Spurious Oscillations are Not Fatal in Computing Microstructures*

Charles R. Collins†

Abstract

Recent interest in materials which form microstructures has lead to efforts to determine the material configurations computationally. One approach has been to use a bilinear element with a one-point quadrature scheme. Although the computational results using this approximation appear to agree with the theory, it has been conjectured that this approach leads to spurious oscillations which contaminate the solution. This paper shows that this hypothesis is false. In fact, all the estimates which hold for conforming approximations, hold in this case.

1 Introduction

One of the key features of modern materials is their ease in forming fine-scale features within the material. In shape memory alloys, a microstructure called twinning, allows the material to recover from plastic-like deformations without any defects. Microstructure has been observed in various materials over the years, but now there is a renewed interest as mathematical theory is playing an increasing role in the study of microstructure and material phenomenon [36, 32]. The theory developed for the microstructure of twinning is elegant (see e.g. [3, 8]), but the results in many cases are very difficult to analyze. So, it is natural to turn to numerical approximations to complement the theoretical and experimental work. For an excellent survey on computing microstructures see the review by Luskin [31]. Besides the computational results, there have been many efforts to prove error estimates for various versions of the problem with different types of approximations. The first works dealt with one-dimensional and scalar valued problems which generated microstructures using piecewise linear approximations [13, 17, 6, 5]. The next papers treated the full rotationally invariant multi-dimensional problem with various types of finite elements. Gremaud [23] provided the first estimates, using a flexible non-conforming element. Chipot, Collins and Kinderlehrer [7]

*This work was supported in part by the Institute for Mathematics and its Applications and by a Professional Development Award from the University of Tennessee.

†Department of Mathematics, University of Tennessee, Knoxville, TN 37996-1300, http://www.math.utk.edu/∼ccollins

and Luskin [30] proved estimates for conforming elements. Recently, Li and Luskin [28, 29] have derived estimates for a physically relevant three-dimensional energy model using a type of non-conforming element.

This paper continues in the trend of the last group of papers by treating the full rotationally invariant case. The difference is that the problem is approximated using a bilinear finite element with reduced integration. This element and quadrature scheme was used in early computations by Collins, Luskin and Riordan [14, 15, 16, 10, 11, 18, 19, 20] but has never been analyzed. Because this approximation uses reduced integration to evaluate the target function, it has been conjectured that, like in the Stokes problem [4], the extra degrees of freedom will produce spurious oscillations in the solution which will lead to 'fatally flawed' results [26, 33]. Inspired by previous analysis [7, 30], this paper shows that this hypothesis is false. In fact, as long as all relevant quantities are evaluated using the reduced integration scheme, all the estimates which hold for conforming elements, (see e.g. [30]), hold in this case with equivalent bounds.

In the sections that follow, we will first give a brief description of twinning, the model problem, some basic properties of the problem and the numerical approximation. Then, we prove existence of a solution, and give an estimate for the energy of an approximate solution. Next, we relate properties of the approximations to the expected values and show that the approximation using the reduced integration scheme satisfies each property with a bound similar to the conforming case in [30]. Finally, in the last section, we provide some numerical experiments which show that the estimates are valid, even in the case which has the largest potential to produce spurious oscillations. This last section also provides some indication as to the optimality of the estimates.

2 Twinning and the Two-Well Problem

Twinning is a type of microstructure which occurs as the result of a solid-to-solid phase transformation which changes the local symmetry of the material. For example in NiTn alloys, the crystalline structure of the material changes from cubic to tetragonal symmetry. To study twinning we use an energy framework based on the work of Ericksen [21, 22] and Pitteri [34, 35]. Starting from the crystallographic description of the material, we have an energy density function which depends on changes in the crystal lattice. Then, applying the Born rule, we convert the discrete model into a continuum model. In this framework, we consider two phases of the material. The high temperature or parent phase of the material, commonly called Austenite, has a certain material symmetry and the continuum energy is invariant under the action of this symmetry group. The lower temperature phase, called Martensite, is marked by a decrease in the symmetry of the material from the Austenitic phase. Since the energy model is invariant under a larger symmetry group, at low temperatures there are many deformations, called variants of Martensite, which minimize the energy. This leads to the energy density containing multiple wells. The microstructure of twinning can then be described as fine-scale laminate, with the bands containing different variants of Martensite, with their arrangement limited only by compatibility between neighboring

variants. Twinning is a two-dimensional phenomena, and involves only variants from two wells at a time, so we will focus on a two-dimensional, two-well problem.

We start with some basic notation and definitions. For $w \in \mathrm{R}^2$ we use the standard Euclidean norm: $|w|^2 = \sum_{i=1}^2 w_i^2 = w \cdot w$. For deformation gradients, $\mathcal{M} = \{A \in \mathrm{R}^{2\times 2} : \det A > 0\}$, the norm $|A|$ is computed as the standard subordinate matrix norm for the vector norm $|\cdot|$, i.e.

$$|A| = \max_{w\in\mathrm{R}^2, w\neq 0} \frac{|Aw|}{|w|}.$$

A special subgroup of these matrices are the rotations. A rotation is a matrix $Q \in \mathcal{M}$ such that $\det Q = 1$ and $Q^T Q = I$, where I is the 2×2 identity matrix. With the norms above, we have for any rotation Q, deformation A and vector w that $|QA| = |A|$ and $|Qw| = |w|$.

In addition, for functions $f \in L^p(\Omega)$ we will use the standard L^p-norm, denoted by $\|f\|_{L^p(\Omega)}$. For sets $\Omega \subset \mathrm{R}^n$, where Ω has dimension $m \leq n$, we use $|\Omega|$ to indicate the m-dimensional measure of the set Ω.

For the two-well energy, we start with an idealized material which, due to a loss of symmetry in the Austenite to Martensite transformation, now prefers two different low-energy orientations. These orientations are represented by deformation gradients W_0 and W_1. We assume that these orientations are distinct and are compatible for twinning, i.e. $W_1 - W_0 = a \otimes n$ with $a \neq 0$ and $|n| = 1$. Next, these low energy deformation gradients may occur, with respect to an observer, at any angle, and thus we define the wells for $i = 0, 1$ by

$$\mathcal{W}_i = \{A \in \mathcal{M} : A = QW_i \text{ for a rotation } Q\}. \tag{1}$$

Thus we would expect that in our material that we would see low-energy configurations whose deformation gradient is in or around the set $\mathcal{W}_0 \cup \mathcal{W}_1$.

To understand the behavior around these wells, we define a projection $\pi : \mathcal{M} \to \mathcal{W}_0 \cup \mathcal{W}_1$ by choosing a rotation Q and an orientation W_i which minimizes $|A - QW_i|$. Then $\pi(A) = QW_i$. We can always choose the rotation and orientation so that π is a well-defined Borel function. We can then decompose $\pi(A)$ into $\theta(A)\Pi(A)$ where, when $\pi(A) = QW_i$, we have $\theta(A) = Q$ and $\Pi(A) = W_i$. In summary,

$$\pi(A) = \theta(A)\Pi(A) = QW_i \tag{2}$$

for some rotation Q and orientation W_i.

To model our material and the expectations expressed above, we use a continuous energy density $\phi : \mathcal{M} \to \mathrm{R}$ with the following three basic properties:

I. ϕ is rotationally invariant, i.e.

$$\phi(QA) = \phi(A) \text{ for all rotations } Q \text{ and deformation gradients } A. \tag{3}$$

II. ϕ is a two-well potential, i.e.

$$\phi(A) \geq 0 \text{ and } \phi(A) = 0 \text{ if and only if } A \in \mathcal{W}_0 \cup \mathcal{W}_1. \tag{4}$$

III. ϕ is 'elliptic' with respect to the projection π, i.e. there is a constant $\nu > 0$ such that

$$\phi(A) \geq \nu |A - \pi A|^2. \tag{5}$$

Next, we associate an undeformed sample of our material in the Austenitic or parent phase with the reference domain, Ω, which we take as a rectangular region in R^2. Then we consider continuous deformations $u : \Omega \to \mathrm{R}^2$ and associate with each deformation a bulk energy given by

$$\mathcal{E}(u) = \int_\Omega \phi(\nabla u)\, dx.$$

Rather than consider all possible deformations, we consider a specific set of admissible deformations corresponding to dead-loading on the boundary, which, in experiments, lead to the formation of a microstructure.

For the boundary conditions, use an deformation based on an average of two variants. Let $0 < \lambda < 1$ and form the matrix

$$F = (1 - \lambda)W_0 + \lambda W_1 = W_0 + \lambda a \otimes n. \tag{6}$$

We expect deformations to choose their gradients from near the wells and thus the gradients should be bounded, but possibly discontinuous. With this in mind, we choose the admissible deformations as

$$\mathcal{A} = \{u \in W^{1,\infty}(\Omega; \mathrm{R}^2) : \det \nabla u > 0, u(x) = Fx \text{ for } x \in \partial\Omega\}. \tag{7}$$

Then the problem is to find a function $u \in \mathcal{A}$ which minimizes the energy $\mathcal{E}$.

With this boundary condition it has been shown (see e.g. [2, 3, 24]), that $\inf_{v \in \mathcal{A}} \mathcal{E}(v) = 0$ yet there is no $u \in \mathcal{A}$ with $\mathcal{E}(u) = 0$. The main analytic approach then is to study energy minimizing sequences. A typical sequence consists of functions whose gradients oscillate between variants of Martensite on a finer and finer scale, so that the total energy decreases to zero in the limit. For boundary conditions based on mixtures of variants, the material can only achieve zero energy through the formation of a microstructure. To study the microstructure associated with the minimizing sequence, a Young measure [37] is used. The Young measure indicates the type and distribution of the gradient values in the limit. In the case we are studying of the two-variant boundary condition, it has been shown that although there may be several possible minimizing sequences, they all generate the same Young measure [3]. In this case the Young measure is given by

$$\nu_x = (1 - \lambda)\delta_{W_0} + \lambda \delta_{W_1} \tag{8}$$

where δ_A is the Dirac delta function with mass at A. This Young measure can be used to estimate properties of the limiting deformation. For example, if we have a minimizing sequence $\{u_k\}$ and we wish to compute $\lim_{k\to\infty} \int_\Omega f(x, \nabla u_k(x))\, dx$ for a continuous function f, then the value of the limit is equal to

$$\int_\Omega \int_{\mathcal{M}} f(x, A)\, d\nu_x(A)\, dx = \int_\Omega (1 - \lambda) f(x, W_0) + \lambda f(x, W_1)\, dx. \tag{9}$$

With the solution represented by a Young measure, which can be thought of as an infinitely fine microstructure [27], it is natural to turn to computations to determine what the structure is on a fixed scale or for a particular form of the energy. Also, for functions $u \in \mathcal{A}$ which are not piecewise linear, $\mathcal{E}(u)$ is an expensive functional to evaluate and an impossible function to minimize, so we can also use computation to find approximate minimizers. To start, we approximate $\mathcal{A}$ using the finite element method based on piecewise bilinear functions (see e.g. Ciarlet [9]). We choose these elements because, when combined with the quadrature rule below, they are more flexible in representing microstructure than conforming piecewise linear elements [11, 12].

