PENALTY METHODS IN FINITE ELEMENT ANALYSIS OF FLUIDS AND STRUCTURES

D. S. MALKUS

*Department of Mathematics, Illinois Institute of Technology,
IIT Center, Chicago, Illinois 60616, U.S.A.*

Abstract

This paper discusses the use of penalty techniques in finite element Galerkin equations. It shows that earlier results relating penalty methods to mixed and Lagrange multiplier methods through the use of reduced/selective numerical integration formulas extend to problems in which there is no extremum principle. It is also shown that penalty techniques can be employed in nonlinear dynamical problems and can lead to the substantial computational savings at each time step. The penalty techniques described here can lead to computational benefits arising from the ability to impose constraints without Lagrange multiplier unknowns. But they also may lead to costs involving the slowing of convergence rates of certain nonlinear iteration schemes, the ill-conditioning of certain iteration matrices and the complication of some mesh refinement schemes. Nevertheless in many problems the benefits well outweigh the costs.

1. Introduction

Penalty methods are a computationally efficient device for imposing difficult constraints in the finite element method. The early applications usually involved steady-state problems in which the solution was characterized as an extremal of a variational functional. In this context, a basic equivalence between the finite element formulation using penalties and the Lagrange multiplier/finite element formulation was discovered[1]. The result was that the computational ease of the penalty formulation could be used to give results with the accuracy attainable using Lagrange multipliers with suitably chosen elements.

This paper shows that these early results extend to nonlinear dynamical problems in which the solution is characterized by a Galerkin statement. It discusses the costs and benefits which can be expected for the employment of penalties. Consider the general mixed Galerkin statement: Find $(u, g) \in S \times T$ such that for all $(\psi, h) \in S \times T$ (with ψ perhaps satisfying some homogeneous b.c.)

$$A(u, \psi) + \int_V B(u, g, \psi) dV + \int_V h(N(u) - \alpha g) dV + F(\psi) = 0 \quad (1)$$

V is the spatial domain of the problem; $F(\psi)$ represents applied loads. $A(u, \psi)$ represents the contribution to Galerkin equations which does not involve the constraints. The second integral in (1) imposes the constraint

$$N(u) = \alpha g \quad (2)$$

where $N: S \rightarrow T$ is an operator (possibly tensor-valued, in which case a dot product is implied in the second integral). B is a real-valued function of u , g and ψ . The statement (1) may apply to a steady-state problem or to a transient problem if $A(u, \psi)$ involves time derivatives of u and (1) is required to hold at each instant t .

(1) admits to finite element spatial discretization (and time differencing for transient problems) by choosing trial spaces for S and T in the usual fashion. The result is a mixed finite element/Galerkin method, which is a Lagrange multiplier method for $\alpha=0$, imposing the nonlinear constraint $N(u) = 0$. The drawback is the need for two trial spaces S and T . An alternative is to impose the constraint $N(u) = \alpha g$ directly by formally substituting $\alpha^{-1} N(u)$ for g . In the $\alpha=0$ case α is taken to be small and α^{-1} becomes the penalty enforcing $N(u) = \alpha g$ without Lagrange multiplier. The second integral in (1) vanishes and the first becomes the penalty term:

$$A(u, \psi) + \int_V B(u, \alpha^{-1} N(u), \psi) dV + F(\psi) = 0 \quad (3)$$

The finite element solution to the penalty form (3) and the solution obtained from the original form (1) may be expected to differ. This follows because in the finite element model of (1), the constraint is only satisfied in the weak sense. In (3) the constraint is imposed exactly.

The key to the successful employment of the penalty method (3) lies in the recognition that the exact imposition of the constraint in (3) overconstrains the problem. It was discovered that the use of reduced integration to evaluate the integral in (3) "softens" the constraint and leads to accurate results [1]. Malkus and Hughes proved an important equivalence theorem which, in essence, stated that the reason reduced integration performs this function is that, for a specific T , it makes the solutions to (1) and (3) the same, and T assures that (1) is a well constrained problem. Recently Oden [2] and also

Bercovier [3] have shown that the need for reduced integration can be related to a specific stability criterion for (3). The Malkus-Hughes equivalence theorem was proved in the context of minimization of a functional. It will now be shown that the equivalence also holds for the more general problems described by equations (1-3).

