# ICES REPORT 09-11

May 2009

# Enforcement of Constraints and Maximum Principles in the Variational Multiscale Method

by

J.A. Evans, T.J.R. Hughes, and G. Sangalli


**The Institute for Computational Engineering and Sciences**
The University of Texas at Austin
Austin, Texas 78712

# Enforcement of Constraints and Maximum Principles in the Variational Multiscale Method

John A. Evans[a,*], Thomas J.R. Hughes[a], and Giancarlo Sangalli[b]

[a] The University of Texas at Austin
Institute for Computational Engineering and Sciences
1 University Station C0200
Austin, TX 78712-0027, U.S.A.

[b] Istituto di Matematica Applicata e Tecnologie Informatiche
Consiglio Nazionale delle Ricerche
Via Ferrata, 1
27100 Pavia, Italy

[*] Corresponding author. E-mail address: evans@ices.utexas.edu

May 5, 2009

**Abstract**

We present a new theoretical framework for the enforcement of constraints in variational multiscale (VMS) analysis. The theory is first presented in an abstract operator format and subsequently specialized for the steady advection-diffusion equation. The approach borrows heavily from results in constrained and convex optimization. An exact expression for the fine-scales is derived in terms of variational derivatives of the constraints, Lagrange multipliers, and a fine-scale Green's function. The methodology described enables the development of numerical methods which satisfy predefined attributes. A practical and effective procedure for solving the steady advection-diffusion equation is presented based on a VMS-inspired stabilized method, weakly enforced Dirichlet boundary conditions, and enforcement of a maximum principle and conservation constraint.

**Key words.** variational multiscale analysis, constrained optimization, convex optimization, Lagrange multipliers, projection, fine-scale Green's function, advection-diffusion, maximum principles, non-negativity, conservation, weak boundary conditions

# Contents

# 1 Introduction

The variational multiscale method [27, 29] was introduced as a framework for incorporating missing unresolved fine-scale effects into numerical problems governing coarse-scale behavior. It has provided a rationale for stabilized methods and a platform for the development of new computational technologies (see, e.g., [1, 4, 7, 14, 26, 31, 32, 33, 35, 38, 46] for application to turbulence modeling and simulation). Construction of the method is simple: decompose the solution $u$ to a partial differential equation into a sum of two components, $\bar{u}$ and $u'$, determine the fine-scale component $u'$ analytically in terms of the coarse-scale component $\bar{u}$, and solve for $\bar{u}$ numerically. The original instantiation of the method was based on variational projection. That is, the decomposition of the solution into a sum of coarse-scale and fine-scale components is uniquely specified by identifying a projector from the space of all scales onto the coarse-scale subspace. The fine-scale component can then be represented as the *fine-scale Green's function* of the coarse-scale residual. The structure of the fine-scale Green's function was studied extensively in [34]. It was discovered that different projectors give rise to different fine-scale Greens functions, and their properties can vary considerably. For example, in the context of a steady advection-diffusion problem, it was found that the projector induced by the $H^1_0$-seminorm results in a highly-attenuated fine-scale Green's function with local support whereas the projector induced by the $L_2$-norm does not. This motivated the construction of local approximations to the fine-scale Green's function for the $H^1_0$-projector [15], and indeed it was shown that the $H^1_0$-optimal variational multiscale method has many features in common with classical stabilized methods [9, 23, 30] and, in particular, SUPG [10].

While a variational multiscale method based on projection guarantees an optimally accurate coarse-scale approximation in terms of the chosen projection, it does not ensure that the numerical approximation exhibits certain characteristics of the solution. For example, an approximation obtained using the $H^1_0$-optimal variational multiscale method may not satisfy maximum principles or non-negativity constraints. Such an approximation may also exhibit undershoots and overshoots, as shown in the two-dimensional example in [34], about transition and boundary layers. This is undesirable for many applications. For example, in transport simulations, one seeks an approximation with non-negative density or concentration. Indeed, most compressible flow calculations terminate upon negative density values which occur due to undershoots about a shock wave. Shock-capturing or discontinuity-capturing procedures are often employed to add enough dissipation to remove these oscillations. Such schemes are usually *ad hoc* procedures involving artificial diffusion, but previous work in variational multiscale analysis has suggested the potential of studying these issues in a more fundamental way.

In [24], an approximation technique was introduced where the $L^p$-norm of the coarse-scale residual is minimized. Numerical tests in $L^1$ showed that the technique could handle discontinuities arising in solutions to first-order partial differential equations without resorting to limiting procedures. Inspired by this work, Demarco et al. introduced a variational multiscale method in [16] where the coarse-scale component $\bar{u}$ was defined through a non-linear optimization problem. In particular, $\bar{u}$ was defined

to be the best fit to $u$ in terms of the $W^{1,1}$-seminorm. However, numerical results revealed that such numerical approximations are not necessarily monotone, especially in the case of higher-order finite element basis functions. To complicate matters, both of the procedures above involve the solution of a difficult nonlinear problem.

