Viewing the Finite Element Residual as a Lagrange Multiplier for Discretization

Wolfgang Bangerth
Institute for Computational Engineering and Sciences
University of Texas at Austin, Austin, TX 78712, USA

December 11, 2004

Abstract

Instead of writing the Galerkin approximation as seeking the stationary point of some functional over a finite-dimensional subspace of the original function space, we write it as the initial, infinite-dimensional variational problem with an explicit constraint that models the fact that we are actually searching in a subspace. By formulating this as a constrained variational problem using a Lagrangian functional, we are led to introduce a Lagrange multiplier for the discreteness constraint. This multiplier turns out to be the residual of the approximation, shedding some light on the basic interpretation of the residual in Galerkin methods.

Since Lagrange multipliers indicate the first order response of a functional to perturbations in the constraint, we consider applications of this relationship to mesh refinement strategies and error estimation for finite element methods. After considering an introductory example for the Laplace equation, the fully nonlinear case is treated, where the accurate computation of some functional of the solution is required.

1 Introduction

Finite element methods, and Galerkin approximations in general, seek to approximate the solution of an infinite-dimensional variational problem by searching only in a finite-dimensional subspace. For finite element methods, this subspace is usually spanned by a nodal basis defined on a grid. While this approach is easy to understand, it is not the only possible formulation of the problem. In this paper, we rewrite it slightly by using the original variation over the infinite-dimensional space, and add the desired discreteness, or finite-dimensionality, of the solution as an explicit constraint. We will then try to interpret the findings we get from this formulation.

It turns out that the Lagrange multiplier associated with the discreteness constraint is actually the residual we get if we enter the approximate solution into the differential equation that is associated with the variational problem. We

will generalize this surprising result to the case of general nonlinear problems where the goal is the accurate computation of some output functional of the solution. In that case, the Lagrange multiplier is the residual of a dual problem. Using this equivalence, we invoke a result from optimization theory that links the Lagrange multiplier to the first-order change in the target functional upon perturbations of the constraint. This suggests methods for "optimal" mesh refinement that we will explore, and we will also briefly touch connections with error estimation.

The formulas appearing in this paper are mostly well-known. However, they are usually derived in different ways, and quantities like the residual, or a dual solution, are usually associated with different notions than in this paper. The purpose of this work is therefore not to produce new results, but only to put them into a different context that allows different interpretations, and to explore the connection between the Lagrange multiplier, the residual, and some of the consequences that are already known for variational problems with constraints.

The layout of the rest of this paper is as follows: In the next section, we describe how Galerkin approximation of the solution of the Poisson equation can be considered as a constraint. Section 3 generalizes this to arbitrary nonlinear problems and the goal oriented case, and we show what consequences perturbations of the constraint have on the target functional applied to the solution. In section 4, we consider application of this to generating "optimal" refinement strategies. Section 5 is then devoted to error estimation.

2 Discretization as a Constraint

In this first section, we would like to lay out the basic idea of this paper using a simple model problem, the Laplace equation. For this, let us consider the following problem: minimize the energy functional $E(v) = \frac{1}{2}\|\nabla v\|^2 - (f, v)$,

$$\min_{u \in V} E(u), \tag{1}$$

where $V = H_0^1(\Omega)$, $(u, v) = \int_\Omega uv\, dx$, and $\|\cdot\|^2 = (\cdot, \cdot)$. For the Ritz projection, we seek an approximation u_h to the solution u of this continuous problem by restricting the search space to a finite dimensional subspace $V_h \subset V$:

$$\min_{u_h \in V_h} E(u_h). \tag{2}$$

For example, in finite elements, the space V_h is chosen as the one spanned by a nodal basis associated with a triangulation of the domain Ω.

We now rewrite the finite dimensional minimization problem (2) in a different form. For this, let us denote by $\{\varphi_h^i\}_{i=1}^N$ a basis of V_h, made up of linearly independent, but not necessarily orthogonal basis functions φ_h^i. Then we can state problem (2) in the following form:

$$\min_{\tilde{u} \in V,\ \alpha \in \mathbb{R}^N} E(\tilde{u}), \qquad \text{subject to} \quad \tilde{u} - \sum_{i=1}^{N} \alpha^i \varphi_h^i = 0. \tag{3}$$

Note that we extend the minimization over the whole space V, and use an explicit constraint for the requirement that the discrete solution should actually belong to the subspace V_h. Discretization is thus considered as a linear constraint to an infinite-dimensional minimization problem.