Let τ_h be a triangulation of Ω into squares K of size $h \times h$. Let $Q_1(K)$ be the set of all bilinear polynomials on K. Then let

$$\mathcal{B}_h = \{u \in C(\Omega; \mathrm{R}^2) : u|_K \in (Q_1(K))^2 \text{ for all } K \in \tau_h\}, \tag{10}$$

and then set

$$\mathcal{A}_h = \mathcal{A} \cap \mathcal{B}_h. \tag{11}$$

Even with piecewise bilinear functions, the energy is still difficult to evaluate efficiently, so we use a one-point quadrature rule based on the midpoint $m_K = (x_m, y_m)$ of the element K to evaluate the energy. We define the approximate energy by

$$\mathcal{E}_h(u) = \sum_{K \in \tau_h} |K| \phi(\nabla u(m_K)). \tag{12}$$

We can view this energy in a different form, as

$$\mathcal{E}_h(u) = \int_\Omega \phi(\nabla_h u)\, dx \tag{13}$$

where for $x \in K \in \tau_h$

$$\nabla_h u(x) = \fint_K \nabla u(y)\, dy = \frac{1}{|K|} \int_K \nabla u(y)\, dy. \tag{14}$$

These two forms of $\mathcal{E}_h$ are the same for $u \in \mathcal{B}_h$. So the approximation problem is to find $u_h \in \mathcal{A}_h$ which minimizes $\mathcal{E}_h$. Since in practice it is difficult to compute such a function, we will also relate the microstructural properties of functions $v \in \mathcal{A}_h$ to their bulk energies $\mathcal{E}_h(v)$.

3 Existence and Energy Estimate

In all the results that follow, the constants $C_1, \ldots, C_7$ are consistent across each theorem, while the constants B_i are local to each theorem. We begin by showing that in this framework there is a solution to the minimization problem and show its energy goes to zero as the mesh size h goes to zero.

Theorem 1 *There exists a $u_h \in \mathcal{A}_h$ such that*

$$\mathcal{E}_h(u_h) = \inf_{v \in \mathcal{A}_h} \mathcal{E}_h(v) \tag{15}$$

and a constant C_1, such that

$$\mathcal{E}_h(u_h) \leq C_1 h^{1/2}. \tag{16}$$

Proof: To show existence, we show that taking the infimum over $\mathcal{A}_h$ is the same as taking the minimum over a bounded, finite-dimensional set. Combining this with the continuity of $\mathcal{E}_h$, we get the desired result.

Using Theorem 3 (59), (which does not depend upon this theorem), we have that for any $v \in \mathcal{A}_h$ with $\mathcal{E}_h(v) \leq 1$ that

$$\int_\Omega |v - Fx|^2 \, dx \leq B_1 \tag{17}$$

where B_1 is a constant independent of v. Since $v \in \mathcal{A}_h$, we can write

$$\int_\Omega |v - Fx|^2 \, dx = \sum_\alpha h^2 w_\alpha |v(x_\alpha) - Fx_\alpha|^2 \tag{18}$$

where $\{x_\alpha\}$ are the nodes of the finite element mesh and w_α are appropriate weights independent of h. Thus for a fixed value of h, and for every α, $|v(x_\alpha) - Fx_\alpha|$ is bounded uniformly. And thus there is a constant B_2, such that

$$|v(x_\alpha)| \leq B_2 \tag{19}$$

for all α. Thus taking the infimum over $\mathcal{A}_h$ is the same as taking the infimum over $\hat{\mathcal{A}}_h$, where

$$\hat{\mathcal{A}}_h = \{v \in \mathcal{A}_h : \mathcal{E}_h(v) \leq 1 \text{ and } |v(x_\alpha)| \leq B_2, \forall \alpha\}. \tag{20}$$

Thus, since ϕ is continuous, so is $\mathcal{E}_h$ and thus the existence of a minimizer is proven. Note: we can still use Theorem 3 (59) to get the bound B_1, as long as $\mathcal{E}_h(v)$ is bounded, and thus the next part guarantees that $\hat{\mathcal{A}}_h \neq \emptyset$.

To estimate the minimum energy, we will use a common approach of constructing an approximate minimizer using a laminar construction [7, 30]. First, we use the fact that $W_1 = W_0 + a \otimes n$ to construct a continuous function u, such that $\nabla u = W_i$, a.e. in a certain proportion and where the change from $\nabla u = W_0$ to $\nabla u = W_1$ occurs on lines with normal n, spaced a distance d apart. The next step is to adjust this function to be an element of $\mathcal{A}_h$, by matching the boundary conditions and the structure of the finite element approximation. The last step is to express the effects of these adjustments on the energy and to choose the spacing d appropriately.

Let $\chi(t)$ be periodic with period 1, with values as follows:

$$\chi(t) = \begin{cases} 0 & \text{for } 0 \leq t \leq 1 - \lambda \\ 1 & \text{for } 1 - \lambda < t < 1 \end{cases}. \tag{21}$$

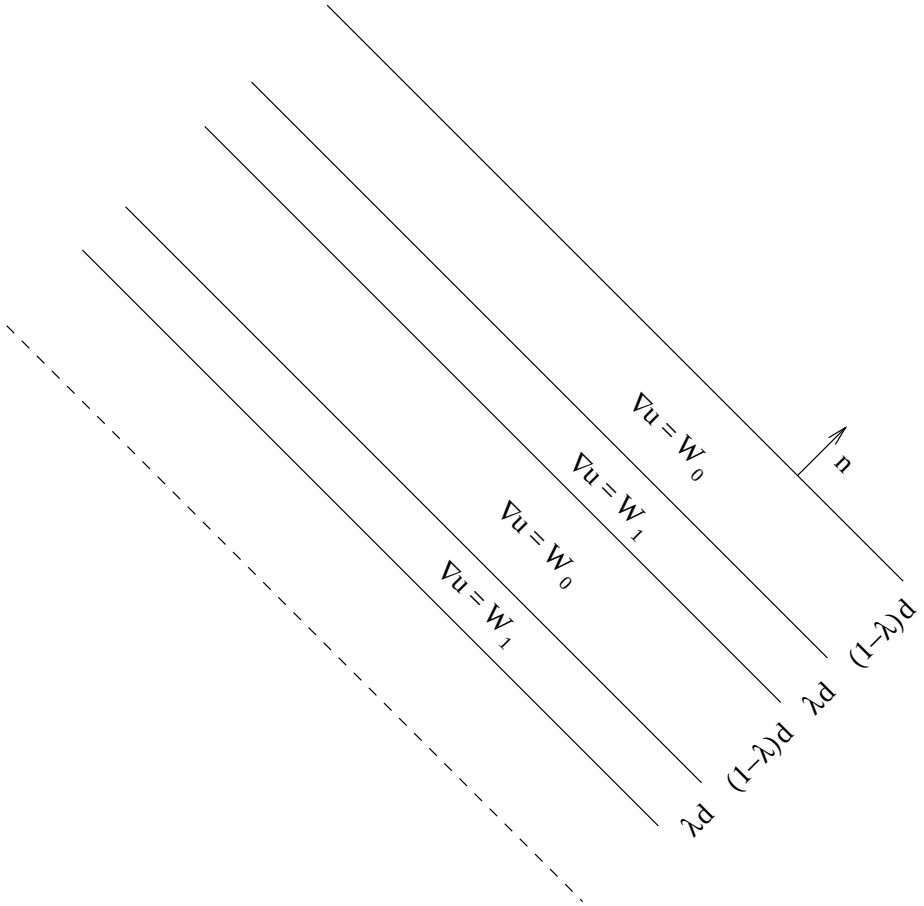

Figure 1: Example of a continuous deformation with $\nabla u = W_i$

Next, let $d > 0$ and construct the function

$$u(x) = W_0 x + \left(\int_0^{x \cdot n} \chi(\tfrac{t}{d})\, dt \right) a. \tag{22}$$

A quick calculation shows that

$$\nabla u(x) = W_0 + \chi(\tfrac{x \cdot n}{d}) a \otimes n \tag{23}$$

and thus $\nabla u = W_i$ a.e. (See Figure 1). And, so $\mathcal{E}(u) = \mathcal{E}_h(u) = 0$.

Next, since $F = W_0 + \lambda a \otimes n$ and $\int_0^1 \chi(t)\, dt = \lambda$ we have that

$$u(x) - Fx = \left(\int_0^{x \cdot n} \chi(\tfrac{t}{d})\, dt - \lambda (x \cdot n) \right) a \tag{24}$$

and letting $s = \frac{x \cdot n}{d} = p + r$ where p is an integer and $0 \le r < 1$, we get

$$\begin{array}{rcl} \left| \int_0^{sd} \chi(\frac{t}{d})\, dt - \lambda(sd) \right| & = & d \left| \int_0^s \chi(t)\, dt - \lambda s \right| \\ & = & d \left| \int_0^r \chi(t)\, dt - \lambda r \right| \\ & \le & d\lambda^2. \end{array} \tag{25}$$

So

$$|u(x) - Fx| \leq d\lambda^2|a|. \tag{26}$$

Now we begin the adjustments, beginning with the boundary conditions: let

$$\psi(x) = \min(1, \tfrac{1}{d}\text{dist}(x, \partial\Omega)) \tag{27}$$

so that $\psi = 0$ on the boundary and $\psi(x) = 1$ if x is at least distance d away from the boundary. Also we have that $|\nabla\psi| \leq \frac{1}{d}$. So, to match the boundary conditions, we use the convex combination

$$\hat{u}(x) = \psi(x)u(x) + (1 - \psi(x))Fx. \tag{28}$$

Note that in $\Omega_d = \{x \in \Omega : \psi(x) = 1\}$ we have that $\nabla\hat{u} = \nabla u = W_i$ and on $\Omega \setminus \Omega_d$ we have that

$$\begin{array}{rcl} |\nabla\hat{u}(x)| & \leq & |u(x) - Fx|\,|\nabla\psi(x)| + |\psi(x)|\,|\nabla u(x)| + (1 - \psi(x))|F| \\ & \leq & \lambda^2|a| + \max(|W_i|) + |F|. \end{array} \tag{29}$$

Thus since ϕ is continuous, $\phi(\nabla\hat{u}(x))$ is bounded by some constant B_3 on $\Omega \setminus \Omega_d$ and $\phi(\nabla\hat{u}(x)) = 0$ on Ω_d.

Next, choose $\hat{u}_h \in \mathcal{A}_h$ such that $\hat{u}_h(x_\alpha) = \hat{u}(x_\alpha)$, for all nodes x_α in the triangulation. i.e. $\hat{u}_h$ interpolates $\hat{u}$.

Now we assess the changes to the energy due to the two adjustments. First, we assume that $d \geq h$ as on τ_h oscillations on a finer scale cannot be represented. From the boundary adjustment the only changes are on the domain $\Omega \setminus \Omega_d$ which has size less than $|\partial\Omega|d$. Thus the contribution to the energy is less than $B_3|\partial\Omega|d$. Next, interpolating $\hat{u}$ changes the value of $\nabla\hat{u}$ in the elements where $\nabla\hat{u}$ changes from W_0 to W_1 or vice versa. This only occurs on the lines where $x \cdot n = pd$ or $x \cdot n = (p - \lambda)d$ for some integer p. In these elements $|\nabla_h\hat{u}_h|$ is bounded, depending only on W_0 and W_1, and so $\phi(\nabla_h\hat{u}_h) \leq B_4$ on these elements. And so each line contributes at most its length ($\leq \text{diam}\,(\Omega)$) times the width affected ($\leq \sqrt{2}h$) times B_4. The number of such lines is less than $2\,\text{diam}\,(\Omega)/d$, so we have

$$\mathcal{E}_h(\hat{u}_h) \leq B_3|\partial\Omega|d + 2\sqrt{2}B_4\text{diam}\,(\Omega)^2 h/d. \tag{30}$$

This bound is minimized by choosing $d = B_5 h^{1/2}$, where

$$B_5 = \sqrt{\frac{2\sqrt{2}B_4\text{diam}\,(\Omega)^2}{B_3|\partial\Omega|}}. \tag{31}$$

With this value of d we get that

$$\mathcal{E}_h(\hat{u}_h) \leq C_1 h^{1/2}, \tag{32}$$

where

$$C_1 = \sqrt{2\sqrt{2}B_3B_4\text{diam}\,(\Omega)^2|\partial\Omega|}. \qquad \square \tag{33}$$

4 Properties of the Solution

Now that we have shown that a minimizer exists and found a bound on its energy, we consider in this section how the properties of a function $v \in \mathcal{A}_h$ depend on the mesh size h and the energy $\mathcal{E}_h(v)$.