In this paper, we develop a new class of variational multiscale methods which allows for the direct enforcement of equality and inequality constraints. The constraints could be, for example, that the coarse-scale solution satisfies a maximum principle, conserves a quantity over a region, or is monotone. In these multiscale methods, the decomposition of a solution into its coarse-scale and fine-scale components is defined through an optimization problem over a subset of the coarse-scale space, the so-called constraint set. Each function in this constraint set satisfies the given equality and inequality constraints. Provided the constrained optimization problem is well-posed, the chosen decomposition is uniquely specified, and an exact expression can be attained for the fine-scales in terms of the coarse-scale residual, variational derivatives of the constraints, Lagrange multipliers, and a fine-scale Green's function. The theory summarizing these ideas is presented in Section 2 in an abstract operator format for a general linear isomorphism.

These ideas are applied to the steady advection-diffusion problem in Section 3. A variational multiscale method is introduced in which maximum principles are enforced. We review the structure of the fine-scale Green's functions in two dimensions for the $L^2$-norm and $H^1_0$-seminorm and, following this, we apply our method to two model problems. Numerical results reveal the quality of our coarse-scale approximations. We end this section with a discussion of conservation and weak boundary conditions and present an additional numerical result.

We draw conclusions in Section 4.

# 2 The abstract setting

## 2.1 The abstract problem

Let $V$ be a Hilbert space, endowed with a scalar product $(\cdot,\cdot)_V$ and induced norm $\|\cdot\|_V$. Let $V^*$ be the dual of $V$ and let ${}_{V^*}\langle\cdot,\cdot\rangle_V$ be the pairing between them. Let $\mathcal{L} : V \to V^*$ be a linear isomorphism. Given $\mathcal{F} \in V^*$, we consider the abstract problem: find $u \in V$ such that

$$\mathcal{L}u = \mathcal{F}. \tag{1}$$

The variational formulation of (1) is: find $u \in V$ such that

$${}_{V^*}\langle \mathcal{L}u, v\rangle_V = {}_{V^*}\langle \mathcal{F}, v\rangle_V, \quad \forall v \in V. \tag{2}$$

The solution $u$ can be formally expressed as $u = \mathcal{G}\mathcal{F}$, where $\mathcal{G} : V^* \to V$ is the Green's (or solution) operator. That is, $\mathcal{G} = \mathcal{L}^{-1}$.

## 2.2 Approximation and constrained optimization

We are interested in finding an approximation $\bar{u}$ belonging to a closed finite-dimensional subspace $\bar{V} \subset V$ to the solution $u$ of (1). In the variational multiscale (VMS) approach, $\bar{V}$ represents the space of computable *coarse scales*. Previous instantiations of the VMS method have focused on finding approximations which minimize the error,

$$u' = u - \bar{u}, \tag{3}$$

or the *fine scales*, with respect to some norm of $V$. While such approximations are accurate, they may not preserve some desirable properties of the exact solution (e.g., positivity, maximum/minimum principles, monotonicity, conservation, etc.). In this work, we instead seek a numerical approximation which, in addition to being accurate, exhibits these desired properties. This naturally leads us to specifying $\bar{u}$ through a *constrained optimization problem*.

To proceed, define the closed constraint set

$$\bar{K} = \left\{\bar{v} \in \bar{V} : \mathbf{f}(\bar{v}) = \mathbf{0}, \mathbf{g}(\bar{v}) \leq \mathbf{0}\right\} \tag{4}$$

where $\mathbf{f} : \bar{V} \to \mathbb{R}^{n_{eq}}$ and $\mathbf{g} : \bar{V} \to \mathbb{R}^{n_{ineq}}$ are continuously differentiable (in the sense of Fréchet) vector functions of the finite-dimensional space $\bar{V}$. We assume that this set is not empty. The constraint set $\bar{K}$ represents the set of admissible approximations. We seek an admissible approximation which will minimize the error with respect to some norm of $V$. It will be useful to assume that this norm is induced from a scalar product as it will allow us to generalize previous results based on variational projection [34].

Let $H$ be a Hilbert space such that $V$ is continuously embedded in $H$. Let $H$ be endowed with a scalar product $(\cdot,\cdot)_H$ and induced norm $\|\cdot\|_H$. The constrained optimization problem for $\bar{u}$ is as follows: find $\bar{u} \in \bar{K}$ such that

$$\frac{1}{2}\|u - \bar{u}\|_H^2 = \min_{\bar{v}\in\bar{K}} \frac{1}{2}\|u - \bar{v}\|_H^2. \tag{5}$$

Two important questions with regards to the above constrained optimization are whether there exists a solution $\bar{u} \in \bar{K}$ and, if so, is such a solution unique. Existence of a solution is a direct result of the following lemma.

**Lemma 1.** *The quadratic functional $q : \bar{V} \to \mathbb{R}$ defined by*

$$q(\bar{v}) = \frac{1}{2}\|u - \bar{v}\|_H^2 \tag{6}$$

*is strictly convex.*

*Proof.* Let $\bar{u}, \bar{v} \in \bar{V}$ such that $\bar{u} \neq \bar{v}$. Let $t \in (0,1)$. Define $\bar{w} = t\bar{u} + (1-t)\bar{v}$. We compute

$$\begin{aligned} 2q(\bar{w}) &= \|u - \bar{w}\|_H^2 \\ &= (u - t\bar{u} - (1-t)\bar{v}, u - t\bar{u} - (1-t)\bar{v})_H \\ &= (u,u)_H + t^2(\bar{u},\bar{u})_H + (1-t)^2(\bar{v},\bar{v})_H \\ &\quad -2t(u,\bar{u})_H - 2(1-t)(u,\bar{v})_H + 2t(1-t)(\bar{u},\bar{v})_H. \end{aligned} \tag{7}$$