We will consider the solution of (3) in a Lagrangian framework. Testing the constraint in the H_0^1-H^{-1} duality product, we are then led to introduce the following Lagrangian:

$$L(\tilde{u}, \alpha, \lambda) = \tfrac{1}{2}\|\nabla\tilde{u}\|^2 - (f, \tilde{u}) + \Big(\tilde{u} - \sum \alpha^i \varphi_h^i, \lambda\Big),$$

where $\lambda \in H^{-1}$ is a Lagrange multiplier.

Proposition 1. *The solution* $\{\tilde{u}, \alpha, \lambda\}$ *of Problem* (3) *satisfies* $\tilde{u} = u_h = \sum \alpha^i \varphi_h^i$ *and* $\lambda = f + \Delta u_h$.

Proof. The solution of (3) is characterized by stationarity of L:

$$L'_{\tilde{u}}(\tilde{u}, \alpha, \lambda)(\varphi) = (\nabla\tilde{u}, \nabla\varphi) - (f, \varphi) + (\lambda, \varphi) = 0 \qquad \forall \varphi \in V, \tag{4}$$

$$L'_{\alpha^i}(\tilde{u}, \alpha, \lambda) = -(\varphi_h^i, \lambda) = 0 \qquad \forall 1 \le i \le N, \tag{5}$$

$$L'_{\lambda}(\tilde{u}, \alpha, \lambda)(\psi) = \Big(\tilde{u} - \sum_i \alpha^i \varphi_h^i, \psi\Big) = 0 \qquad \forall \psi \in H^{-1}, \tag{6}$$

where L' denotes differentiation of L with respect to the subscripted variable. Considering (4) for discrete test functions $\varphi_h \in V_h$ and using (5) to delete the last term, we see that indeed $\tilde{u}$ satisfies the same equation that u_h satisfies. Together with the condition of discreteness, (6), it follows that $\tilde{u} = u_h$, proving the first claim of the proposition.

Next, considering (4) for arbitrary test functions $\varphi \in V$ and integrating the first term by parts, it follows that the Lagrange multiplier for the discretization constraint is actually the residual: $\lambda = f + \Delta\tilde{u} = f + \Delta u_h$. (In the following, for finite elements, we will denote by the term *residual* both the cell and the edge residuals. λ is thus in H^{-1}, or if associated with a function representation, contains also Dirac measures on cell edges.) Then, (5) states that the residual is orthogonal to the subspace V_h in which we look for our solution; this property is commonly refered to as *Galerkin orthogonality*. □

Alternatively, one could have enforced the constraint in the H^1 scalar product, with the Lagrangian $L(\tilde{u}, \alpha, \lambda) = \frac{1}{2}\|\nabla\tilde{u}\|^2 - (f, \tilde{u}) + \Big(\nabla(\tilde{u} - \sum_i \alpha^i \varphi_h^i), \nabla\lambda\Big)$. This leads to similar formulas as above, and corresponding conclusions. The Lagrange multiplier is now $\lambda = (-\Delta)^{-1}(f + \Delta u_h) = e \in H^1$, where $e = u - u_h$ is the error.

3 Goal orientation and nonlinear problems

After this introduction, let us generalize the above formulation to nonlinear and possibly indefinite problems, and a general goal functional instead of the natural

energy norm. Assume we wanted to find the solution $u \in V$ of the variational problem

$$A(u)(\varphi) = 0 \qquad \forall \varphi \in V,$$

where A is some semilinear form that includes the operator as well as the right hand side and boundary terms. For the Laplace equation above, $A(u)(\varphi) = (\nabla u, \nabla \varphi) - (f, \varphi)$. Let the goal of the computation be the accurate computation of some functional $J(u)$ of the solution. Assuming that the semilinear form introduced above allows for a unique solution, we can write this problem as an optimization problem (see, for example, Becker and Rannacher [2001]):

$$\min_{u \in V} J(u), \qquad \text{subject to } A(u)(\varphi) = 0 \qquad \forall \varphi \in V.$$