2. The Equivalence Theorem

The following construction is crucial to the establishment of the equivalence:

- (A) $\left\{ \begin{array}{l} \text{Given an integration formula for each element, construct a trial} \\ \text{space } T \text{ with the integration points for nodes.} \\ \text{Construct the element assemblage without interelement continuity} \\ \text{of the } g\text{'s (separate nodal values on the opposite sides of bound-} \\ \text{aries if necessary).} \end{array} \right.$

Note that such a T has a nodal interpolate to functions discontinuous across element boundaries.

Theorem 1: The following two procedures produce the same finite element solution:

- I. Choose an appropriate trial space S . Construct T according to (A). Use the integration formula described in (A) to evaluate the two integrals appearing in (1), and any appropriate formula to evaluate $A(u, \psi)$. Solve (1) to produce (u, g) .
- II. Use the same trial space S and evaluate $A(u, \psi)$ exactly as in I. Use the integration formula of (A) to evaluate the integral in (3). Solve (3) to produce u . Interpolate to $\alpha^{-1} N(u)$ at the integration points of (A) to produce nodal values of g in T .

Proof: If (1) is satisfied and the integrals are evaluated by a quadrature formula based on the nodes of the g 's and h 's, then the second integral in (1) can be seen to imply that $N(u)$ must interpolate to αg at the nodes of T . Substituting this result into the first integral shows that (3) is satisfied for any $\alpha > 0$. On the other hand a similar argument shows that if II. is satisfied so also is I. *q.e.d.*

Clearly II. is a simpler procedure than I., involving fewer unknowns and avoiding the "saddle-point" introduced by the g and h variables. I. has the advantage that for $\alpha = 0$, S and T can be chosen in such a way that the problem is well-constrained. The analysis in [1] shows that when T is constructed according to (A), i.e. when T is based on the reduced integration points of S , this produces a well constrained pairing of S and T . Furthermore Oden [2] and Bercovier [3] have shown that for a variety of specific B 's and N 's, the use of reduced integration implies that for a fixed mesh, the sequence of g 's is bounded as $\alpha \rightarrow 0$. This means that α can be chosen small enough so that $N(u)$ is negligible. The penalty form of (3) may be solved as described in II. to produce the solution which would be obtained with more work from (1) using procedure I.

In Section 4. specific examples of applications of these techniques will be described. Before considering the specific applications, it is shown that if certain restrictions on the form of $B(u, g, \psi)$ are imposed - restrictions which are satisfied in all of the described applications - the solution procedures of I. and II. are essentially equivalent.

3. Equivalence of the Incremental Equations

It follows from the derivation of the Galerkin equations of form (1) - either from an operator equation or from the variational derivative of a functional - that A , B and F

are linear in ψ . If B is assumed to have the form

$$B(u, g, \psi) = gC(u, \psi) \quad (4)$$

then a stronger result that Theorem 1. can be established. In this case, the incremental matrix obtained by differentiation of the left-hand side of (1) with respect to the nodal variables u_i of S and g_j of T is closely related to the incremental matrix of (3).

First some convenient notation: $\partial u / \partial u_i$ denotes the vector of finite element basis functions of S taken in the order specified by u. Likewise $\partial g / \partial g_j$ can be used to denote the vector of basis functions of T. Let $\int_V$ denote the numerical integral of construction (A). Now consider the following matrices obtained by differentiating the left-hand side of (1) with respect to u_i :

$$K_2 = \text{mat} \left[\frac{\partial A}{\partial u_i} \left(u, \frac{\partial u}{\partial u_j} \right) \right] \quad (5)$$

$$\bar{K}_0 = \text{mat} \left[\int_V \alpha g \frac{\partial C}{\partial u_i} \left(u, \frac{\partial u}{\partial u_j} \right) dV \right] \quad (6)$$

$$\bar{K}_1 = \int_V \frac{\partial N(u)}{\partial u_i} \left(\frac{\partial g}{\partial g_j} \right)^T dV \quad (7)$$

$$\bar{H}_1 = \text{mat} \left[\int_V \frac{\partial g}{\partial g_j} C \left(u, \frac{\partial u}{\partial u_i} \right) dV \right] \quad (8)$$

$$Q = \int_V \frac{\partial g}{\partial g_j} \left(\frac{\partial g}{\partial g_i} \right)^T dV \quad (9)$$