Note that

$$0 < \|\bar{u} - \bar{v}\|_H^2 = (\bar{u},\bar{u})_H + (\bar{v},\bar{v})_H - 2(\bar{u},\bar{v})_H. \tag{8}$$

Combining (7) with (8), we obtain

$$\begin{aligned} 2q(\bar{w}) &< (u,u)_H + t^2(\bar{u},\bar{u})_H + (1-t)^2(\bar{v},\bar{v})_H \\ &\quad -2t(u,\bar{u})_H - 2(1-t)(u,\bar{v})_H + t(1-t)(\bar{u},\bar{u})_H + t(1-t)(\bar{v},\bar{v})_H \\ &= t\left((u,u)_H - 2(u,\bar{u})_H + (\bar{u},\bar{u})_H\right) \\ &\quad +(1-t)\left((u,u)_H - 2(u,\bar{v})_H + (\bar{v},\bar{v})_H\right) \\ &= t\|u - \bar{u}\|_H^2 + (1-t)\|u - \bar{v}\|_H^2 \\ &= 2tq(\bar{u}) + 2(1-t)q(\bar{v}). \end{aligned} \tag{9}$$

This completes the proof. □

**Theorem 1. (Existence)** *Problem (5) has a solution.*

*Proof.* In a finite-dimensional setting, a strictly convex quadratic functional is bounded from below and attains its greatest lower bound on a nonempty closed set. □

Uniqueness is not as immediate a result as existence. To prove uniqueness, one must impose more structure on the constraint set $\bar{K}$. A reasonable, and powerful, assumption to make is that the constraint set is convex. Then, our constrained optimization problem becomes a *convex optimization problem.* All of the examples presented in this paper involve such constraint sets.

**Theorem 2. (Uniqueness for Convex Constraint Sets)** *Suppose $\boldsymbol{f}$ is an affine function of $\bar{V}$ and $\boldsymbol{g}$ is a convex function of $\bar{V}$. That is, suppose $\bar{K}$ is a convex set.*

*Then problem (5) has a unique solution.*

*Proof.* Suppose that problem (5) has two solutions $\bar{u}, \bar{v} \in \bar{K}$. Since $\bar{K}$ is convex, then $\bar{w} = t\bar{u} + (1-t)\bar{v} \in \bar{K}$ for any $t \in (0,1)$. Further, we have

$$\begin{aligned} \frac{1}{2}\|u - \bar{w}\|_H^2 &< \frac{t}{2}\|u - \bar{u}\|_H^2 + \frac{1-t}{2}\|u - \bar{v}\|_H^2 \\ &= \frac{1}{2}\|u - \bar{u}\|_H^2 \end{aligned} \tag{10}$$

due to Lemma 1. This is a contradiction. Hence, (5) has a unique solution. □

The Karush-Kuhn-Tucker conditions are a set of conditions that a solution to a constrained optimization problem must satisfy (subject to some regularity restrictions). In the context of a convex optimization problem, they are necessary and sufficient. These conditions allow one to characterize a solution to (5) through a set of algebraic relations.

**Theorem 3. (Karush-Kuhn-Tucker Conditions for Convex Constraint Sets)** *Suppose $\bar{K}$ is a convex set. Further, suppose $\bar{K}$ has a strict feasible point. That is, let there exist a point $\bar{v} \in \bar{K}$ such that $\boldsymbol{g}(\bar{v}) < \boldsymbol{0}$. Then $\bar{u} \in \bar{K}$ is the unique solution to (5) if and only if there exist Lagrange multipliers $\lambda_i \in \mathbb{R}$ $(i = 1, \ldots, n_{eq})$ and $\mu_i \in \mathbb{R}$ $(i = 1, \ldots, n_{ineq})$ such that the following conditions hold:*

***Stationary***

$$(\bar{u} - u, \bar{v})_H + \sum_{i=1}^{n_{eq}} \lambda_i \, {}_{\bar{V}^*}\langle Df_i, \bar{v}\rangle_{\bar{V}} + \sum_{i=1}^{n_{ineq}} \mu_i \, {}_{\bar{V}^*}\langle Dg_i(\bar{u}), \bar{v}\rangle_{\bar{V}} = 0 \tag{11}$$

***Primal feasibility***

$$f_i(\bar{u}) = 0, \quad \text{for } i = 1, 2, \ldots, n_{eq} \tag{12}$$

$$g_i(\bar{u}) \leq 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq} \tag{13}$$

***Dual feasibility***

$$\mu_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq} \tag{14}$$

***Complementary slackness***

$$\mu_i g_i(\bar{u}) = 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq} \tag{15}$$

*Above, $\bar{V}^*$ is the dual of $\bar{V}$, ${}_{\bar{V}^*}\langle \cdot, \cdot \rangle_{\bar{V}}$ is the pairing between $\bar{V}$ and its dual, $Df_i \in \bar{V}^*$ is the Fréchet derivative of $f_i$, which is constant, and $Dg_i(\bar{u}) \in \bar{V}^*$ is the Fréchet derivative of $g_i$ at $\bar{u}$.*

*Proof.* This is a classical result in convex optimization. For a discussion of the proof, see [8]. □

Theorem 3 will prove instrumental in the development of VMS formulations which directly enforce equality and inequality constaints. Before proceeding, let us comment that in a particular situation, the Lagrange multipliers associated with Theorem 3 are unique. This occurs when the solution is a *regular point* of the constraint set $\bar{K}$.