This problem is a trivial optimization problem, since the constraint has only one feasible point. After discretization, the finite-dimensional solution u_h would be determined in a similar way:

$$\min_{u_h \in V_h} J(u_h), \qquad \text{subject to } A(u_h)(\varphi_h) = 0 \qquad \forall \varphi_h \in V_h.$$

In the same way as in the previous section, let us rewrite this finite-dimensional problem such that we minimize over the full, infinite-dimensional space, and add discreteness as an explicit constraint:

$$\begin{aligned} &\min_{\tilde{u} \in V, \alpha \in \mathbb{R}^N} J(\tilde{u}), \\ \text{subject to} \quad &A(\tilde{u})(\varphi_h) = 0 && \forall \varphi_h \in V_h, \\ \text{and} \quad &\left\langle \tilde{u} - \sum \alpha^i \varphi_h^i, \psi \right\rangle = 0 && \forall \psi. \end{aligned} \tag{7}$$

Here, $\langle \cdot, \cdot \rangle$ denotes either the H_0^1-H^{-1} duality pairing or the Dirichlet scalar product $(\nabla \cdot, \nabla \cdot)$, putting the two possibilities discussed at the end of the previous section into a common framework. Depending on the choice of this product, the test functions ψ are then either from $\Lambda = H^{-1}$ or $\Lambda = H_0^1$. Note that in (7) the state equation constraint alone does *not* determine the solution $\tilde{u}$ uniquely, since we search in an infinitely dimensional space but only test with functions from a finitely dimensional one. The second constraint is therefore necessary to guarantee a unique solution.

In order to solve (7), we define a Lagrangian as

$$L(\tilde{u}, \alpha, z_h, \lambda) = J(\tilde{u}) - A(\tilde{u})(z_h) + \left\langle \tilde{u} - \sum \alpha^i \varphi_h^i, \lambda \right\rangle.$$

The optimality conditions are now

$$L'_{\tilde{u}} = J'(\tilde{u})(\varphi) - A'(\tilde{u})(\varphi, z_h) + \langle \varphi, \lambda \rangle = 0 \qquad \forall \varphi \in V, \tag{8}$$

$$L'_{\alpha^i} = -\left\langle \varphi_h^i, \lambda \right\rangle = 0 \qquad \forall 1 \le i \le N, \tag{9}$$

$$L'_{z_h} = -A(\tilde{u})(\varphi_h) = 0 \qquad \forall \varphi_h \in V_h, \tag{10}$$

$$L'_{\lambda} = \left\langle \tilde{u} - \sum \alpha^i \varphi_h^i, \psi \right\rangle = 0 \qquad \forall \psi \in \Lambda. \tag{11}$$

From the discreteness constraint (11) and equation (10) we see that indeed $\tilde{u} = u_h$. On the other hand, choosing discrete test functions in (8) and using (9) shows that z_h is the discrete solution of the dual problem $A'(\tilde{u})(\varphi_h, z_h) = J'(\tilde{u})(\varphi_h)$ for all test functions $\varphi_h \in V_h$. Then, it is obvious that λ is the *residual of this dual problem*. Note that if we take the natural energy functional as goal functional as we did in the previous section, then primal and dual solution coincide, and the Lagrange multiplier λ is also the primal residual. It is also worth mentioning that in the case of finite element methods, the solution of the discrete dual problem, z_h, is defined on the same mesh as the primal solution $\tilde{u}$, and need not be computed on a finer mesh.

As a last observation in this section, let us consider how perturbations in the constraint and corresponding perturbations in the solution $\tilde{u}$ affect the value of the functional $J(\cdot)$ at the optimum of the unperturbed and the perturbed problems. Let $\tilde{u}_1, z_h, \lambda$ be the solution of problem (7), and $\tilde{u}_2 \in V$ an arbitrary function that is presumably close to $\tilde{u}_1$. Then, by Taylor expansion and using (8), we have with $\delta\tilde{u} = \tilde{u}_2 - \tilde{u}_1$:

$$\begin{aligned} J(\tilde{u}_2) - J(\tilde{u}_1) &= J'(\tilde{u}_1)(\delta\tilde{u}) + \mathcal{R}_1 \\ &= A'(\tilde{u}_1)(\delta\tilde{u}, z_h) - \langle \delta\tilde{u}, \lambda \rangle + \mathcal{R}_1, \end{aligned} \tag{12}$$

where $\mathcal{R}_1$ is a higher order term at least quadratic in $\delta\tilde{u}$.