Using the ordering (u_i, g_j) of the unknowns, the incremental matrix of (1) is

$$\begin{bmatrix} K_2 + \bar{K}_0 & \bar{K}_1 \\ \bar{H}_1^T & -\alpha Q \end{bmatrix} \quad (10)$$

To derive the incremental matrix of (3) consider the additional matrices:

$$K_0 = \text{mat} \left[\int_V N(u) \frac{\partial}{\partial u_i} C \left(u, \frac{\partial u}{\partial u_j} \right) dV \right] \quad (11)$$

$$K_1 = \text{mat} \left[\int_V \frac{\partial N(u)}{\partial u_i} C \left(u, \frac{\partial u}{\partial u_j} \right) dV \right] \quad (12)$$

In terms of these matrices and K_2 as defined in (5), the incremental matrix obtained by differentiating the left-hand side of (3) with respect to u_i is

$$K_2 + \alpha^{-1}(K_1 + K_0) \quad (13)$$

The incremental nodal variables g_j may be eliminated from (10) to yield the "reduced hessian":

$$K_2 + \alpha^{-1}(\bar{K}_1 Q^{-1} \bar{H}_1^T + \bar{K}_0) \quad (14)$$

Theorem 2: If construction (A) is used and the procedure I. is followed

$$K_1 \equiv \bar{K}_1 Q^{-1} \bar{H}_1^T \quad (15)$$

and

$$\alpha^{-1}(K_0 - \bar{K}_0)_{ij} = \int_V [\alpha^{-1}N(u) - g] \frac{\partial C(u, \frac{\partial u}{\partial u_i})}{\partial u_j} dv \quad (16)$$

Proof:

lemma: For a given vector of nodal values ξ defining $s \in S$

$$h = Q^{-1} \bar{H}_1 \xi \quad (17)$$

defines $h \in T$ which interpolates at the nodes to $[\partial N(u)/\partial u]_i^T \xi$.

Pf: (17) implies

$$\int_V \frac{\partial g}{\partial \xi} \left[\frac{\partial N(u)}{\partial u} \right]^T \xi dv = \int_V \frac{\partial g}{\partial \xi} h dv \quad (18)$$

and by the reasoning of Theorem 1., the result follows.

Now consider for ξ, χ arbitrary nodal-value vectors of S , (12) implies

$$\xi^T K_1 \chi = \int_V \xi^T \frac{\partial N(u)}{\partial u} C(u, v) dv = \int_V h C(u, v) dv = h \bar{H}_1^T \chi \quad (19)$$

from (8). But according to the lemma, the right-hand side of (19) is what is required in (15). Note that (16) follows directly from (6) and (11). q.e.d.

Corollary 1: At a solution to (1) or (3) the reduced hessian and hessian of (3) are equal.

Corollary 2: If $C(u, \psi)$ is independent of u then the conclusion of Cor. 1. holds.

Corollary 3: If $C(u, \psi) = \partial N(u, \psi)$, the Frechet derivative of $N(u)$, then $\bar{H}_1 = \bar{K}_1$, and $\bar{K}_0$, K_0 and K_1 are all symmetric.

Corollary 4: Theorem 2. holds on the element level.

The proofs of the corollaries are simple. 1. makes $\bar{K}_0 = \bar{K}_1$, 2. implies both are zero. 3. follows from the fact that on a finite dimensional space $\partial N(u, \cdot) = \partial N/\partial u$. The final corollary follows from the fact that construction (A) gives a trial space T which has no interelement connectivity.