**Definition. (Regular Point)** Let $\bar{v} \in \bar{K}$ and let $J$ be the set of indices $j$ for which $g_j(\bar{v}) = 0$. Then $\bar{v}$ is said to be a regular point of the constraint set $\bar{K}$ if the Fréchet derivatives $Df_i(\bar{v}), Dg_j(\bar{v}), 1 \leq i \leq n_{eq}, j \in J$ are linearly independent.

**Theorem 4. (Uniqueness of Lagrange Multipliers for Regular Solutions)** *Suppose $\bar{K}$ is a convex set with a strict feasible point. Further, suppose (5) has a unique solution $\bar{u}$ which is a regular point of the constraint set $\bar{K}$. Then, the Lagrange multipliers associated with Theorem 3 are unique.*

*Proof.* Let $\tilde{J}$ be the set of indices $j$ for which $g_j(\bar{u}) < 0$. Then, $\mu_j = 0$ for $j \in \tilde{J}$ due to (15). The remaining Lagrange multipliers may be uniquely determined from (11) since $\bar{u}$ is a regular point. □

We can repose the stationary condition of Theorem 3 in a form that will allow us to use previous results based on variational projection. Let $\mathcal{P}$ be the projection operator from $V$ onto $\bar{V}$ defined by

$$(\mathcal{P}u, \bar{v})_H = (u, \bar{v})_H, \quad \forall \bar{v} \in \bar{V}. \tag{16}$$

We define $\tilde{V} = \mathrm{Ker}(\mathcal{P})$, which is also a closed subspace of $V$, and notice that

$$V = \bar{V} \oplus \tilde{V}. \tag{17}$$

That is, each $v \in V$ can be written uniquely as $v = \bar{v} + \tilde{v}$ where $\bar{v} \in \bar{V}$ and $\tilde{v} \in \tilde{V}$. Let $\mathcal{R}$ be the Riesz operator from $\bar{V}$ onto $\bar{V}^*$ defined by

$${}_{\bar{V}^*}\langle \mathcal{R}\bar{u}, \bar{v} \rangle_{\bar{V}} = (\bar{u}, \bar{v})_H, \quad \forall \bar{v} \in \bar{V}. \tag{18}$$

We immediately have the following theorem.

**Theorem 5. (Alternative Stationary Condition)** *Let $\bar{K}$ be a convex constraint set with a strict feasible point. The stationary condition of Theorem 3 is equivalent to*

$$\mathcal{P}u' = \mathcal{P}(u - \bar{u}) = \mathcal{P}u - \bar{u} = \sum_{i=1}^{n_{eq}} \lambda_i b_i + \sum_{i=1}^{n_{ineq}} \mu_i c_i(\bar{u}) \tag{19}$$

*where $b_i \in \bar{V}$ is defined as*

$$b_i = \mathcal{R}^{-1} Df_i \tag{20}$$

*and $c_i : \bar{V} \to \bar{V}$ is the nonlinear operator defined by*

$$c_i(\bar{v}) = \mathcal{R}^{-1} Dg_i(\bar{v}), \quad \forall \bar{v} \in \bar{V}. \tag{21}$$

From Theorem 5, we deduce a corollary which describes the structure of the error in terms of the finite-dimensional coarse-scale space $\bar{V}$ and the infinite-dimensional space $\tilde{V}$.

**Corollary 1. (Decomposition of the Error)** *Let $\bar{K}$ be a convex constraint set with a strict feasible point. Then, the error associated with the unique solution of (5), namely, $u' = u - \bar{u}$, can be uniquely decomposed as*

$$u' = \tilde{u}' + \hat{u}' \tag{22}$$

*where $\tilde{u}' \in \tilde{V}$ and $\hat{u}' \in \bar{V}$ satisfies*

$$\hat{u}' = \sum_{i=1}^{n_{eq}} \lambda_i b_i + \sum_{i=1}^{n_{ineq}} \mu_i c_i(\bar{u}), \tag{23}$$

*in which $\lambda_i, \mu_j$ are Lagrange multipliers associated with Theorem 3 and $b_i$ and $c_i$ are defined by (20) and (21) respectively.*

To conclude this subsection, note that Theorem 5 suggests a route towards enforcing the given constraints without modifying the classical VMS method. Since the classical VMS method in theory returns the projection $\mathcal{P}u$, Theorem 5 can be invoked to *postprocess* the numerical solution. That is, once the projection $\mathcal{P}u$ is found, the solution to our constrained optimization problem can be found using the Karush-Kuhn-Tucker conditons utilizing the alternative stationary condition presented in (19) as it does not involve the unknown exact solution $u$. Such a procedure is outlined in Appendix A. This further suggests the advantage of postprocessing solutions obtained using stabilized methods in order to arrive at accurate approximations satisfying the given constraints. Note, however, that postprocessing is not advised in all situations. For example, in a compressible flow code, a negative density value occurring in an iterate of a Newton-Raphson solve will most likely cause a numerical simulation to fail. This implies that a non-negativity condition needs to be enforced throughout the solution process.