Let us consider the special case that not only $\tilde{u}_1 = u_h$ satisfies $A(\tilde{u}_1)(\varphi_h) = 0$ for all discrete test functions $\varphi_h \in V_h$, but that the same holds true for $\tilde{u}_2$ as well. For example, this is the case if $\tilde{u}_2$ is the solution of the same problem on a finer grid, or the exact solution u. Then,

$$0 = A(\tilde{u}_2)(\varphi_h) - A(\tilde{u}_1)(\varphi_h) = A'(\tilde{u}_1)(\delta\tilde{u}, \varphi_h) + \mathcal{R}_2, \tag{13}$$

where again $\mathcal{R}_2$ is a higher order term at least quadratic in $\delta\tilde{u}$. Inserting this into (12), we get

$$J(\tilde{u}_2) - J(\tilde{u}_1) = -\langle \delta\tilde{u}, \lambda \rangle + \mathcal{R}_1 + \mathcal{R}_2, \tag{14}$$

From it we infer that the Lagrange multiplier indicates, to first order, changes in the minimization functional under (Galerkin) perturbations of the constraint. It is this aspect of the residual/Lagrange multiplier that we will stress in the rest of this work. As a side note, given what λ denotes in the two cases discussed above, it is obvious that the above representation does not depend on which type of product the bilinear form $\langle \cdot, \cdot \rangle$ refers to.

4 Application to "optimal" mesh refinement

Considering the special case of finite element schemes with a basis defined over a grid, let us apply the above ideas to devise strategies for "optimal" mesh refinement. Here, by "optimal" we mean that we want to refine those cells that bring the value of the goal functional $J(\cdot)$ at the discrete solution closest to its

optimal value when minimized over the full space V. We approach this question by asking: *if we refine cell K, how much would the value of J change*. We will repeat this for every cell in the triangulation, and pick those cells for refinement for which the change is largest.

As stated, we have already made an approximation, since we are considering the change due to refining each cell separately from all others. Indeed, since the Ritz projection is not local, refining two cells does not yield a change in J that is simply the sum of the changes due to each of these. We will ignore these second-order effects for practical purposes, and will also make other approximations.

Of course, we can not simply evaluate the change in J upon refining cell K, since this would require the solution of a full problem for each K. Rather, we will make use of the fact that the change is, to first order, given through the Lagrange multiplier. Doing so amounts to a kind of saturation assumption: we drop the second order terms in (14) because we assume the error to be small already.

If we denote by $\tilde{u}_1 = u_h$ the solution on the "coarse" mesh with associated space V_h, and $\tilde{u}_2 = u_h^{+K}$ the solution on the mesh with cell K refined and associated space $V_h \oplus V^{+K}$, where V^{+K} is the space spanned by the newly added shape functions, then we have by (14) and using the assumption stated above:

$$J(u_h^{+K}) - J(u_h) \approx -\left\langle \delta\tilde{u}^{+K}, \lambda \right\rangle,$$

where $\delta\tilde{u}^{+K} = u_h^{+K} - u_h$. We can obviously evaluate the residual/Lagrange multiplier λ by solving the discrete dual problem (8) for z_h only once, i.e. independently of the cell K presently under consideration; note again that λ is the residual of the primal problem if J is the natural energy functional induced by $A(\cdot)(\cdot)$. The question is then whether we have any chance to evaluate the change in the solution $\delta\tilde{u}^{+K}$. For simplicity of notation, assume that we are dealing with a linear problem, in which case $A(\cdot)(\cdot)$ can be written as $A(u)(\varphi) = a(u,\varphi) - (f,\varphi)$. Then $\delta\tilde{u}^{+K}$ satisfies