First, we start with some properties of the bilinear functions. For this we need the function approximation denoted by the operator A_h, which acts as follows:

$$A_h u(x) = \fint_K u(y)\, dy, \tag{34}$$

when $x \in K \in \tau_h$.

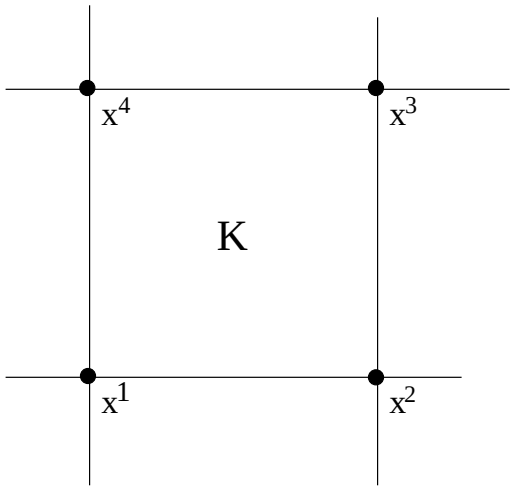

Figure 2: Typical Element K with local node numbers

Lemma 1 *Let $v \in \mathcal{B}_h$, $K \in \tau_h$ and $m_K = (x_m, y_m)^T$ be the midpoint of K. Then for $x \in K$ we can write*

$$v(x) = A_h v + \nabla_h v \cdot (x - m_K) + D_h(v)(x - x_m)(y - y_m) \tag{35}$$

where

$$\begin{aligned} A_h v_l &= \tfrac{1}{4}(v_l^1 + v_l^2 + v_l^4 + v_l^3) \\ \nabla_h v_l &= \tfrac{1}{2h}[v_l^3 + v_l^2 - (v_l^4 + v_l^1), v_l^3 + v_l^4 - (v_l^2 + v_l^1)] \\ D_h(v)_l &= \tfrac{1}{h^2}(v_l^3 + v_l^1 - (v_l^4 + v_l^2)), \end{aligned} \tag{36}$$

and where $v(x) = (v_1(x), v_2(x))$, $v_l^i = v_l(x^i)$ and the x^i are the four nodes defining K starting with the lower left corner and numbering them counterclockwise (see Figure 2). From this, we also have

$$\nabla v = \nabla_h v + D_h(v)(y - y_m, x - x_m) = \nabla_h v + D_h(v) \otimes (y - y_m, x - x_m). \tag{37}$$

Proof: The first expression, (35), simply comes from writing $v \in \mathcal{B}_h$ in terms of the standard basis functions for bilinear approximations (see e.g. Ciarlet [9]) and then rewriting them in the form given. The second expression, (37), comes from taking the derivative of equation (35). □

The main difference between the approximation we use and the conforming approximation is measured by the value of $D_h(v)$ as defined above. Now we estimate the size of $|D_h(v)|$.

Lemma 2 *For $v \in \mathcal{A}_h$ and using the above notation (35-36), we have that*

$$|D_h(v)| \leq \frac{2}{h}|\nabla_h v|. \tag{38}$$

Proof: From the definition of $\mathcal{A}$, we have that $\det \nabla v > 0$ and thus $\det \nabla_h v > 0$, so using Lemma 1 (37) we have

$$\det \nabla v = (\det \nabla_h v)(1 + (\nabla_h v)^{-1} D_h(v) \cdot (y - y_m, x - x_m)) > 0. \tag{39}$$

For this to hold, we must have

$$(\nabla_h v)^{-1} D_h(v) \cdot (y - y_m, x - x_m) > -1, \tag{40}$$

for all $(x, y) \in K$. We can choose $(x, y) \in K$ such that $|(y - y_m, x - x_m)| = \frac{h}{2}$ and

$$(\nabla_h v)^{-1} D_h(v) \cdot (y - y_m, x - x_m) = -|(\nabla_h v)^{-1} D_h(v)| \frac{h}{2}. \tag{41}$$

Thus using this (x, y) pair, we get

$$|(\nabla_h v)^{-1} D_h(v)| < \frac{2}{h}. \tag{42}$$

From this we have, as desired,

$$|D_h(v)| = |\nabla_h v (\nabla_h v)^{-1} D_h(v)| \leq |\nabla_h v| \, |(\nabla_h v)^{-1} D_h(v)| < \frac{2}{h}|\nabla_h v|. \qquad \square \tag{43}$$

This result shows that ∇v can vary significantly from $\nabla_h v$ on each element and thus provides the potential for spurious oscillations. However, we will see that the variance is not large enough to significantly change any of the needed results. Let us look at some more general properties of A_h and ∇_h.

Lemma 3 *If $\omega \subset \Omega$ is a union of elements in τ_h then*

$$\int_\omega A_h f(x)\, dx = \int_\omega f(x)\, dx, \tag{44}$$

for any f for which the integrals exist.

Similarly, we have that

$$\int_\omega \nabla_h f(x)\,dx = \int_\omega \nabla f(x)\,dx, \tag{45}$$

when the integrals exist.

Also, we have for any functions f and g for which these integrals exist, and function h which is constant on each element $K \in \tau_h$, that

$$\int_\omega g(x)(f(x) - A_h f(x))\,dx = \int_\omega (g(x) - h(x))(f(x) - A_h f(x))\,dx. \tag{46}$$

Proof: Let $S \subset \tau_h$, so that $\omega = \cup_{K\in S} K$. Then applying the definition of A_h (34), we have

$$\begin{array}{rcl} \int_\omega A_h f(x)\,dx &=& \sum_{K\in S} \int_K A_h f(x)\,dx \\ &=& \sum_{K\in S} \int_K f(x)\,dx \\ &=& \int_\omega f(x)\,dx. \end{array} \tag{47}$$

For (45), we note that $\nabla_h f = A_h \nabla f$ and apply (44) to $\nabla_h f$.

Let f and g be functions for which these integrals exist and let h be a function such that $h(x) = h_K$ for $x \in K$, then

$$\begin{array}{rcl} \int_\omega g(x)(f(x) - A_h f(x))\,dx &=& \sum_{K\in S} \int_K g(x)(f(x) - A_h f(x))\,dx \\ &=& \sum_{K\in S} \int_K (g(x) - h_K)(f(x) - A_h f(x))\,dx \\ && + \sum_{K\in S} h_K \int_K f(x) - A_h f(x)\,dx \\ &=& \int_\omega (g(x) - h(x))(f(x) - A_h f(x))\,dx, \end{array} \tag{48}$$

since the last sum is zero via (44) with $\omega = K$. □

Continuing, we establish some useful bounds involving ∇_h.

Lemma 4 *For all $v \in \mathcal{B}_h$, we have*

$$\int_\Omega |\nabla_h v - \pi \nabla_h v|^2\,dx \le \nu^{-1} \mathcal{E}_h(v). \tag{49}$$

If, in addition, $\mathcal{E}_h(v) \le 1$, then

$$\begin{array}{rcl} \int_\Omega |\nabla_h v|^2\,dx &\le& C_2 \\ \int_\Omega |\nabla_h v - F|^2\,dx &\le& C_3 \\ \int_\Omega |\nabla v - \nabla_h v|^2\,dx &\le& C_2 \end{array} \tag{50}$$

where

$$\begin{array}{rcl} C_2 &=& \frac{2}{\nu} + 2|\Omega| \max(|W_i|^2) \\ C_3 &=& 2C_2 + 2|\Omega|\,|F|^2. \end{array} \tag{51}$$

Proof: For any $v \in \mathcal{B}_h$ and from the 'ellipticity' property of ϕ (5), we have

$$|\nabla_h v - \pi \nabla_h v|^2 \le \nu^{-1} \phi(\nabla_h v).$$

Multiplying this inequality by $|K|$ and then summing over all $K \in \tau_h$ as in (12), gives (49).

For the others, assume $\mathcal{E}_h(v) \leq 1$, and we have

$$\begin{array}{rcl} \int_\Omega |\nabla_h v|^2\,dx & \leq & 2\int_\Omega |\nabla_h v - \pi\nabla_h v|^2\,dx + 2\int_\Omega |\pi\nabla_h v|^2\,dx \\ & \leq & \frac{2}{\nu}\mathcal{E}_h(v) + 2|\Omega|\max(|W_i|^2) \\ & \leq & C_2. \end{array} \tag{52}$$

Using this result, we get

$$\begin{array}{rcl} \int_\Omega |\nabla_h v - F|^2\,dx & \leq & 2\int_\Omega |\nabla_h v|^2\,dx + 2|\Omega|\,|F|^2 \\ & \leq & C_3. \end{array} \tag{53}$$

For the last estimate, we have that $|\nabla v - \nabla_h v| = |D_h(v)|\,|(y - y_m, x - x_m)|$ and since $|(y - y_m, x - x_m)| = |x - m_K|$ and $|K| = h^2$ we get

$$\begin{array}{rcl} \int_\Omega |\nabla v - \nabla_h v|^2\,dx & = & \sum_{K\in\tau_h}\int_K |\nabla v - \nabla_h v|^2\,dx \\ & = & \sum_{K\in\tau_h} |D_h(v)|^2\int_K |x - m_K|^2\,dx \\ & \leq & \sum_{K\in\tau_h} \frac{4}{h^2}|\nabla_h v|^2\frac{h^4}{12} \\ & \leq & \sum_{K\in\tau_h}\int_K |\nabla_h v|^2\,dx \\ & = & \int_\Omega |\nabla_h v|^2\,dx = C_2, \end{array} \tag{54}$$

via Lemma 2 (38). □

Now we begin a series of results which relate the properties of an arbitrary $v \in \mathcal{A}_h$ with the value of its energy $\mathcal{E}_h(v)$. First, we expect oscillations in the direction n, but in directions perpendicular to n we expect low energy functions to be smooth (in fact, planar) and close to Fx. With this in mind we have the following estimate:

Theorem 2 *For all $v \in \mathcal{A}_h$, with $\mathcal{E}_h(v) \leq 1$ and for all $m \in R^2$ with $m \cdot n = 0$ and $|m| = 1$ we have*

$$\int_\Omega |(\nabla_h v - F)m|^2\,dx \leq C_2\mathcal{E}_h(v)^{1/2}. \tag{55}$$

Proof: Let $v \in A_h$ with $\mathcal{E}_h(v) \leq 1$ and $m \in \mathrm{R}^2$ with $m \cdot n = 0$ and $|m| = 1$. Using the divergence theorem and Lemma 3 (45) we have that the average value of $\nabla_h v$ is F, i.e.

$$F = \fint_\Omega \nabla v\,dx = \fint_\Omega \nabla_h v\,dx, \tag{56}$$

and so

$$\begin{array}{rcl} \fint_\Omega |(\nabla_h v - F)m|^2\,dx & = & \fint_\Omega |(\nabla_h v)m|^2 - 2(\nabla_h v)m \cdot Fm + |Fm|^2\,dx \\ & = & \fint_\Omega |\nabla_h v m|^2\,dx - \fint_\Omega |Fm|^2\,dx \\ & = & \fint_\Omega |(\nabla_h v - \pi\nabla_h v)m + (\pi\nabla_h v)m|^2\,dx - |Fm|^2. \end{array} \tag{57}$$

