}(\omega) \left(\sum_{K\in\mathcal{T}_h} \int_K \left|\!\left|\!\left| \Pi(\nabla u_h(x)) - \nabla u_h(x) \right|\!\right|\!\right|^2 dx \right)^{1/2} \\
&\qquad + \left|\!\left|\!\left| \sum_{K\in\mathcal{T}_h} \int_{K\cap\omega} \big\{\nabla u_h(x) - F_\lambda\big\}\, dx \right|\!\right|\!\right| \\
&\qquad \le C\Big\{ \big(\mathcal{E}_h^\beta(u_h)\big)^{1/8} + h^{\beta/4} + h^{(1-\beta)/4} \Big\}.
\end{aligned}
\tag{2.28}
$$

The second term can be bounded according to the definition

$$
\begin{aligned}
&\left|\!\left|\!\left| \sum_{K\in\mathcal{T}_h} \int_{K\cap(\omega-\{\omega_\rho^1\cup\omega_\rho^2\})} \Pi(\nabla u_h(x))\, dx \right|\!\right|\!\right| \le C\mu\big(\omega - \{\omega_\rho^1 \cup \omega_\rho^2\}\big) \\
&\qquad \le \frac{C}{\rho} \sum_{K\in\mathcal{T}_h} \int_{K\cap(\omega-\{\omega_\rho^1\cup\omega_\rho^2\})} \left|\!\left|\!\left| \Pi(\nabla u_h(x)) - \nabla u_h(x) \right|\!\right|\!\right| dx \\
&\qquad \le C \left(\sum_{K\in\mathcal{T}_h} \int_K \left|\!\left|\!\left| \Pi(\nabla u_h(x)) - \nabla u_h(x) \right|\!\right|\!\right|^2 dx \right)^{1/2} \\
&\qquad \le C\Big\{ \big(\mathcal{E}_h^\beta(u_h)\big)^{1/4} + h^{\beta/2} + h^{(1-\beta)/2} \Big\},
\end{aligned}
\tag{2.29}
$$

with the last bound being a consequence of Theorem 2.11. The statement of the theorem now follows from the linear independence of F_1 and F_2 in (2.27). ∎

3. Computational experiments. In order to demonstrate the theoretical results of the previous sections numerically, we consider the prototype problem

$$
\mathcal{E}(v) = \int_\Omega \Big((v_x)^2 - 1 \Big)^2 + (v_y)^2 \, dx \tag{3.1}
$$

on the domain $\Omega = (0,1)\times(0,1) \subset \mathbb{R}^2$ and the deformation $v : \Omega \to \mathbb{R}$. This problem exhibits the crucial characteristics of the full problem in two or three dimensions [4, 5, 12]. However, it is more instructive for numerical experiments, since due to its low dimensionality it is possible to study the deformation v itself and not just its gradient norm.

In order to test the theory for nonaligned meshes, we consider the following generalization of the energy functional:

$$
\mathcal{E}(v) = \int_\Omega \Big((\nabla v \cdot n(\gamma))^2 - 1 \Big)^2 + (\nabla v \cdot w(\gamma))^2 \, dx \tag{3.2}
$$

with

$$n(\gamma) = \begin{pmatrix} \cos(\gamma) \\ \sin(\gamma) \end{pmatrix}, \quad w(\gamma) \in n(\gamma)^{\perp}, \quad -45^\circ \le \gamma \le +45^\circ. \tag{3.3}$$

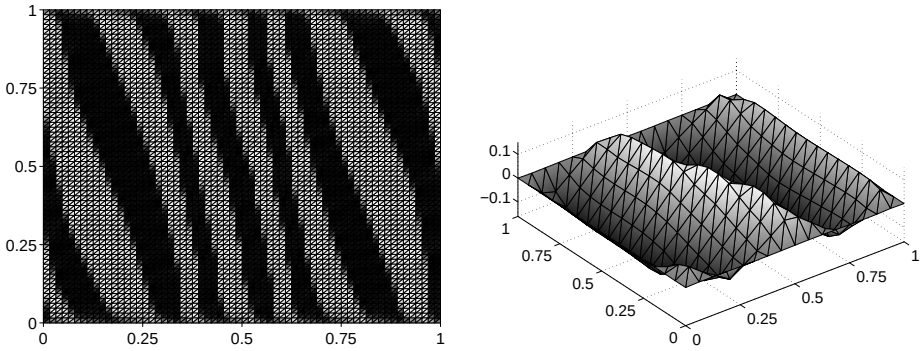
Here, $n = n(\gamma)$ denotes the vector normal to the laminate direction and $w = w(\gamma)$ is a vector along the laminates. Both depend on the angle γ, which denotes the angle between the positive x-axis and the vector n. The domain for the angle was chosen as $[-45^\circ, +45^\circ]$, because it covers the full range of alignments with any regular triangular mesh. Notice that for $\gamma = 0^\circ$, the scaled energy in (3.2) reduces to the original prototype problem (3.1).