**Remark.** In previous works, the notation $\mathrm{Ker}(\mathcal{P}) = V'$ was used as $\mathrm{Ker}(\mathcal{P})$ was associated with the fine-scales. In this paper, the fine-scale component of the solution does not necessarily belong to $\mathrm{Ker}(\mathcal{P})$ so we use the notation $\tilde{V}$ instead.

## 2.3 The variational multiscale formulation

The aim of the VMS approach is to obtain a finite set of equations, independent of $u$, which will allow us to obtain a solution $\bar{u} \in \bar{K}$ to the constrained optimization problem (5). The first step in deriving a VMS formulation is to decompose (2) as follows: find $u \in V$ such that

$$\begin{aligned} {}_{V^*}\langle \mathcal{L}u, \bar{v} \rangle_V &= {}_{V^*}\langle \mathcal{F}, \bar{v} \rangle_V, \quad \forall \bar{v} \in \bar{V}, & (24) \\ {}_{V^*}\langle \mathcal{L}u, \tilde{v} \rangle_V &= {}_{V^*}\langle \mathcal{F}, \tilde{v} \rangle_V, \quad \forall \tilde{v} \in \tilde{V}. & (25) \end{aligned}$$

The problem defined by (24) and (25) is equivalent to the problem defined by (2) and has the same unique solution, $u = \mathcal{G}f$.

The next step in deriving a VMS formulation is to decompose $u$ uniquely into a coarse-scale component $\bar{u} \in \bar{V}$ and a fine-scale component $u'$. In the first instantiation of the variational multiscale method, this was accomplished by enforcing that $u' \in \tilde{V}$. Here, we obtain a unique decomposition, assuming that $\bar{K}$ is convex, by enforcing that $\bar{u}$ is the unique solution to (5). This leads to the following variational formulation.

$$
(\dagger)\quad \left\{
\begin{array}{l}
\text{Find } \bar{u} \in \bar{V}, \tilde{u}' \in \tilde{V}, \lambda_i \in \mathbb{R}\ (i = 1, \ldots, n_{eq}),\ \mu_i \in \mathbb{R}\ (i = 1, \ldots, n_{ineq}) \text{ such that} \\
\qquad {}_{V^*}\langle \mathcal{L}(\bar{u} + \tilde{u}' + \hat{u}'), \bar{v} \rangle_V = {}_{V^*}\langle \mathcal{F}, \bar{v} \rangle_V, \quad \forall \bar{v} \in \bar{V} \qquad (26) \\
\qquad {}_{V^*}\langle \mathcal{L}(\bar{u} + \tilde{u}' + \hat{u}'), \tilde{v} \rangle_V = {}_{V^*}\langle \mathcal{F}, \tilde{v} \rangle_V, \quad \forall \tilde{v} \in \tilde{V} \qquad (27) \\
\qquad f_i(\bar{u}) = 0, \quad \text{for } i = 1, 2, \ldots, n_{eq} \qquad (28) \\
\qquad g_i(\bar{u}) \leq 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq} \qquad (29) \\
\qquad \mu_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq} \qquad (30) \\
\qquad \mu_i g_i(\bar{u}) = 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq} \qquad (31) \\
\text{where } \hat{u}' \in \bar{V} \text{ is defined as} \\
\qquad \hat{u}' = \sum_{i=1}^{n_{eq}} \lambda_i b_i + \sum_{i=1}^{n_{ineq}} \mu_i c_i(\bar{u}) \qquad (32) \\
\text{and } b_i \text{ and } c_i \text{ are defined by (20) and (21) respectively.}
\end{array}
\right.
$$

We now attempt to solve for the fine-scales analytically in terms of the coarse-scales and the Lagrange multipliers so we can arrive at a finite set of equations for the coarse-scales. Suppose we have the inf-sup conditions for $\mathcal{L}$ on $\bar{V}$ and $\tilde{V}$

$$\inf_{\bar{w} \in \bar{V}} \sup_{\bar{v} \in \bar{V}} \frac{{}_{V^*}\langle \mathcal{L}\bar{w}, \bar{v} \rangle_V}{\|\bar{w}\|_V \|\bar{v}\|_V} > 0, \qquad (33)$$

$$\inf_{\tilde{w} \in \tilde{V}} \sup_{\tilde{v} \in \tilde{V}} \frac{{}_{V^*}\langle \mathcal{L}\tilde{w}, \tilde{v} \rangle_V}{\|\tilde{w}\|_V \|\tilde{v}\|_V} > 0, \qquad (34)$$

$$\sup_{\bar{w} \in \bar{V}} \frac{{}_{V^*}\langle \mathcal{L}\bar{w}, \bar{v} \rangle_V}{\|\bar{w}\|_V} > 0, \quad \text{for all } \bar{v} \in \bar{V}, \bar{v} \neq 0, \qquad (35)$$

$$\sup_{\tilde{w} \in \tilde{V}} \frac{{}_{V^*}\langle \mathcal{L}\tilde{w}, \tilde{v} \rangle_V}{\|\tilde{w}\|_V} > 0, \quad \text{for all } \tilde{v} \in \tilde{V}, \tilde{v} \neq 0. \qquad (36)$$