$$\begin{aligned} a(\delta\tilde{u}^{+K}, \varphi_h^i) &= 0 && \forall \varphi_h^i \in V_h, \\ a(\delta\tilde{u}^{+K}, \varphi_i^{+K}) &= (f, \varphi_i^{+K}) - a(u_h, \varphi_i^{+K}) && \forall \varphi_i^{+K} \in V^{+K}, \end{aligned}$$

where φ_i^{+K} are the basis functions of V^{+K}, i.e. for lowest-order elements one hat function for the cell interior and each edge and face. Unfortunately, since the Ritz projection is not local, $\delta\tilde{u}^{+K}$ is not local either, and in particular not in V^{+K} only. However, due to the decay properties of the Ritz projector, we may *approximate* it locally by some $\widehat{\delta\tilde{u}}^{+K} \in V^{+K}$. For this, two standard approaches exist (although usually smooth bubbles are taken as additional test functions, rather than the hierarchical enrichment). In the first approach, we solve a local Dirichlet problem on the patch $\hat{K}$ including the cell K and its neighbors:

$$a(\widehat{\delta\tilde{u}}^{+K}, \varphi_i^{+K})_{\hat{K}} = (f, \varphi_i^{+K})_{\hat{K}} - a(u_h, \varphi_i^{+K})_{\hat{K}}. \tag{15}$$

Zero boundary conditions are imposed at the boundary $\partial\hat{K}$. This approach has first been proposed in the context of error estimation in Babuška and Rheinboldt [1978].

An alternative is to solve local Neumann problems. For more information on this, see Ladevèze and Leguillon [1983], Bank and Weiser [1985], Ainsworth and Oden [1993]. In both cases, $\widehat{\delta\tilde{u}}^{+K}$ can be interpreted as the local correction on cell K to u_h obtained by a single iteration of a Jacobi preconditioned form of a defect correction scheme, since the right hand side of the defining equation (15) is the residual of the original problem.

With the so-defined approximations $\widehat{\delta\tilde{u}}^{+K}$ for each cell K, we can then set out to compare refinement indicators $\eta_K = |\langle\widehat{\delta\tilde{u}}^{+K}, \lambda\rangle|$, defined as the amount by which the target functional $J(\cdot)$ would change when refining cell K.

Remark 1. *The strategy of constructing local approximations to the error $e = u - u_h$ and using them for mesh refinement is well known, and described in a plethora of papers. For a general overview, see, for example Verfürth [1996], Ainsworth and Oden [2000], Babuška and Strouboulis [2001]. However, there the refinement strategy is usually based on picking those cells for which the error is largest, using error estimates. The derivation above demonstrates the non-trivial conjecture that this is actually equivalent to picking those cells for which the change in the target functional is largest under refinement of a single cell.*

5 Error estimation

In the previous section, we chose $\tilde{u}_2$ equal to the solution of a problem where we refined a single cell. In this section, let us instead set it equal to the exact solution: $\tilde{u}_2 = u$. Then $\delta\tilde{u} = \tilde{u}_2 - \tilde{u}_1 = u - u_h = e$ is the error. From (14) we then see that

$$J(u) - J(u_h) = -\langle e, \lambda\rangle + \mathcal{R}_1 + \mathcal{R}_2. \tag{16}$$

Note in particular that if both A and J are linear, then the quadratic remainders $\mathcal{R}_i$, $i = 1, 2$ vanish.

In this section, let us briefly indicate how this formula can be placed among the various ways errors are estimated in the literature. Basically, we may use this error representation in two ways. In the first, we approximate the error e by some $\hat{e}$, to get

$$J(u) - J(u_h) \approx -\langle\hat{e}, \lambda\rangle\,. \tag{17}$$

Since λ can be evaluated after solving a discrete dual problem for z_h, there remains the choice of $\hat{e}$. For this, various strategies have been devised in the literature. For example, one may compute the solution u_h^+ on a finer grid, and set $\hat{e} = u_h^+ - u_h$, see for example Becker and Rannacher [2001], Bangerth and Rannacher [2003]. Alternatively, the local corrections defined in the previous

section may be used to set

$$\hat{e} = \sum_K \widehat{\delta\tilde{u}}^{+K}.$$

Many other possibilities of various sophistication are possible, which we will all not discuss further. For more information, see the general references mentioned above. If such an estimate $\hat{e}$ has been computed, the numerical evaluation of the *goal oriented error approximation* (17) is straightforward.