Next, in the direction m, W_0, W_1 and F are indistinguishable. From $W_1 = W_0 + a \otimes n$ and $F = W_0 + \lambda a \otimes n$, we have that $W_1 m = W_0 m = Fm$ and thus for any $A \in \mathcal{M}$,

$|\pi(A)m| = |QW_i m| = |W_i m| = |Fm|$. Combining this with the last result gives for any $\mu > 0$ that

$$\begin{aligned} ⨍_\Omega |(\nabla_h v - F)m|^2\,dx &\le ⨍_\Omega (1+\tfrac{1}{\mu})|(\nabla_h v - \pi\nabla_h v)m|^2 \\ &\quad +(1+\mu)|(\pi\nabla_h v)m|^2\,dx - |Fm|^2 \\ &\le \tfrac{\mu+1}{\mu}\ ⨍_\Omega |\nabla_h v - \pi\nabla_h v|^2\,dx + \mu|Fm|^2 \\ &\le \tfrac{(\mu+1)\mathcal{E}_h(v)}{\mu\nu|\Omega|} + \mu|Fm|^2, \end{aligned} \tag{58}$$

using Lemma 4 (50). Then (55) follows by noting that $|Fm|^2 \le \max|W_i|^2$ and taking $\mu = \mathcal{E}_h(v)^{1/2}$. □

Now with this result, we can estimate the difference between v and Fx over the entire domain Ω, thus showing that any $v \in \mathcal{A}_h$ with low energy cannot have significant spurious oscillations.

Theorem 3 *For $v \in \mathcal{A}_h$ with $\mathcal{E}_h(v) \le 1$, we have*

$$\int_\Omega |v(x) - Fx|^2\,dx \le 8M^2 C_2 \mathcal{E}_h(v)^{1/2} + \mathcal{O}(h), \tag{59}$$

where

$$M = \max_{K\in\tau_h} |m_K|. \tag{60}$$

Proof: Since $v \in \mathcal{A}_h$ we have $v = Fx$ on $\partial\Omega$, so for any m with $|m| = 1$ and $m \cdot n = 0$, applying the divergence theorem we get

$$\begin{aligned} 0 &= \textstyle\int_{\partial\Omega} |v(x) - Fx|^2 m\cdot x\,ds \\ &= \textstyle\int_\Omega \nabla(|v(x) - Fx|^2 m\cdot x)\cdot m\,dx \\ &= \textstyle\int_\Omega |v - Fx|^2\,dx + \int_\Omega (m\cdot x)\nabla|v - Fx|^2\cdot m\,dx. \end{aligned} \tag{61}$$

Now we can express

$$\begin{aligned} \textstyle\int_\Omega (m\cdot x)\nabla|v - Fx|^2\cdot m\,dx &= 2\textstyle\sum_{K\in\tau_h} m\cdot m_K \int_K (\nabla_h v - F)m\cdot(v - Fx)\,dx \\ &\quad +S(v), \end{aligned} \tag{62}$$

where $|S(v)|$ can be bounded. We will prove this later, but for now, assume that it is true and combining this with (61) we get

$$\begin{aligned} \textstyle\int_\Omega |v - Fx|^2\,dx &= -\textstyle\int_\Omega (m\cdot x)\nabla|v - Fx|^2\cdot m\,dx \\ &= -2\textstyle\sum_{K\in\tau_h}\{m\cdot m_K \int_K (\nabla_h v - F)m\cdot(v - Fx)\,dx\} - S(v) \\ &\le 2M\textstyle\int_\Omega |(\nabla_h v - F)m|\,|v - Fx|\,dx + |S(v)| \\ &\le 2M\left(\textstyle\int_\Omega |(\nabla_h v - F)m|^2\,dx\right)^{1/2}\left(\textstyle\int_\Omega |v - Fx|^2\,dx\right)^{1/2} + |S(v)| \\ &\le 2MC_2^{1/2}\mathcal{E}_h(v)^{1/4}\left(\textstyle\int_\Omega |v - Fx|^2\,dx\right)^{1/2} + |S(v)| \end{aligned} \tag{63}$$

where the last inequality comes from Theorem 2 (55).

Now returning to (62) we have $S(v) = \sum_{K\in\tau_h} S_K(v)$, where

$$\begin{aligned} S_K(v) &= \int_K (m\cdot x)\nabla|v-Fx|^2\cdot m\,dx - 2\int_K m\cdot m_K(\nabla_h v - F)m\cdot(v-Fx)\,dx \\ &= 2\int_K \left[(m\cdot x)(\nabla v - F) - (m\cdot m_K)(\nabla_h v - F)\right] m\cdot(v-Fx)\,dx. \end{aligned} \tag{64}$$

Now, in this expression, we expand $\nabla v - F$ using Lemma 1 (37) and we expand $m\cdot x$ and $v - Fx$ using (35). After multiplying the resulting expression out and using the fact that $\int_K (x-x_m)^p(y-y_m)^q\,dx\,dy = 0$ if p or q is odd, only 4 terms remain. They are

$$\begin{aligned} \tfrac{1}{2}S_K(v) &= \int_K m\cdot(x-m_K)(\nabla_h v - F)m\cdot(\nabla_h v - F)(x-m_K)\,dx \\ &\quad + \int_K m\cdot m_K D_h(v)(y-y_m, x-x_m)m\cdot(\nabla_h v - F)(x-m_K)\,dx \\ &\quad + \int_K m\cdot(x-m_K)D_h(v)(y-y_m,x-x_m)m\cdot A_k(v-Fx)\,dx \\ &\quad + \int_K m\cdot(x-m_K)D_h(v)(y-y_m,x-x_m)m\cdot D_h(v)(x-x_m)(y-y_m)\,dx. \end{aligned} \tag{65}$$

Evaluating each of these, using Lemma 2 (38), the estimates from Lemma 4 (50) and summing it all up, we get

$$\begin{aligned} \tfrac{1}{2}|S(v)| &\le C_2\tfrac{h^2}{6}\mathcal{E}_h(v)^{1/2} + M\tfrac{h}{6}(C_2C_3)^{1/2} + \tfrac{h}{6}|\Omega|^{1/2}\left(\int_\Omega |v-Fx|^2\,dx\right)^{1/2} + \tfrac{h^2}{36}|\Omega| \\ &\le \tfrac{1}{6}|\Omega|^{1/2}h\left(\int_\Omega |v-Fx|^2\,dx\right)^{1/2} + \mathcal{O}(h). \end{aligned} \tag{66}$$

Putting this result into (63) we get

$$\int_\Omega |v-Fx|^2\,dx \le \left[2MC_2^{1/2}\mathcal{E}_h(v)^{1/4} + \tfrac{1}{3}|\Omega|^{1/2}h\right]\left(\int_\Omega |v-Fx|^2\,dx\right)^{1/2} + \mathcal{O}(h). \tag{67}$$

From this we get that

$$\left(\int_\Omega |v-Fx|^2\,dx\right)^{1/2} \le 2MC_2^{1/2}\mathcal{E}_h(v)^{1/4} + \mathcal{O}(h^{1/2}) \tag{68}$$

and thus

$$\int_\Omega |v-Fx|^2\,dx \le 8M^2C_2\mathcal{E}_h(v)^{1/2} + \mathcal{O}(h). \qquad \square \tag{69}$$

With this estimate we can now estimate the difference between $\nabla_h v$ and F in any direction.

Theorem 4 *For any domain $\omega \subset \Omega$ which consists only of full elements of τ_h, there is a constant C_4 depending only on ω and other constants, such that*

$$\left|\int_\omega (\nabla_h v - F)\,dx\right| \le C_4\mathcal{E}_h(v)^{1/8} + \mathcal{O}(h^{1/4}), \tag{70}$$

for all $v \in \mathcal{A}_h$, with $\mathcal{E}_h(v) < 1$.

Proof: Let m be the unit outward normal on $\partial\omega$, then by Lemma 3 (45) and the divergence theorem, we have

$$\begin{aligned} |\textstyle\int_\omega(\nabla_h v - F)\,dx| &= |\textstyle\int_\omega(\nabla v - F)\,dx| \\ &= |\textstyle\int_{\partial\omega}(v(x) - Fx)\otimes m\,ds| \\ &\le \textstyle\int_{\partial\omega}|v(x) - Fx|\,ds \\ &\le |\partial\omega|^{1/2}\,(\textstyle\int_{\partial\omega}|v(x) - Fx|^2\,ds)^{1/2} \end{aligned} \tag{71}$$

with the last inequality coming from the Cauchy-Schwarz inequality. Next, applying the trace theorem [1], there is a constant C_T, depending only on ω, such that

$$\begin{aligned} \textstyle\int_{\partial\omega}|v(x) - Fx|^2\,ds &\le C_T\,[\textstyle\int_\omega|v(x) - Fx|^2\,dx + \int_\omega|\nabla(|v(x) - Fx|^2)|\,dx] \\ &\le C_T\textstyle\int_\omega|v - Fx|^2\,dx + 2C_T\int_\omega|\nabla v - F|\,|v - Fx|\,dx \\ &\le C_T\textstyle\int_\omega|v - Fx|^2\,dx + 2C_T\,(\int_\omega|v - Fx|^2\,dx)^{1/2}\,(\int_\Omega|\nabla v - F|^2\,dx)^{1/2}. \end{aligned} \tag{72}$$

Now using Lemma 4 (50) we have

$$\begin{aligned} (\textstyle\int_\Omega|\nabla v - F|^2\,dx)^{1/2} &\le (\textstyle\int_\Omega|\nabla_h v - F|^2\,dx)^{1/2} + (\int_\Omega|\nabla v - \nabla_h v|^2\,dx)^{1/2} \\ &\le C_3^{1/2} + C_2^{1/2}. \end{aligned} \tag{73}$$

Thus we get

$$\begin{aligned} |\textstyle\int_\omega(\nabla_h v - F)\,dx| &\le |\partial\omega|^{1/2}\,[C_T\textstyle\int_\omega|v - Fx|^2\,dx \\ &\quad +2C_T(C_3^{1/2} + C_2^{1/2})\,(\textstyle\int_\omega|v - Fx|^2\,dx)^{1/2}\Big]^{1/2}. \end{aligned} \tag{74}$$

Applying Theorem 3 (59), we get that

$$\left|\int_\omega(\nabla_h v - F)\,dx\right| \le C_4\mathcal{E}_h(v)^{1/8} + \mathcal{O}(h^{1/4}), \tag{75}$$

for some constant C_4 which depends on C_2, C_3, C_T and $|\partial\omega|$. □

Corollary 1 *With the same assumptions on v as in Theorem 4, we can extend the results of Theorem 4 to any smooth $\omega \subset \Omega$.*

Proof: Let $\omega \subset \Omega$ and let ω_h be the smallest subset of Ω which consists only of full elements of τ_h and contains ω. Then we can write

$$\left|\int_\omega(\nabla_h v - F)\,dx\right| \le \left|\int_{\omega_h}(\nabla_h v - F)\,dx\right| + \left|\int_{\omega_h-\omega}(\nabla_h v - F)\,dx\right|. \tag{76}$$

Then using the fact that $|\omega_h - \omega| \le |\partial\omega|\sqrt{2}h$ we have

$$\begin{aligned} \left|\textstyle\int_{\omega_h-\omega}(\nabla_h v - F)\,dx\right| &\le \textstyle\int_{\omega_h-\omega}|\nabla_h v - F|\,dx \\ &\le |\omega_h - \omega|^{1/2}\left(\textstyle\int_{\omega_h-\omega}|\nabla_h v - F|^2\,dx\right)^{1/2} \\ &\le |\partial\omega|^{1/2}\sqrt[4]{2}h^{1/2}C_3^{1/2}. \end{aligned} \tag{77}$$

Combining this result with the result from Theorem 4 (70), we get that

$$\left|\int_\omega (\nabla_h v - F)\,dx\right| \le C_4 \mathcal{E}_h(v)^{1/8} + \mathcal{O}(h^{1/4}). \qquad \square \tag{78}$$

Next, we show that although $\nabla_h v$ could take on any values anywhere in the wells $\mathcal{W}_0 \cup \mathcal{W}_1$ to attain low energy, that, if it matches the boundary condition, then it must take most values near W_0 and W_1.