Then, from Theorem 1 of [34], we know we can solve (27) uniquely for $\tilde{u}' \in \tilde{V}$ given $\bar{u}, \hat{u}' \in \bar{V}$ as

$$\tilde{u}' = \mathcal{G}' \left( \mathcal{F} - \mathcal{L}(\bar{u} + \hat{u}') \right) \qquad (37)$$

where $\mathcal{G}' : V^* \to \tilde{V}$ is the *fine-scale Green's operator*

$$\mathcal{G}' = \mathcal{G} - \mathcal{G}\mathcal{P}^T \left( \mathcal{P}\mathcal{G}\mathcal{P}^T \right)^{-1} \mathcal{P}\mathcal{G}. \qquad (38)$$

Substituting (32) and (37) into (22), we find that the entire fine-scale component of the solution is

$$\begin{aligned} u' &= \tilde{u}' + \hat{u}' \\ &= \mathcal{G}'\left(\mathcal{F} - \mathcal{L}(\bar{u} + \hat{u}')\right) + \hat{u}' \\ &= \mathcal{G}'\left(\mathcal{F} - \mathcal{L}(\bar{u})\right) + \sum_{i=1}^{n_{eq}} \lambda_i \left(I - \mathcal{G}'\mathcal{L}\right) b_i + \sum_{i=1}^{n_{ineq}} \mu_i \left(I - \mathcal{G}'\mathcal{L}\right) c_i(\bar{u}). \end{aligned} \tag{39}$$

The first term on the third line of (39) is the representation of the fine-scales in the absence of constraints. The other terms "adjust" the fine-scales so that the coarse-scale solution satisfies the constraints. Substituting (37) into (†) leads to the following variational formulation.

(‡)

Find $\bar{u} \in \bar{V}, \lambda_i \in \mathbb{R}$ $(i = 1, \ldots, n_{eq})$, $\mu_i \in \mathbb{R}$ $(i = 1, \ldots, n_{ineq})$ such that

$$ {}_{V^*}\langle \mathcal{L}(\bar{u} + \hat{u}') - \mathcal{L}\mathcal{G}'\mathcal{L}(\bar{u} + \hat{u}'), \bar{v} \rangle_V = {}_{V^*}\langle \mathcal{F} - \mathcal{L}\mathcal{G}'\mathcal{F}, \bar{v} \rangle_V, \quad \forall \bar{v} \in \bar{V} \tag{40}$$

$$f_i(\bar{u}) = 0, \quad \text{for } i = 1, 2, \ldots, n_{eq} \tag{41}$$

$$g_i(\bar{u}) \leq 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq} \tag{42}$$

$$\mu_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq} \tag{43}$$

$$\mu_i g_i(\bar{u}) = 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq} \tag{44}$$

where $\hat{u}' \in \bar{V}$ is defined as

$$\hat{u}' = \sum_{i=1}^{n_{eq}} \lambda_i b_i + \sum_{i=1}^{n_{ineq}} \mu_j c_i(\bar{u}). \tag{45}$$

and $b_i$ and $c_i$ are defined by (20) and (21) respectively.

We immediately have the following theorem.

**Theorem 6.** *Let $\bar{K}$ be a convex set with a strict feasible point. Further, suppose the inf-sup conditons (33)-(36) hold. Then problem (‡) has a solution $\bar{u} \in \bar{V}, \lambda_i \in \mathbb{R}$ $(i = 1, \ldots, n_{eq})$, $\mu_i \in \mathbb{R}$ $(i = 1, \ldots, n_{ineq})$. Further, $\bar{u}$ is uniquely determined and is the unique solution to (5) with $u = \mathcal{G}\mathcal{F}$, and the Lagrange multipliers $\lambda_i, \mu_i$ are unique if $\bar{u}$ is a regular point of $\bar{K}$.*

The variational formulation (‡) defines a new class of VMS methods which allow for the direct enforcement of equality and inequality constraints. It consists of a finite set of equations for the coarse-scale solution $\bar{u}$ and Lagrange multipliers associated with the constraints. Theorem 6 guarantees that this formulation is well-posed under fairly modest conditions. The formulation can be solved using either an active set method or interior point method in the presence of inequality constraints.

Note that the values of $b_i$ and the $c_i(\bar{u})$ can be realistically computed due to the finite-dimensional nature of the constraints. Specifically, we can evaluate (20) and

(21) by a linear solve involving a symmetric positive-definite matrix. This is outlined in Appendix B. Thus, in practical applications, only the fine-scale Green's operator need be approximated in the above formulation. This is a definitive advantage of our approach and suggests a path for constructing new stabilized methods which enforce important constraints such as positivity. In fact, by recognizing the relationship between stabilized methods and variational multiscale analysis (see [34]), this implies that classical stabilized methods such as SUPG [10] can be easily modified in order to enforce equality and inequality constraints without sacrificing accuracy.