Alternatively, introducing the *continuous* dual problem

$$J'(\tilde{u}) - A'(\tilde{u})(\varphi, z) = 0 \qquad \forall \varphi \in V,$$

and using it together with (8), we see that $\langle \varphi, \lambda \rangle = -A'(u_h)(\varphi, z - z_h)$. Thus,

$$J(u) - J(u_h) = A'(u_h)(e, z - z_h) + \mathcal{R}_1 + \mathcal{R}_2. \tag{18}$$

This leads us to residual based error estimators, since the first term on the right is actually a residual weighted by $z - z_h$. For example, for the Laplace equation, using cell-wise integration by part, this equation attains the well-known form

$$J(u) - J(u_h) = \sum_K (f + \Delta u_h, z - z_h)_K + \tfrac{1}{2}(n \cdot [\nabla u_h], z - z_h)_{\partial K}, \tag{19}$$

where the normal jump at a face ∂K between cells K and K' is defined as usual by

$$n \cdot [\nabla u_h]|_K = \begin{cases} n \cdot (\nabla u_h|_{K'} - \nabla u_h|_K) & \text{if } \partial K \not\subset \partial\Omega, \\ -2n \cdot \nabla u_h|_K & \text{otherwise,} \end{cases}$$

and the outward normal vector is denoted by n.

Instead of finding an approximation for the *primal error* e, the challenge lies here in approximating the *dual error* $z - z_h$. We will not discuss this approach either, but instead refer to the book Bangerth and Rannacher [2003] which is entirely devoted to error estimation and adaptivity based on weighted residuals.

6 Conclusions

We have given a different statement of Galerkin approximation, by not assuming that we perform variation only on a finite-dimensional subspace, but putting the discreteness into an explicit constraint. This leads, in a Lagrangian formulation, to a multiplier for the constraint, that then turns out to be the residual of either the primal problem (if we minimize the natural energy functional associated with the differential equation) or of a dual problem (in the general goal-oriented setting).

Well-known properties of Lagrange multipliers were then used to analyze refinement strategies and error estimates. In particular, some common strategies for mesh refinement based on the size of an error indicator are re-interpreted

to be "optimal" in the sense that they pick those cells which will generate the largest change in the target functional, thus bringing the value of the desired quantity closest to its exact value. The formulas resulting from the formulation as a constrained variational problem were then also used to derive error estimation formulas.

The results of this paper are not new in the sense that the representations have been known previously. However, it seems that their derivation using the discreteness-as-a-constraint idea have not been given in the literature so far. Insofar, this paper offers a different perspective on otherwise known facts, and may help clarify our intuition what the residual is in finite element computations and what it can be used for. Our results also allow us to appreciate the role of the residual in error estimates and mesh refinement criteria, as it not only denotes a measure of the error if viewed in appropriate norms, but also indicates how restrictive the discreteness constraint is, or conversely how much the solution would change if the constraint didn't exist.

Acknowledgments The author acknowledges help in preparing this paper by F. Nobile, J.-L. Guermond and B. Barth, as well as the encouragement to write it given by J. T. Oden, I. Babuška and T. Hughes.

References

M. Ainsworth and J. T. Oden. A unified approach to a posteriori error estimation using element residual methods. *Numer. Math.*, 65:23–50, 1993.

M. Ainsworth and J. T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. John Wiley and Sons, 2000.

I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15:736–754, 1978.

I. Babuška and T. Strouboulis. *The Finite Element Method and its Reliability*. Clarendon Press, New York, 2001.

W. Bangerth and R. Rannacher. *Adaptive Finite Element Methods for Differential Equations*. Birkhäuser Verlag, 2003.

R. E. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Math. Comp.*, 44:283–301, 1985.

R. Becker and R. Rannacher. An optimal control approach to error estimation and mesh adaptation in finite element methods. *Acta Numerica*, 10:1–102, 2001.

P. Ladevèze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM J. Numer. Anal.*, 20:485–509, 1983.

R. Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*. Wiley/Teubner, New York, Stuttgart, 1996.