Theorem 5 *For all $v \in \mathcal{A}_h$ with $\mathcal{E}_h(v) \le 1$, we have*

$$\int_\Omega |I - \theta(\nabla_h v)|^2 \le C_5 \mathcal{E}_h(v)^{1/2}, \tag{79}$$

where I is the 2×2 identity matrix, and

$$\int_\Omega |\nabla_h v - \Pi(\nabla_h v)|^2\,dx \le C_6 \mathcal{E}_h(v)^{1/2}, \tag{80}$$

where

$$\begin{aligned} C_5 &= 2(\nu^{-1} + C_2) \\ C_6 &= 2(\nu^{-1} + C_5 \max(|W_i|^2)). \end{aligned} \tag{81}$$

Proof: For any $m \in \mathrm{R}^2$ with $|m| = 1$ and $m \cdot n = 0$ and for any $A \in \mathcal{M}$ we have

$$\Pi(A)m = W_0 m = W_1 m = Fm. \tag{82}$$

Thus

$$\begin{aligned} (\theta(A) - I)Fm &= (\theta(A) - I)\Pi(A)m \\ &= \pi(A)m - \Pi(A)m \\ &= \pi(A)m - Fm \\ &= (\pi(A) - A)m + (A - F)m. \end{aligned} \tag{83}$$

Thus setting $A = \nabla_h v$, taking the norm, squaring the results and integrating over Ω we have

$$\begin{aligned} \textstyle\int_\Omega |(\theta(\nabla_h v) - I)Fm|^2\,dx &\le 2\textstyle\int_\Omega |\pi\nabla_h v - \nabla_h v|^2\,dx + 2\int_\Omega |(\nabla_h v - F)m|^2\,dx \\ &\le 2\nu^{-1}\mathcal{E}_h(v) + 2C_2\mathcal{E}_h(v)^{1/2} \\ &\le 2(\nu^{-1} + C_2)\mathcal{E}_h(v)^{1/2} \\ &= C_5\mathcal{E}_h(v)^{1/2}. \end{aligned} \tag{84}$$

where the last inequalities come from Lemma 4 (49) and Theorem 2 (55).

Next, since F is non-singular, $g = Fm \neq 0$ and thus we can rewrite this inequality, with

$$\theta(\nabla_h v) = \begin{pmatrix} c & -s \\ s & c \end{pmatrix}, \tag{85}$$

since $\theta(A)$ is a rotation, as

$$\int_\Omega [(c-1)^2 + s^2](g_1^2 + g_2^2) \le C_5 \mathcal{E}_h(v)^{1/2}. \tag{86}$$

Note also, for $g^\perp = (-g_2, g_1)$, we have $g^\perp \cdot g = 0$ and computing, we have

$$\int_\Omega |(\theta(\nabla_h v) - I)g^\perp|^2 \, dx = \int_\Omega [(c-1)^2 + s^2](g_1^2 + g_2^2) \le C_5 \mathcal{E}_h(v)^{1/2}. \tag{87}$$

Thus, since $(g, g^\perp)$ form a basis for R^2, we have that

$$\int_\Omega |\theta(\nabla_h v) - I|^2 \le C_5 \mathcal{E}_h(v)^{1/2}. \tag{88}$$

For the second estimate (80) we use the definition of the projections (2) to write

$$\begin{aligned} \nabla_h v - \Pi(\nabla_h v) &= (\nabla_h v - \pi(\nabla_h v)) + (\pi(\nabla_h v) - \Pi(\nabla_h v)) \\ &= (\nabla_h v - \pi(\nabla_h v)) + (\theta(\nabla_h v) - I)\Pi(\nabla_h v). \end{aligned} \tag{89}$$

Now integrating over Ω, and using (79) and Lemma 4 (49), we get

$$\begin{aligned} \textstyle\int_\Omega |\nabla_h v - \Pi(\nabla_h v)|^2 \, dx &\le 2\int_\Omega |\nabla_h v - \pi(\nabla_h v)|^2 \, dx + 2\int_\Omega |\theta(\nabla_h v) - I|^2 |\Pi(\nabla_h v)|^2 \, dx \\ &\le 2\nu^{-1}\mathcal{E}_h(v) + 2C_5 \max(|W_i|^2)\mathcal{E}_h(v)^{1/2} \\ &\le 2(\nu^{-1} + C_5 \max(|W_i|^2)\mathcal{E}_h(v)^{1/2} \\ &= C_6 \mathcal{E}_h(v)^{1/2}. \qquad \square \end{aligned} \tag{90}$$

For a more general and elegant proof of (79) see [30].

Now that we have shown that the gradients take on the values W_0 and W_1 primarily, we need to study the distribution of the gradients. For any $\omega \subset \Omega$, $\rho > 0$ and $v \in \mathcal{A}_h$, we define the sets for $i = 0, 1$:

$$\omega_\rho^i = \omega_\rho^i(v) = \{x \in \omega : \Pi(\nabla_h v(x)) = W_i \text{ and } |\Pi \nabla_h v - \nabla_h v| < \rho\}, \tag{91}$$

and the set $\omega_\rho^e = \omega - \omega_\rho^0 - \omega_\rho^1$. From these sets, we can define the approximate probabilities:

$$\mu_h^i = \frac{|\omega_\rho^i|}{|\omega|} \text{ for } i = 0, 1 \text{ and } \mu_h^e = \frac{|\omega_\rho^e|}{|\omega|}. \tag{92}$$

So that μ_h^i is the probability that $\nabla_h v$ is within ρ of W_i and μ_h^e is the probability that $\nabla_h v$ is not with ρ of either W_0 or W_1.

Theorem 6 *For any $\omega \subset \Omega$, $v \in \mathcal{A}_h$ with $\mathcal{E}_h(v) \le 1$ and the sets and probabilities defined as above, we have*

$$|\mu_h^1 - \lambda| \le C_7 \mathcal{E}_h(v)^{1/8} + \mathcal{O}(h^{1/4}), \tag{93}$$

where

$$C_7 = \frac{1}{|a|}\left(|\omega|^{1/2} C_6^{1/2} + C_4 + \frac{2\max(|W_i|)}{\nu\rho^2|\omega|}\right) \tag{94}$$

and

$$\mu_h^e \le \frac{C_6}{\rho^2|\omega|}\mathcal{E}_h(v)^{1/2}. \tag{95}$$

Proof: First, we can express w_ρ^e as

$$\omega_\rho^e = \{x \in \omega : |\Pi\nabla_h v - \nabla_h v| \geq \rho\}, \tag{96}$$

and thus

$$\begin{aligned} |\omega_\rho^e| &= \textstyle\int_{\omega_\rho^e} 1\, dx \\ &\leq \tfrac{1}{\rho}\textstyle\int_{\omega_\rho^e} |\Pi\nabla_h v - \nabla_h v|\, dx \\ &\leq \tfrac{1}{\rho}|\omega_\rho^e|^{1/2}\left(\textstyle\int_\Omega |\Pi\nabla_h v - \nabla_h v|^2\, dx\right)^{1/2} \\ &\leq \tfrac{|\omega_\rho^e|^{1/2}}{\rho} C_6^{1/2}\mathcal{E}_h(v)^{1/4}, \end{aligned} \tag{97}$$

using Theorem 5 (80). So

$$|\omega_\rho^e| \leq \frac{C_6}{\rho^2}\mathcal{E}_h(v)^{1/2}, \tag{98}$$

and (95) follows immediately.

Next, from the fact that

$$\omega = \omega_\rho^0 \cup \omega_\rho^1 \cup \omega_\rho^e \tag{99}$$

we have that

$$\mu_h^0 + \mu_h^1 + \mu_h^e = 1 \tag{100}$$

and that μ_h^0, μ_h^1 and $\mu_h^e \geq 0$. Using the properties of the projection Π (2), we get

$$⨍_\omega \Pi\nabla_h v - F\, dx = \mu_h^0 W_0 + \mu_h^1 W_1 + \frac{1}{|\omega|}\int_{\omega_\rho^e} \Pi\nabla_h v\, dx - F. \tag{101}$$

Next, since $F = W_0 + \lambda a \otimes n$ and $W_1 = W_0 + a \otimes n$, we can rewrite this last equality as

$$\begin{aligned} ⨍_\omega \Pi\nabla_h v - F\, dx - \tfrac{1}{|\omega|}\textstyle\int_{\omega_\rho^e} \Pi\nabla_h v\, dx &= \mu_h^0 W_0 + \mu_h^1 W_1 - F \\ &= (\mu_h^0 + \mu_h^1 - 1)W_0 + (\mu_h^1 - \lambda)a \otimes n \\ &= -\mu_h^e W_0 + (\mu_h^1 - \lambda)a \otimes n. \end{aligned} \tag{102}$$

Next, we estimate the two integrals on the left. First, via Cauchy-Schwarz, Theorem 5 (80) and Corollary 1 (70) we get that

$$\begin{aligned} \left|⨍_\omega \Pi\nabla_h v - F\, dx\right| &\leq \left|⨍_\omega \Pi(\nabla_h v) - \nabla_h v\, dx\right| + \left|⨍_\omega \nabla_h v - F\, dx\right| \\ &\leq |\omega|^{1/2}\left(\textstyle\int_\omega |\Pi(\nabla_h v) - \nabla_h v|^2\, dx\right)^{1/2} + \left|⨍_\omega \nabla_h v - F\, dx\right| \\ &\leq |\omega|^{1/2} C_6^{1/2}\mathcal{E}_h(v)^{1/4} + C_4\mathcal{E}_h(v)^{1/8} + \mathcal{O}(h^{1/4}) \\ &\leq (|\omega|^{1/2} C_6^{1/2} + C_4)\mathcal{E}_h(v)^{1/8} + \mathcal{O}(h^{1/4}). \end{aligned} \tag{103}$$

Second, we have that

$$\begin{aligned} \tfrac{1}{|\omega|}\left|\textstyle\int_{\omega_\rho^e} \Pi\nabla_h v\, dx\right| &\leq \tfrac{1}{|\omega|}\max(|W_i|)|\omega_\rho^e| \\ &\leq \max(|W_i|)\mu_h^e. \end{aligned} \tag{104}$$

Combining these last two results with (102) gives

$$\begin{aligned} |\mu_h^1 - \lambda|\,|a| &\leq \left|\int_\omega \Pi\nabla_h v - F\,dx\right| + \mu_h^e|W_0| + \tfrac{1}{|\omega|}\left|\int_{\omega_\rho^e} \Pi\nabla_h v\,dx\right| \\ &\leq (|\omega|^{1/2}C_6^{1/2} + C_4)\mathcal{E}_h(v)^{1/8} + \mathcal{O}(h^{1/4}) + 2\max(|W_i|)\mu_h^e \\ &\leq |a|C_7\mathcal{E}_h(v)^{1/8} + \mathcal{O}(h^{1/4}). \qquad \square \end{aligned} \tag{105}$$

Finally we estimate how well the properties of $\nabla_h v$ match with the Young measure associated with the problem.