It should be noted that in the examples of Section 4., the conditions of Cor. 3. are always met. This means that the Hessians referred to in this section are the sum of K_2 and a symmetric matrix. This can lead to substantial computational storage savings and algorithmic stability. When the conditions of Cor. 3. are met, this says that the constraint is of the same form as the constraint in a constrained minimization problem, whether or not the Galerkin equations were so derived. The penalty term is the same as would be derived by the variational differentiation of

$$\frac{\alpha^{-1}}{2} \int_V N^2(u) dv \quad (20)$$

assuming of course, that $N(u)$ is Fréchet differentiable.

4. Applications

The applications presented here have Galerkin equations of the form

$$A(u, \psi) + \int_V g \partial N(u, \psi) dv + \int_V h (N(u) - \alpha g) dv + F(\psi) = 0 \quad (21)$$

$F(\psi)$ is derived from arbitrary boundary and surface loads and will not be specified further. Therefore to specify (21) it is only required to identify A , N and α . The penalty form (3) is easily deduced from (1) and (3).

4.1 A Simple Minimization Problem

An elastic beam:

$$u = \begin{bmatrix} W \\ \theta \end{bmatrix}$$

$$A(u, \psi) = 2 \int_0^L \theta, x \psi, x dx$$

$$N(u) = 2(W, x - \theta)$$

$$\alpha = t^2(1+\nu)/3$$

there W is deflection, θ rotation and t is the thickness of the beam. The shear correction factors have been taken to be 1. g is the shear force. More about this and related problems can be found in [1].

4.2 Large-Strain, Incompressible Elasto-Statics

In this problem the Galerkin statement comes from a variational principle as with 4.1, but $N(u)$ is highly nonlinear.

$$A(u, \psi) = \int_V \frac{\partial W}{\partial u_{i,j}} \psi_{i,j} dV$$

$$N(u) = 2G(III(u) - 1)$$

$$\alpha = \text{a modulus of compressibility.}$$

Here $W = W(I, II)$ where I , II and III are the three principal invariants of the Cauchy-Green tensor of the displacement field u_i . G is a function such that $G(x) = 0$ iff $x = 0$. g is $G'(0)$ times the hydrostatic pressure. III is of sixth degree in strains and G may be highly nonlinear. In fact, G may be a function of III which is singular at $III = 0$ to assure that the energy is infinite at zero volume. More about this problem can be found in [1].

4.3 The Transient, Incompressible, Navier-Stokes Equations

$$A(u, \psi) = - \int_V [1/2 (V_{i,j} + V_{j,i}) (\psi_{i,j} + \psi_{j,i}) + R_e (\frac{\partial V_i}{\partial t} \psi_i + V_j V_{j,i} \psi_i)] dV \quad (22)$$

$$N(u) = V_{i,i}$$

$$\alpha = \text{artificial modulus of compressibility}$$

Here nondimensional units have been employed and R_e is the Reynolds number. $u = V_i$, the nondimensional velocity field, and g is the hydrostatic pressure.

The following version of the Newmark difference scheme has been employed in an implicit-explicit form by Hughes [4]. He reports successful use of this scheme in a variety of high-Reynolds number transient flow problems. The time-differenced version of (1) is

$$\text{Let } V_0 \text{ and } g_0 \text{ be start-up values}$$

$$a_0 = -M^{-1}[R_e^{-1}(K_2 V_0 + \bar{K}_1 g_0) + D(V_0) - R_e^{-1}F] \quad (23a)$$

$$\tilde{V}_{n+1} = V_n + \Delta t(1-\gamma)a_n$$

$$a_{n+1} = (V_{n+1} - \tilde{V}_{n+1})/\gamma \Delta t$$

$$\begin{bmatrix} \frac{K_2 + \lambda M}{K_1} & \bar{K}_1 \\ \bar{K}_1^T & -\alpha Q \end{bmatrix} \begin{bmatrix} \bar{V}_{n+1} \\ \bar{S}_{n+1} \end{bmatrix} = \begin{bmatrix} \lambda M \bar{V}_{n+1} - R_e D(\bar{V}_{n+1}) + F \\ 0 \end{bmatrix} \quad (23b)$$