A case study for five angles γ, five mesh parameters h, and five finite elements has been performed. (1) The angle γ varied through five values, which cover all degrees of alignment of the physical laminates with the numerical mesh; namely, $\gamma = -45^\circ, -22.5^\circ, 0^\circ, +22.5^\circ, +45^\circ$. (2) The numerical grid is given by a regularly refined triangular mesh, which is independent of the angle γ. To determine the convergence rate for the energy, the mesh parameter was chosen as $h = \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}$.

(3) Classical conforming elements both with exact as well as with relaxed boundary conditions, and classical nonconforming elements with exact as well as relaxed boundary conditions have been tested; in order for the generalized energy functional in Algorithm 1.2 to yield the same scaling as the original energy functional for the classical elements, we set $\beta = 0$ and $\alpha_{11} = \alpha_{12} = 0$. The simulations for the discontinuous finite elements have used the scaled energy functional in Algorithm 1.2 with $\beta = 1/2$. Only graphs for the classical conforming element with exact boundary conditions and the discontinuous elements are included in the following; more information about the case study as well as a more extensive report on the results for all finite elements is contained in [13].

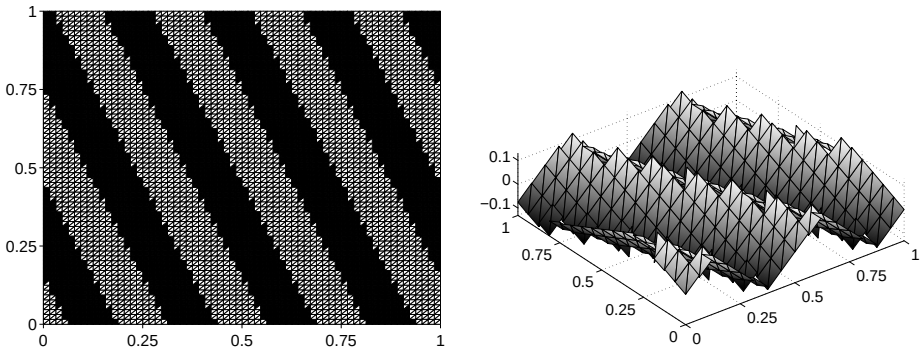
For the graphs presented in the following, the angle $\gamma = +22.5^\circ$ is chosen, since it is the "least" aligned case between the laminates and the grid. It is known that this case poses significant problems for the classical finite element discretizations; see [9, 10, 22]. These problems are clearly visible in Figure 3.1. Figure 3.1(a) shows a plot of the gradient norm for $h = 1/64$. The white color in the graph indicates that the gradient on that finite element is close to $+1$; the black color indicates values that are close to -1. The laminates do not follow their correct direction of $\gamma = +22.5^\circ$ anymore but are rather distorted, and no reliable information concerning the volume fractions can be obtained. This shows the dependence of the numerical solution on the alignment of the grid with the physical laminates. Figure 3.1(b) shows the deformation itself for $h = 1/16$. This size for h was chosen for the plot to allow the features to be large enough to show sufficient detail for observation. This figure shows that the penalty method was successful in enforcing the boundary conditions exactly, as appropriate for this classical conforming element.

Correspondingly, Table 3.1(a) shows the values of the energy functional that were observed for the classical conforming finite element. We notice that in the first and the third column for the angles $\gamma = -45^\circ$ and $\gamma = 0^\circ$, respectively, the convergence rate can be seen to be nearly linear. This is explained by the fact that in these two cases the numerical mesh is aligned with the direction of the physical laminates. In the general case however, as seen in the remaining columns of the table including for $\gamma = +22.5^\circ$, the convergence rate is observed to be (much) worse than linear. Table 3.1(b) lists the corresponding energy values obtained for the classical nonconforming element. While the values themselves are lower than for the conforming element, the same observations hold with respect to the convergence behavior as for the conforming element.



(a) (b)

FIG. 3.1. (a) *Deformation gradient in normal direction for angle* $\gamma = +22.5^\circ$ *using the classical conforming element with exact boundary conditions with* $h = 1/64$. (b) *Deformation for angle* $\gamma = +22.5^\circ$ *using the classical conforming element with exact boundary conditions with* $h = 1/16$.