Theorem 7 *For any measurable function $f : \Omega \times \mathcal{M} \to R$ and $v \in \mathcal{A}_h$ we have the approximation error given by*

$$err(f,v) = \int_\Omega (f(x,\nabla_h v(x)) - (1-\lambda)f(x,W_0) - \lambda f(x,W_1))\,dx. \tag{106}$$

If $\mathcal{E}_h(v) \leq 1$ then

$$|err(f,v)| \leq C_8\mathcal{E}_h(v)^{1/4} + \mathcal{O}(h^{1/2}), \tag{107}$$

where

$$C_8 = C_6^{1/2}\left\|\frac{\partial f}{\partial A}\right\|_{L^2(\Omega)\times L^\infty(\mathcal{M})} + 2MC_2^{1/2}|a|^{-1}\|\nabla G\|_{L^\infty(\Omega)} + C_6^{1/2}|a|^{-1}\|G\|_{L^2(\Omega)}. \tag{108}$$

and $G(x) = f(x,W_0) - f(x,W_1)$.

Proof: This proof proceeds by dividing the expression up into parts which we can or have estimated:

$$\begin{aligned} \mathrm{err}(f,v) &= \int_\Omega f(x,\nabla_h v(x)) - f(x,\Pi(\nabla_h v(x)))\,dx \\ &\quad + \int_\Omega f(x,\Pi(\nabla_h v(x))) - (1-\lambda)f(x,W_0) - \lambda f(x,W_1)\,dx \\ &= I_1 + I_2. \end{aligned} \tag{109}$$

Using Theorem 5 (80), we get

$$\begin{aligned} |I_1| &\leq \int_\Omega |f(x,\nabla_h v(x)) - f(x,\Pi(\nabla_h v(x)))|\,dx \\ &\leq \int_\Omega \left\|\tfrac{\partial f}{\partial A}(x,\cdot)\right\|_{L^\infty(\mathcal{M})} |\nabla_h v - \Pi(\nabla_h v)|\,dx \\ &\leq \left\|\tfrac{\partial f}{\partial A}\right\|_{L^2(\Omega)\times L^\infty(\mathcal{M})} \left(\int_\Omega |\nabla_h v - \Pi(\nabla_h v)|^2\,dx\right)^{1/2} \\ &\leq C_6^{1/2}\left\|\tfrac{\partial f}{\partial A}\right\|_{L^2(\Omega)\times L^\infty(\mathcal{M})} \mathcal{E}_h(v)^{1/4}. \end{aligned} \tag{110}$$

Next, for I_2, we let $G(x) = f(x,W_0) - f(x,W_1)$ and note that when $\Pi(A) = W_0$ we have

$$\begin{aligned} f(x,\Pi(A)) - (1-\lambda)f(x,W_0) - \lambda f(x,W_1)\,dx &= \lambda f(x,W_0) - \lambda f(x,W_1) \\ &= \lambda G(x), \end{aligned} \tag{111}$$

and when $\Pi(A) = W_1$ we have

$$\begin{aligned} f(x,\Pi(A)) - (1-\lambda)f(x,W_0) - \lambda f(x,W_1))\,dx &= (\lambda-1)f(x,W_0) - (\lambda-1)f(x,W_1) \\ &= (\lambda-1)G(x). \end{aligned} \tag{112}$$

Since $F = W_0 + \lambda a \otimes n$, we have that $\lambda a \otimes n = F - W_0$ and $(\lambda-1)a\otimes n = F - W_1$. Thus for all A,

$$[f(x,\Pi(A)) - (1-\lambda)f(x,W_0) - \lambda f(x,W_1)]\, a\otimes n = G(x)(F - \Pi(A)). \tag{113}$$

So

$$\begin{aligned} I_2 a\otimes n &= \textstyle\int_\Omega f(x,\Pi(\nabla_h v(x))) - (1-\lambda)f(x,W_0) - \lambda f(x,W_1))\,dx(a\otimes n) \\ &= \textstyle\int_\Omega G(x)(F - \Pi(\nabla_h v(x)))\,dx \\ &= \textstyle\int_\Omega G(x)\nabla(Fx - v(x))\,dx + \int_\Omega G(x)(\nabla v(x) - \nabla_h v(x))\,dx \\ &\quad \textstyle + \int_\Omega G(x)(\nabla_h v(x) - \Pi(\nabla_h v(x)))\,dx \\ &= I_3 + I_4 + I_5. \end{aligned} \tag{114}$$

To estimate I_3 we use integration by parts and the divergence theorem to write

$$\begin{aligned} I_3 &= \textstyle\int_\Omega G(x)\nabla(Fx - v(x))\,dx \\ &= \textstyle\int_\Omega \nabla G(x)\otimes(Fx - v(x))\,dx, \end{aligned} \tag{115}$$

noting that $v = Fx$ on $\partial\Omega$. Thus using Cauchy-Schwarz and Theorem 3 (59) we get

$$\begin{aligned} |I_3| &\le \textstyle\int_\Omega |\nabla G(x)|\,|Fx - v(x)|\,dx \\ &\le \|\nabla G\|_{L^2(\Omega)} \left(\textstyle\int_\Omega |v(x) - Fx|^2\,dx\right)^{1/2} \\ &\le \|\nabla G\|_{L^2(\Omega)} \left(2MC_2^{1/2}\mathcal{E}_h(v)^{1/4} + \mathcal{O}(h^{1/2})\right). \end{aligned} \tag{116}$$

Next, using Lemma 3 (46) we have that

$$\begin{aligned} I_4 &= \textstyle\sum_{K\in\tau_h}\int_K G(x)(\nabla v - \nabla_h v)\,dx \\ &= \textstyle\sum_{K\in\tau_h}\int_K (G(x) - G(m_K))(\nabla v - \nabla_h v)\,dx. \end{aligned} \tag{117}$$

Thus applying Lemma 2 (38), Cauchy-Schwarz and Lemma 4 (50) we get

$$\begin{aligned} |I_4| &\le \textstyle\sum_{K\in\tau_h}\int_K |G(x) - G(m_K)|\,|\nabla v - \nabla_h v|\,dx \\ &\le \|\nabla G\|_{L^\infty(\Omega)} \textstyle\sum_{K\in\tau_h}\int_K |D_h(v)|\,|x - m_K|^2\,dx \\ &\le \|\nabla G\|_{L^\infty(\Omega)} \tfrac{h^2}{3}\textstyle\int_\Omega |\nabla_h v|\,dx \\ &\le \|\nabla G\|_{L^\infty(\Omega)} \tfrac{h^2}{3}|\Omega|^{1/2}C_2^{1/2}. \end{aligned} \tag{118}$$

Finally, using Theorem 5 (80) and Cauchy-Schwarz, we have that

$$\begin{aligned} |I_5| &\le \textstyle\int_\Omega |G(x)|\,|\nabla_h v - \Pi\nabla_h v|\,dx \\ &\le \|G\|_{L^2(\Omega)} \left(\textstyle\int_\Omega |\nabla_h v - \Pi\nabla_h v|^2\,dx\right)^{1/2} \\ &\le C_6^{1/2}\|G\|_{L^2(\Omega)}\mathcal{E}_h(v)^{1/4}. \end{aligned} \tag{119}$$

Combining the results for $|I_1|$, $|I_3|$, $|I_4|$ and $|I_5|$ and that fact that $|a \otimes n| = |a|$, we get

$$\begin{aligned} |\mathrm{err}(f,v)| &\leq C_6^{1/2} \left\| \tfrac{\partial f}{\partial A} \right\|_{L^2(\Omega)\times L^\infty(\mathcal{M})} \mathcal{E}_h(v)^{1/4} + |a|^{-1} \|\nabla G\|_{L^2(\Omega)} \left(2MC_2^{1/2} \mathcal{E}_h(v)^{1/4} + \mathcal{O}(h^{1/2}) \right) \\ &\quad + |a|^{-1} \|\nabla G\|_{L^\infty(\Omega)} \tfrac{h^2}{3} |\Omega|^{1/2} C_2^{1/2} + |a|^{-1} C_6^{1/2} \|G\|_{L^2(\Omega)} \mathcal{E}_h(v)^{1/4} \\ &\leq C_8 \mathcal{E}_h(v)^{1/4} + \mathcal{O}(h^{1/2}). \qquad \square \end{aligned} \tag{120}$$

The bounds of Theorems 1–7 agree with the bounds determined by Luskin [30] for conforming elements, up to the power of $\mathcal{E}_h(v)$. The only difference is in the dependence on the mesh size h, which is typical when using a quadrature scheme.

5 Computational Results

In this section we describe some computational experiments and comment on the optimality of the error estimates from the last section. We consider the specific case when $W_1 - W_0 = a \otimes n$ with $n = (1, 0)$, as this case has the highest potential to produce spurious oscillations. This combination of element and quadrature rule was chosen to produce extra degrees of freedom so that the twinned configurations would not be as dependent on the triangulation orientation as in the conforming cases [12]. When $n = (1, 0)$, the extra degrees of freedom are not needed to form the microstructure and thus are free to cause spurious oscillations.

First we describe the wells and the energy density. Let

$$W_0 = \begin{pmatrix} 1 & 0 \\ -\epsilon & 1 \end{pmatrix} \text{ and } W_1 = \begin{pmatrix} 1 & 0 \\ \epsilon & 1 \end{pmatrix} \tag{121}$$

then

$$W_1 - W_0 = (0, 2\epsilon) \otimes (1, 0) = a \otimes n. \tag{122}$$

For the energy density, let $\phi(F) = \Phi(F^T F) = \Phi(C)$, where

$$\Phi(C) = \kappa_1 (C_{11} - (1 + \epsilon^2))^2 + \kappa_2 (C_{22} - 1)^2 + \kappa_3 (C_{12}^2 - \epsilon^2)^2. \tag{123}$$

This energy satisfies all the basic properties. For this trial, we take the reference domain to $\Omega = [0, 1]^2$ and we choose $\epsilon = 0.2$, $\lambda = 0.5$ and $\kappa_1 = \kappa_2 = \kappa_3 = 1$.

For this case, we will look at two types of results. First we will look at how the errors depend on the mesh size h, by computing a near optimal solution for various h's in the range from $\frac{1}{8}$ to $\frac{1}{128}$. Second, we will fix $h = \frac{1}{128}$ and look at how the errors depend on the energy $\mathcal{E}_h$ as the iterative solution process goes from the initial guess to the final, near optimal, answer. The algorithm we use to compute the minimizers is a variation of the conjugate gradient method, see Collins [11] for more details. For the initial guess in each case we take a uniformly twinned approximation comparable to the function created in the proof of Theorem 1. For each deformation, $v \in \mathcal{A}_h$, which we test, we compute 4 values (V_1-V_4):

$V_1 = E_h(v)$ (the energy).

$V_2 = \int_\Omega |(\nabla_h v - F)m|^2 \, dx$ with $m = (0, 1)$ (from Theorem 2).

$V_3 = \int_\Omega |v(x) - Fx|^2 \, dx$ (from Theorem 3).

$V_4 = \left| \int_\Omega \sum_{ij} (\nabla_h v)_{ij}^2 \, dx - (2 + \epsilon^2) \right|$ (from Theorem 7, with $f(x, A) = \sum_{ij} A_{ij}^2$).