Here Δt is the time-step, γ can be chosen to enhance stability, $\lambda = R_e / \gamma \Delta t$, and $D(\bar{V}_{n+1})$ is the nonlinear part of $A(u, \psi)$ coming from the inertial terms in (22). M is the mass matrix of S , usually lumped. Hughes devised this splitting so that D is treated explicitly and the viscous terms implicitly. Cor. 3. implies that the implicit part of the operator is symmetric. But he does not use the form corresponding to (1), but rather the penalty form (3). Since the nonlinear part of A is put on the right-hand side, K_0 and $\bar{K}_0$ of Theorem 2. are not assembled. Theorem 2. and Cors. 2.-4. imply that the pressure can be eliminated on the element level, and denoting the viscous-term-with-penalty matrix of the implicit part of the operator by K , (23a) and (23b) become

$$a_0 = -M^{-1}[R_e^{-1}KV_0 + D(V_0) - R_e^{-1}F] \quad (24a)$$

$$(K + \lambda M)\bar{V}_{n+1} = \lambda M\bar{V}_{n+1} - R_e D(\bar{V}_{n+1}) + F \quad (24b)$$

The number of unknowns is substantially reduced, and the implicit matrix operator is positive-definite.

5. Advantages and Drawbacks

The advantages of using the penalty techniques have already been pointed out, but nevertheless bear repeating: Advantage #1 is the reduction in the number of unknowns, which was observed in all of the applications of the previous section. Advantage #2 is the removal of the saddlepoint indefiniteness of the Lagrange multiplier formulation. If the operator is otherwise positive definite, as is the case in 4.1, near a minimum in 4.2, cleverly arranged in 4.3, numerical stability and simplicity are gained. Advantage #3 is the simplicity of computer coding of the penalty formulation with its single trial space — as opposed to the pairing required by Lagrange multipliers.

The major disadvantages are also three in number: Disadvantage #1 is the ill-conditioning with respect to inversion which is inherent in the incremental matrices with penalties. This ill-conditioning cannot be scaled out. Disadvantage #2 is that the presence of the penalty slows down some iterative algorithms for matrix inversion or non-linear iteration. This is particularly the case with the conjugate-gradient and Polak-Ribière algorithms. Disadvantage #3 is that the g 's calculated from $\alpha^{-1}N(u)$ are only accurate at the reduced integration points. Therefore schemes which refine meshes by interpolating to the solutions on crude meshes for starting values (a very common practice in application 4.2) can produce very poor starting values for g .

It is the opinion of this author that in most applications to fluid and structural problems, the benefits of penalty methods outweigh the costs. This is particularly the case in large scale dynamical problems such as application 4.3. More work is needed in application 4.2 in order to minimize the disadvantages mentioned above, but the penalty approach nevertheless seems promising.

Acknowledgement

The ideas presented here are an extension of original ideas arrived at independently by T. J. R. Hughes and myself. He has become perhaps the most skilled practitioner of the methods I have described here. The Navier-Stokes application comes entirely from his work and that of his colleagues.

References

- [1] MALKUS, D.S., HUGHES, T.J.R., "Mixed finite element methods — reduced selective integration techniques: a unification of concepts," Comp. Meth. Appl. Mech. and Engng., 15(1), pp. 63-81 (1978).
- [2] ODEN, BECKER, CAREY, Manuscript in preparation for an advanced course in finite element methods.
- [3] BERCOVIER, M., "Perturbations of mixed variational problems. Applications to mixed finite element methods," R.A.I.R.O. Analyse Numérique/Numerical Analysis, 12(3), pp. 211-236 (1978).
- [4] HUGHES, T.J.R., LIU, W.K., BROOKS, A., "Finite element analysis of incompressible viscous flows by the penalty function formulation," J. Comp. Phys., (to appear).

Errata

Unfortunately since this paper went to press, I have discovered several notational errors in the matrix definitions in Section 3. The definitions of $\bar{H}_1$ and $\bar{K}_1$ are reversed, i.e. the left-hand side of (7) should be $\bar{H}_1$ and that of (8) should be $\bar{K}_1$. In (10) $\bar{K}_0$ should be multiplied by α^{-1} . In (17), the transpose of $\bar{H}_1$ should appear. Likewise in (19) the far left-hand side should be the transpose of K_1 and the far right-hand side should be $h^{T-T} K_1^T \chi$. I apologize for any confusion this may have caused.