(a) (b)

FIG. 3.2. (a) *Deformation gradient in normal direction for angle* $\gamma = +22.5^\circ$ *using the discontinuous element with* $h = 1/64$. (b) *Deformation for angle* $\gamma = +22.5^\circ$ *using the discontinuous element with* $h = 1/16$.

Hence, the nonconforming element does not significantly alleviate the problems found with the conforming one.

Figure 3.2 plots the result for the discontinuous finite elements in Algorithm 1.2. These elements use the vertices as degrees of freedom without any continuity requirement in the element definition; the amount of discontinuity allowed is controlled via a penalty technique implemented by the α_{11}- and α_{12}-terms in the scaled energy functional in Algorithm 1.2.

Figure 3.2(a) shows the norm of the deformation gradient in normal direction to the laminates. It is observed that the bands of the laminates are represented crisply and running straight along the direction defined by $\gamma = +22.5^\circ$. The laminates themselves are wider than in the classical cases as required by the energy functional in Algorithm 1.2 with $\beta = 1/2$. Concretely, each laminate has a width of order $\mathcal{O}(h^{1-\beta}) = \mathcal{O}(h^{1/2})$; see section 1. Figure 3.2(b) shows the deformation u. First, we can see also here the value of the deformation gradients in normal direction is close to $+1$ and -1 as exhibited by the hat shape of the solution. As expected, the solution exhibits a degree of discontinuity, but this effect is limited to the element edges corresponding to transitions from one laminate to another, while the laminates

TABLE 3.1

Values of the energy norm for the elements used. (a) *Classical conforming element with exact boundary conditions.* (b) *Classical nonconforming element with exact boundary conditions.* (c) *Discontinuous element.*

	$\gamma = -45°$	$\gamma = -22.5°$	$\gamma = 0°$	$\gamma = +22.5°$	$\gamma = +45°$
$h = 1/4$	0.5754	0.5305	0.4532	0.6690	0.8920
$h = 1/8$	0.3497	0.3604	0.3891	0.4617	0.6182
$h = 1/16$	0.1884	0.2634	0.1939	0.3444	0.4260
$h = 1/32$	0.0972	0.1958	0.0963	0.2512	0.2911
$h = 1/64$	0.0488	0.1587	0.0474	0.1956	0.2320

(a)

	$\gamma = -45°$	$\gamma = -22.5°$	$\gamma = 0°$	$\gamma = +22.5°$	$\gamma = +45°$
$h = 1/4$	0.2297	0.2629	0.3733	0.3833	0.3265
$h = 1/8$	0.1532	0.1744	0.1996	0.2315	0.2200
$h = 1/16$	0.0885	0.0980	0.0992	0.1438	0.1507
$h = 1/32$	0.0470	0.0652	0.0491	0.0960	0.1085
$h = 1/64$	0.0238	0.0392	0.0243	0.0765	0.0679

(b)

	$\gamma = -45°$	$\gamma = -22.5°$	$\gamma = 0°$	$\gamma = +22.5°$	$\gamma = +45°$
$h = 1/4$	0.0165	0.0521	0.0173	0.0849	0.1156
$h = 1/8$	0.0052	0.0222	0.0047	0.0324	0.0481
$h = 1/16$	0.0007	0.0052	0.0007	0.0108	0.0192
$h = 1/32$	0.0005	0.0014	0.0007	0.0033	0.0038
$h = 1/64$	0.0000	0.0003	0.0000	0.0009	0.0013

(c)

themselves are represented continuously.

Table 3.1(c) lists the values of the energy functional for the discontinuous element. It is observed that we have $\mathcal{O}(h^2)$ convergence rate. This is remarkable, in particular, since this result is clearly independent of the degree of alignment of the physical laminates with the numerical grid for all angles γ. That demonstrates the superiority of the discretization using discontinuous finite elements.

The calculations were performed on a Silicon Graphics Challenge XL workstation at the University of Maryland, Baltimore County. The computer program implements the nonlinear conjugate gradient method for the minimization with initial conditions chosen close to the assumed solution. The scaled energy functional of Algorithm 1.2 is discretized using the package FEAT2D [3] for the underlying finite element discretization.