The first results, shown in Figure 3, show the values of V_1-V_4 plotted against various values of h. The first thing to note is that all the values appear to be converging towards zero. Also, each is converging at a rate higher than predicted. The energy $\mathcal{E}_h(u_h)$ behaves closely to $\mathcal{O}(h^{7/4})$. This is much better than the estimate of Theorem 1, but this is expected as there is no adjustment needed to form the twin bands and so the width of the bands can be as small as h. The adjustments to match the boundary condition are still needed, however, the results are still better than this would predict. This exceptional convergence is due to the elements flexibility to adjust to match the boundary conditions using very low energy configurations. The convergence of V_2 and V_4 is similar to that for V_1 and is significantly better than the $\mathcal{E}_h(v)^{1/2}$ and $\mathcal{E}_h(v)^{1/8}$ predicted. V_3 converges at a rate close to $\mathcal{O}(h^2)$, indicating that Theorem 3 also underestimates the convergence rate in this case.

The second set of results, shown in Figures 4, 5 and 6, show V_2, V_3 and V_4 plotted against V_1 for various approximations with $h = \frac{1}{128}$. In this case we also see that V_2 and V_4 have near linear relationships with V_1, indicating that the estimates in Theorems 2 and 7 should depend linearly on $\mathcal{E}_h(v)$. V_3 has a more erratic relationship, but also indicates that the optimal bound should depend on a larger power of $\mathcal{E}_h(v)$ than the $1/2$ of Theorem 3. These results show that in this special case that we have convergence, with better than predicted rates. In more general cases, we expect that the estimates of the previous section will hold more closely.

The estimates of the previous section and these computational results show that the threat of spurious oscillations to the quality of the computed microstructures is non-existent and thus this combination of the bilinear element and one-point quadrature can be used confidently for computing microstructures.

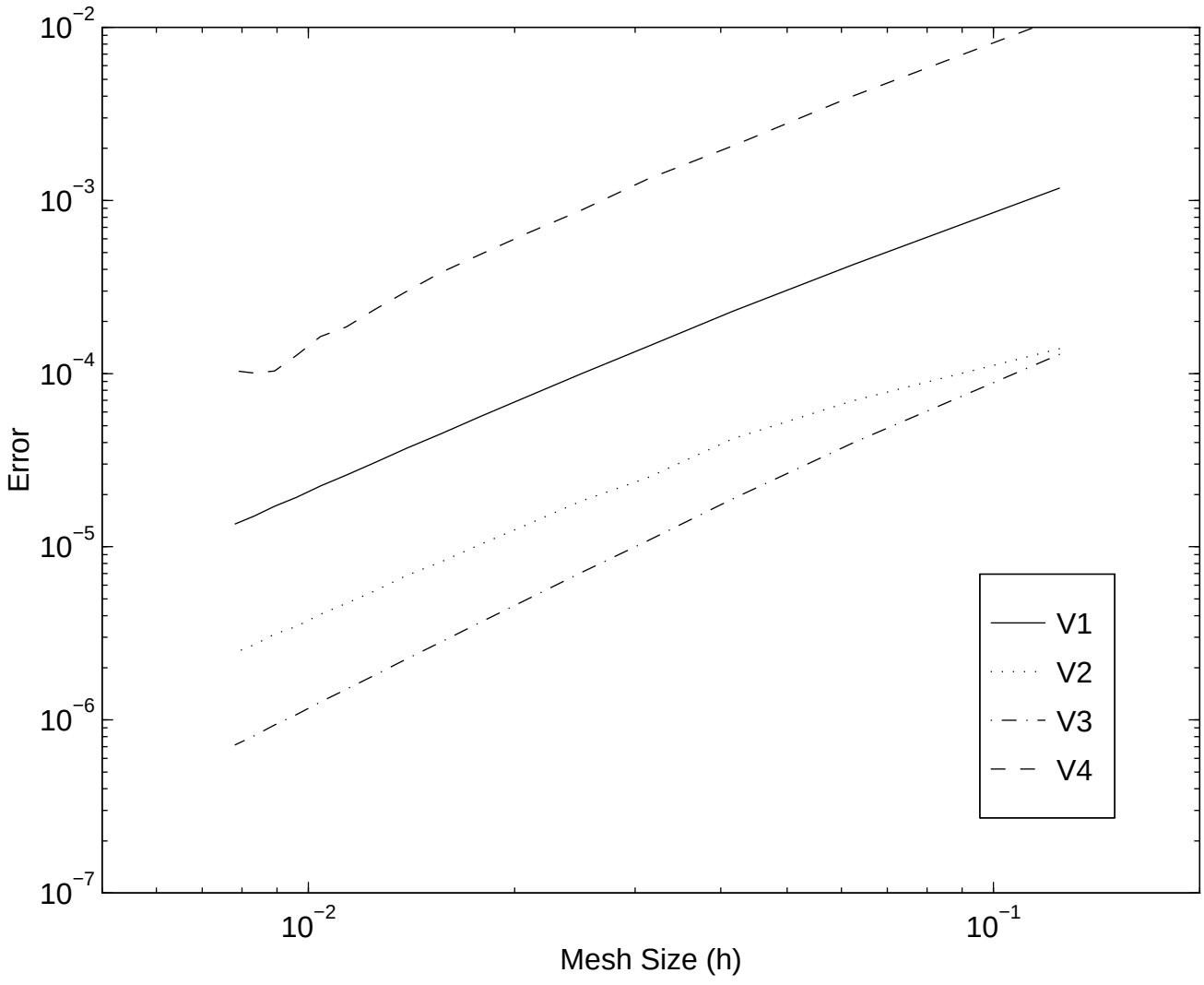

Figure 3: Errors for special case, plotted against mesh size

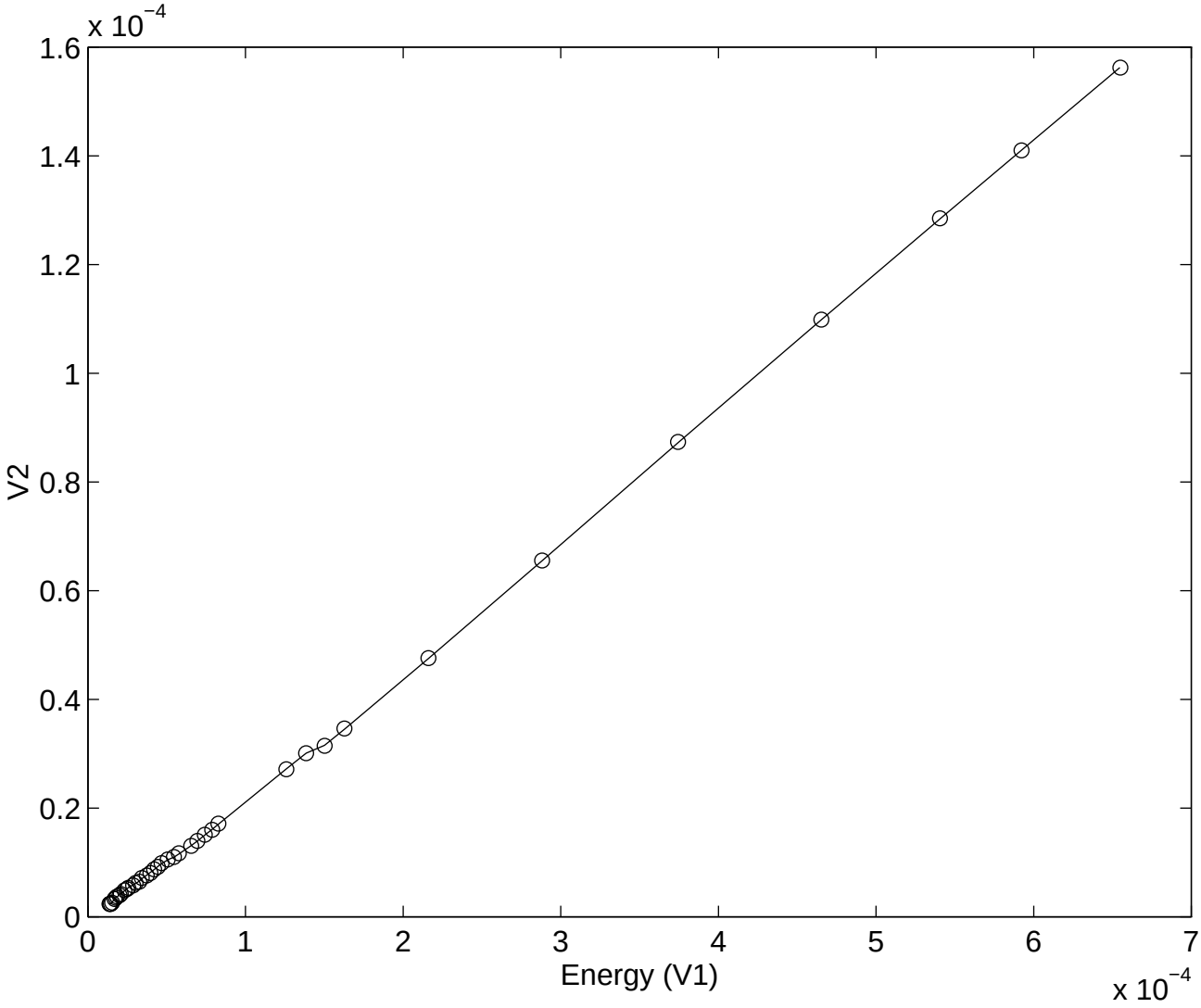

Figure 4: V_2 (Theorem 2 Error) plotted against V_1 (Energy) with $h = \frac{1}{128}$