# 3 The steady advection-diffusion problem

## 3.1 Problem description

Let $\Omega \subset \mathbb{R}^2$ be a regular domain. Let $\partial\Omega$ be the boundary of $\Omega$. We consider the steady advection-diffusion problem with homogeneous Dirichlet boundary conditions. The strong formulation of the problem is as follows.

$$\text{(S)} \begin{cases} \text{Find } u : \overline{\Omega} \to \mathbb{R} \text{ such that} \\ \qquad -\kappa\Delta u + \beta \cdot \nabla u = f \text{ in } \Omega, \quad \text{with } u|_{\partial\Omega} = 0 \qquad (46) \\ \text{where } f : \Omega \to \mathbb{R} \text{ is the source term, } \kappa > 0 \text{ is the scalar diffusivity, and } \beta : \overline{\Omega} \to \\ \mathbb{R}^2 \text{ is the advection velocity, for which we assume } \nabla \cdot \beta = 0. \end{cases}$$

We assume that the specified data is sufficiently smooth such that (S) is well-defined. The corresponding variational formulation is the following.

$$\text{(V)} \begin{cases} \text{Find } u \in H_0^1(\Omega) \text{ such that} \\ \qquad B(u,v) = F(v), \ \ \forall v \in H_0^1(\Omega) \qquad (47) \\ \text{where} \\ \qquad B(u,v) = (\kappa\nabla u - \beta u, \nabla v)_{L^2(\Omega)}, \qquad (48) \\ \qquad F(v) = {}_{H^{-1}(\Omega)}\langle f, v\rangle_{H_0^1(\Omega)}. \qquad (49) \end{cases}$$

We see the above formulation fits within the framework of Section 2 with $V = H_0^1(\Omega)$, $V^* = H^{-1}(\Omega)$, $\mathcal{L} : H_0^1(\Omega) \to H^{-1}(\Omega)$ defined by

$${}_{H^{-1}(\Omega)}\langle \mathcal{L}u, v\rangle_{H_0^1(\Omega)} = B(u,v), \ \ \forall u, v \in H_0^1(\Omega) \qquad (50)$$

or equivalently,

$$\mathcal{L} = -\kappa\Delta + \beta \cdot \nabla, \qquad (51)$$

and $\mathcal{F} \in H^{-1}(\Omega)$ defined by

$${}_{H^{-1}(\Omega)}\langle \mathcal{F}, v\rangle_{H_0^1(\Omega)} = F(v), \ \ \forall v \in H_0^1(\Omega). \qquad (52)$$

The invertibility of $\mathcal{L}$ is due to the coercivity of $B(\cdot,\cdot)$. In this context, it is convenient to represent the Green's operator $\mathcal{G}$ through the Green's function $g : \Omega \times \Omega \to \mathbb{R}$ such that

$$Gr(y) = \int_\Omega g(x,y) r(x) dx. \qquad (53)$$

for all $r \in H^{-1}(\Omega)$. We have that $g|_{\partial(\Omega\times\Omega)} = 0$ and for all $y \in \Omega$, $\mathcal{L}^* g(\cdot, y) = \delta(\cdot - y)$, where $\delta$ is the Dirac mass at the origin and $\mathcal{L}^*$ denotes the dual of $\mathcal{L}$. We immediately see that the solution of (V) is

$$\begin{aligned} u(y) &= \int_\Omega g(x,y)\mathcal{F}(x)dx \\ &= \int_\Omega g(x,y)f(x)dx. \end{aligned} \tag{54}$$

Note that here and in what follows, integrals are to be understood in the sense of distributions.

**Remark.** We will subsequently, as above, focus attention on the case of homogeneous boundary conditions, but a problem with prescribed, nonzero boundary values can also easily be transformed into this setting. Suppose that $u \in H^1(\Omega)$ is a weak solution of

$$\left\{ \begin{array}{rcll} \mathcal{L}u &=& f, & \text{in } \Omega, \\ u &=& g, & \text{on } \partial\Omega. \end{array} \right.$$

Choosing a $w \in H^1(\Omega)$ for which $w = g$ on $\partial\Omega$, we have that $\tilde{u} := u - w \in H_0^1(\Omega)$ is a weak solution to the problem

$$\left\{ \begin{array}{rcll} \mathcal{L}\tilde{u} &=& \tilde{f}, & \text{in } \Omega, \\ \tilde{u} &=& 0, & \text{on } \partial\Omega. \end{array} \right.$$

where $\tilde{f} = f - \mathcal{L}w \in H^{-1}(\Omega)$.

## 3.2 The maximum principle and non-negativity constraint

Since the differential operator associated with the steady advection-diffusion problem is elliptic, we have the following maximum principle.

**Theorem 7. (Maximum principle)** *Suppose $u \in C^2(\Omega) \cap C(\bar{\Omega})$.*

1. *If $\mathcal{L}u \leq 0$ in $\Omega$, then*
$$\max_{\bar{\Omega}} u = \max_{\partial\Omega} u. \tag{55}$$
2. *If $\mathcal{L}u \geq 0$ in $\Omega$, then*
$$\min_{\bar{\Omega}} u = \min_{\partial\Omega} u. \tag{56}$$

*Proof.* The result is classical. For a discussion of the proof, see Chapter 6 of [22]. □.

Note that in the presence of non-negative forcing, the maximum principle provides a non-negativity constraint for the solution to the steady advection-diffusion equation with homogeneous boundary conditions.