4. Conclusions. This paper proposes the use of discontinuous finite elements for the numerical simulation of crystalline microstructure. This becomes possible via the introduction of a new, generalized energy functional that rests on three fundamental ideas: (1) The boundary conditions are enforced only up to the order of the mesh parameter. (2) The degree of discontinuity is controlled by penalty terms in the energy functional. (3) The laminates and laminate transitions are scaled appropriately relative to each other and relative to the mesh parameter. Using this functional, excellent convergence behavior (second order) for the energy is shown both theoretically and by numerical test calculations. This energy estimate implies much improved estimates for other quantities of interest—for instance for the most crucial volume fractions of the variants of the crystalline microstructure.

Acknowledgments. We would like to thank the Institute for Mathematics and its Applications at the University of Minnesota for the stimulating environment during our stay in 1996–1997 as well as for the use of its computer facilities, on which the numerical code was developed. We also thank the University of Maryland, Baltimore County, on whose computers the case study was performed.

REFERENCES

[1] J. Ball and R. James, *Fine phase mixtures as minimizers of energy*, Arch. Rational. Mech. Anal., 100 (1987), pp. 13–52.
[2] J. Ball and R. James, *Proposed experimental tests of a theory of fine microstructure and the two-well problem*, Philos. Trans. Roy. Soc. London Ser. A, 338 (1992), pp. 389–450.
[3] H. Blum, J. Harig, S. Müller, and S. Turek, *FEAT2D: Finite Element Analysis Tools. User Manual. Release* 1.3, Tech. Rep., Universität Heidelberg, 1992.
[4] C. Carstensen and P. Plecháč, *Numerical solution of the scalar double-well problem allowing microstructure*, Math. Comp., 66 (1997), pp. 997–1026.
[5] M. Chipot and C. Collins, *Numerical approximations in variational problems with potential wells*, SIAM J. Numer. Anal., 29 (1992), pp. 1002–1019.
[6] M. Chipot and S. Müller, *Sharp Energy Estimates for Finite Element Approximations of Non-convex Problems*, Tech. Rep. 8, Max-Planck-Institute for Mathematics in the Natural Sciences, Leipsig, 1997.
[7] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[8] C. Collins, *Computation and Analysis of Twinning in Crystalline Solids*, Ph.D. thesis, University of Minnesota, Minneapolis, MN, 1990.
[9] C. Collins, *Computation of twinning*, in Microstructure and Phase Transitions, IMA Vol. Math. Appl. 54, Springer-Verlag, New York, 1993, pp. 39–50.
[10] C. Collins, *Comparison of computational results for twinning in the two-well problem*, in Proceedings of the 2nd International Conference on Intelligent Materials, C. Rogers and G. Wallace, eds., Technomic, 1994, pp. 391–401.
[11] C. Collins, *Convergence of a reduced integration method for computing microstructures*, SIAM J. Numer. Anal., 35 (1998), pp. 1271–1298.
[12] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, New York, 1989.
[13] M. K. Gobbert and A. Prohl, *A Survey of Classical and New Finite Element Methods for the Computation of Crystalline Microstructure*, Tech. Rep. 1576, IMA, 1998.
[14] P. Gremaud, *Numerical analysis of a nonconvex variational problem related to solid-solid phase transitions*, SIAM J. Numer. Anal., 31 (1994), pp. 111–127.
[15] P. Klouček, B. Li, and M. Luskin, *Analysis of a class of nonconforming finite elements for crystalline microstructure*, Math. Comp., 67 (1996), pp. 1111–1125.
[16] P. Klouček and M. Luskin, *The computation of the dynamics of martensitic microstructure*, Contin. Mech. Thermodyn., 6 (1994), pp. 209–240.
[17] M. Kružík, *Oscillations, Concentrations and Microstructure Modeling*, Ph.D. thesis, Charles University, Prague, 1996.
[18] M. Kružík, *Numerical approach to double well problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1833–1849.
[19] B. Li and M. Luskin, *Finite element analysis of microstructure for the cubic to tetragonal transformation*, SIAM J. Numer. Anal., 35 (1998), pp. 376–392.
[20] B. Li and M. Luskin, *Nonconforming finite element approximation of crystalline microstructure*, Math. Comp., 67 (1998), pp. 917–946.
[21] M. Luskin, *Approximation of a laminated microstructure for a rotationally invariant, double well energy density*, Numer. Math., 75 (1996), pp. 205–221.
[22] M. Luskin, *On the computation of crystalline microstructure*, Acta Numerica, 5 (1996), pp. 191–257.
[23] A. Prohl, *An adaptive finite element method for solving a double well problem describing crystalline microstructure*, RAIRO Modél. Math. Anal. Numér., to appear.
[24] T. Roubíček, *Relaxation in Optimization Theory and Variational Calculus*, Walter de Gruyter, Berlin, New York, 1997.