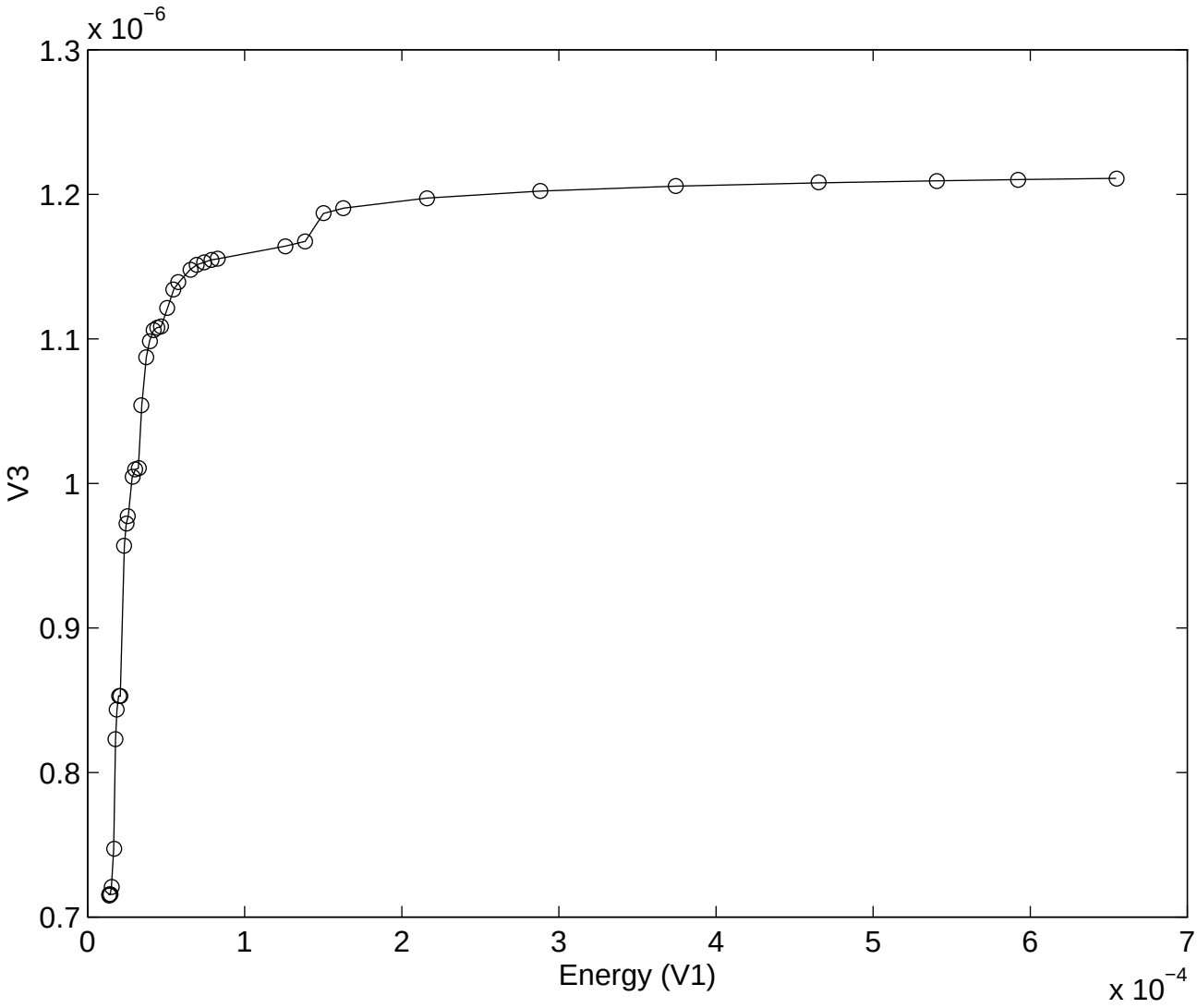

Figure 5: V_3 (Theorem 3 Error) plotted against V_1 (Energy) with $h = \frac{1}{128}$

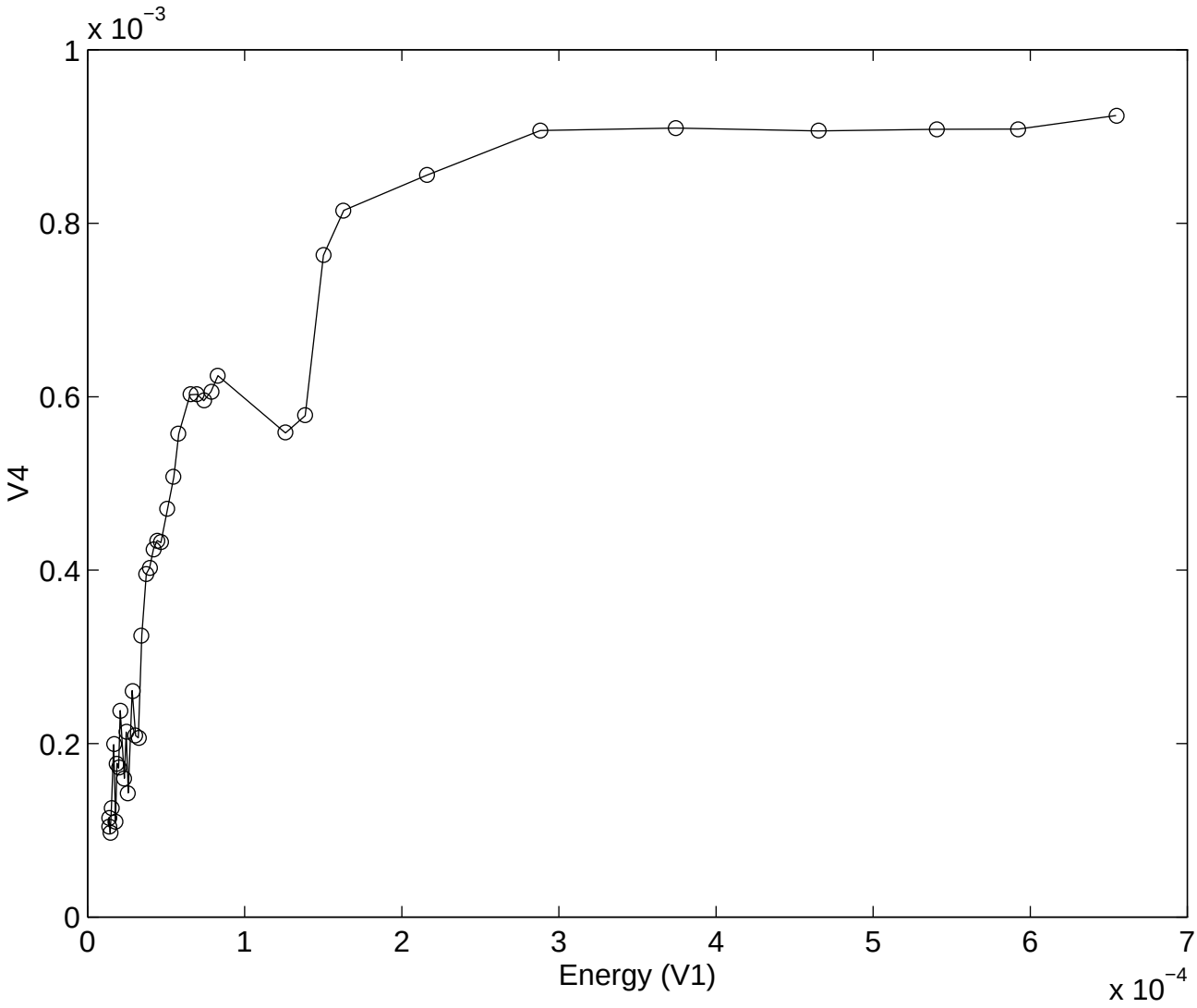

Figure 6: V_4 (Theorem 7 Error) plotted against V_1 (Energy) with $h = \frac{1}{128}$

References

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, 1975.

[2] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Rat. Mech. Anal., 100 (1987), pp. 13–52.

[3] ——, *Proposed experimental tests of a theory of fine microstructure and the two well problem*, Phil. Trans. Roy. Soc. Lond., 338 (1992), pp. 389–450.

[4] J. M. Boland and R. A. Nicolaides, *Stable and semistable low order finite elements for viscous flows*, SIAM J. Numer. Anal., 22 (1985), pp. 474–492.

[5] M. Chipot, *Numerical analysis of oscillations in nonconvex problems*, Numer. Math., 59 (1991), pp. 747–767.

[6] M. Chipot and C. Collins, *Numerical approximations in variational problems with potential wells*, SIAM J. Numer. Anal., 29 (1992), pp. 1002–1019.

[7] M. Chipot, C. Collins, and D. Kinderlehrer, *Numerical analysis of oscillations in multiple well problems*, Numerisch Mathematik, 70 (1995), pp. 259–282.

[8] M. Chipot and D. Kinderlehrer, *Equilibrium configurations of crystals*, Arch. Rat. Mech. Anal., 103 (1988), pp. 237–277.

[9] P. G. Ciarlet, *Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1980.

[10] C. Collins, *Computation and Analysis of Twinning in Crystalline Materials*, PhD thesis, University of Minnesota, 1990.

[11] ——, *Computation of twinning*, in Kinderlehrer et al. [25], pp. 39–50.

[12] ——, *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.

[13] C. Collins, D. Kinderlehrer, and M. Luskin, *Numerical approximation of the solution of a variational problem with a double well potential*, SIAM J. Numer. Anal., 28 (1991), pp. 321–332.

[14] C. Collins and M. Luskin, *The computation of the austenitic-martensitic phase transition*, in Partial Differential Equations and Continuum Models of Phase Transitions, M. Rascle, D. Serre, and M. Slemrod, eds., vol. 344 of Lecture Notes in Physics, Springer-Verlag, 1989, pp. 34–50.

[15] ——, *Computational results for phase transitions in shape memory materials*, in Rogers [36], pp. 198–215.

[16] ——, *Numerical modeling of the microstructure of crystals with symmetry-related variants*, in Proceedings of the ARO US-Japan Workshop on Smart/Intelligent Materials and Systems, Lancaster, Pennsylvania, 1990, Technomic Publishing Co., pp. 309–318.

[17] ——, *Optimal order error estimates for the finite element approximation of the solution of a nonconvex variational problem*, Math. Comp., 57 (1991), pp. 621–637.

[18] ——, *Computational results for martensitic twinning*, in Proceedings of International Conference on Martensitic Transformations (ICOMAT-92), 1992.

[19] C. Collins, M. Luskin, and J. Riordan, *Computational images of crystalline microstructure*, in Computing Optimal Geometries, J. E. Taylor, ed., AMS Video, 1991.

[20] ——, *Computational results for a two-dimensional model of crystalline microstructure*, in Kinderlehrer et al. [25], pp. 51–56.

[21] J. L. Ericksen, *Constitutive theory for some constrained elastic crystals*, Int. J. Solids and Structures, 22 (1986), pp. 951–964.

[22] ——, *Stable equilibrium configurations of elastic crystals*, Arch. Rat. Mech. Anal., 94 (1986), pp. 1–14.

[23] P. A. Gremaud, *Numerical analysis of a nonconvex variational problem related to solid-solid phase transitions*, SIAM J. Numer. Anal., 31 (1994), pp. 111–127.

[24] D. Kinderlehrer, *Twinning in crystals, II*, in Metastability and Incompletely Posed Problems, S. Antman, J. L. Ericksen, D. Kinderleher, and I. Muller, eds., IMA Vol Math Appl 3, Springer-Verlag, 1987, pp. 185–211.

[25] D. Kinderlehrer, R. James, M. Luskin, and J. L. Ericksen, eds., *Microstructure and Phase Transition*, IMA Volumes in Mathematics and its Applications, Springer-Verlag, 1993.

[26] D. Kinderlehrer, R. A. Nicolaides, and H. Wang, *Spurious oscillations in computing microstructures*, in Smart Structures and Materials 1993: Mathematics in Smart Structures, H. T. Banks, ed., Proc. SPIE 1919, 1993, pp. 38–46.

[27] D. Kinderlehrer and P. Pedregal, *Characterizations of Young measures generated by gradients*, Arch. Rat. Mech. Anal., (1991), pp. 329–365.

[28] B. Li and M. Luskin, *Finite element analysis of microstructure fo the cubic to tetragonal transformation*, Preprint 1373, IMA, 1996.

[29] ——, *Nonconforming finite element approximation of crystalline microstructure.* Manuscript, 1996.

[30] M. Luskin, *Approximation of a laminated microstructure for a rotationally invariant, double well energy density*, Numer. Math., (1996).
(to appear).

[31] ——, *On the computation of crystalline microstructure*, Acta Numerica, 5 (1996), pp. 191–257.

[32] National Research Council, *Mathematical research in materials science: opportunities and perspectives*, National Academy Press, 1993.

[33] R. A. Nicolaides, *Nodal minimizers and other false solutions in computational microstructures*, in Proceedings of AFOSR Workshop, May 1993, pp. 44–46.

[34] M. Pitteri, *Reconciliation of local and global symmetries of crystals*, J. Elasticity, 14 (1984), pp. 175–190.

[35] ——, *On the kinematics of mechanical twinning in crystals*, Arch. Rat. Mech. Anal., 88 (1985), pp. 25–57.

[36] C. Rogers, ed., *Smart Materials, Structures and Mathematical Issues*, Technomic Publishing Co., Lancaster, PA, 1989.

[37] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, Chelsea, 1980.