**Corollary 2. (Non-negativity constraint)** *Suppose $u \in C^2(\Omega) \cap C(\bar{\Omega})$ is a strong solution to (S). Further, suppose $f \geq 0$ in $\Omega$. Then:*

$$\min_{\bar{\Omega}} u = 0. \tag{57}$$

The maximum principle has been a very powerful tool as it has allowed theoreticians to obtain information about solutions to differential equations without any explicit knowledge of the solutions themselves. While the maximum principle holds analytically, it is not necessarily true that a numerical approximation obeys this property. From a physical standpoint, though, it is important that this principle holds in the discrete setting as violations of this principle may lead to unphysical negative densities, temperatures, concentrations, or electric charges. However, not even numerical solutions to self-adjoint elliptic boundary value problems obtained using the Galerkin finite element method are guaranteed to satisfy the maximum principle. Consequently, much work has been done to understand when numerical solutions satisfy or violate the maximum principle, a discrete equivalent known as the *discrete maximum principle*, or weaker versions of these principles [12, 13, 18, 25, 43, 48], and further work has been done to develop new methodologies which automatically preserve a discrete maximum principle [2, 11, 17, 19, 36, 37, 39, 40, 41, 42, 45, 47]. Of particular interest is a recent paper by Liska and Shashkov [40] in which a method based on constrained optimization is developed to arrive at a numerical solution to self-adjoint elliptic problems satisfying a discrete maximum principle. In fact, this technique is identically a variational multiscale method directly enforcing the maximum principle in the context of a projector based on the energy norm.

We find that, indeed, numerical approximations obtained using the classical VMS method may not preserve the maximum principle because a linear projection is not guaranteed to preserve a maximum principle. Because of this, we will introduce a VMS method in the next subsection in which the maximum principle is automatically enforced.

**Remark.** The piecewise bilinear nodal interpolant of the solution to the advection-diffusion equation is guaranteed to preserve the maximum principle. However, one notes the nodal interpolant is not a true variational projection as it is not continuous in $H_0^1(\Omega)$.

## 3.3 The classical and constrained variational multiscale methods

Let $\bar{V} \subset H_0^1(\Omega)$ be a finite-dimensional space. The space $\bar{V}$ is simultaneously the space of weighting functions and the set of trial solutions. Let us introduce the fine-scale Green's function $g' : \Omega \times \Omega \to \mathbb{R}$, which represents the fine-scale Green's operator $\mathcal{G}'$ and satisfies

$$\mathcal{G}'r(y) = \int_\Omega g'(x,y)r(x)dx \tag{58}$$

for all $r \in H^{-1}(\Omega)$. Since the bilinear form $B(\cdot,\cdot)$ is coercive, the inf-sup conditons (33)-(36) hold and thus the above fine-scale Green's function is well-defined. The classical VMS method can be stated as follows.

*Classical variational multiscale method*:

Find $\bar{u} \in \bar{V}$ such that

$$\int_\Omega (\kappa \nabla \bar{u}(x) - \beta(x)\bar{u}(x)) \cdot \nabla \bar{v}(x) dx - \int_\Omega \int_\Omega \mathcal{L}\bar{u}(x) g'(x,y) \mathcal{L}^* \bar{v}(y) dy dx = \int_\Omega f(x)\bar{v}(x) dx - \int_\Omega \int_\Omega f(x) g'(x,y) \mathcal{L}^* \bar{v}(y) dy dx \tag{59}$$

for all $\bar{v} \in \bar{V}$.

To define the constrained VMS method, we need a discrete analogue of the maximum principle. That is, we need a finite set of inequality constraints that is equivalent to (55) and (56) for the space of functions $\bar{V}$. This is a very difficult thing to do in general. However, if the space $\bar{V}$ corresponds to piecewise bilinear finite element functions satisfying homogeneous boundary conditions, then it is sufficient to require that the nodal values satisfy a *discrete maximum principle*.

Let a piecewise bilinear function $\bar{v}$ have the representation

$$\bar{v}(x) = \sum_{i=1}^{n_{nd}} \bar{v}_i N_i(x), \tag{60}$$

where $\bar{v}_i$ is the nodal value of $\bar{v}$ at $x_i$ and $N_i(x)$ is the piecewise bilinear finite element function satisfying

$$N_i(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{61}$$

We immediately have the following theorem.

**Theorem 8. (Discrete maximum principle)** *Let $\bar{v}$ have the form of (60). Then:*

1. *The function $\bar{v}$ satisfies*
$$\max_{\bar{\Omega}} \bar{v} = \max_{\partial\Omega} \bar{v} \tag{62}$$
*if and only if*
$$\max_i \bar{v}_i = \max_{\partial\Omega} \bar{v}. \tag{63}$$
2. *The function $\bar{v}$ satisfies*
$$\min_{\bar{\Omega}} \bar{v} = \min_{\partial\Omega} \bar{v} \tag{64}$$
*if and only if*
$$\min_i \bar{v}_i = \min_{\partial\Omega} \bar{v}. \tag{65}$$

*Proof.* A piecewise bilinear finite element function attains its minima and maxima at nodes. □

Theorem 8 suggests the following representation for our constrained variational formulation which imposes a discrete maximum principle on the numerical solution $\bar{u}$ in the case when $\bar{V} \subset V$ is a space of piecewise bilinear finite element functions satisfying homogeneous boundary conditions.

*Constrained variational multiscale method for bilinear finite elements*:
Find $\bar{u} \in \bar{V}, \mu_{i,1} \in \mathbb{R}, \mu_{i,2} \in \mathbb{R}$ such that

$$\begin{aligned}\int_\Omega \left(\kappa \nabla \left(\bar{u}(x) + \hat{u}'(x)\right) - \beta(x)\left(\bar{u}(x) + \hat{u}'(x)\right)\right) \cdot \nabla \bar{v}(x) dx \\ - \int_\Omega \int_\Omega \mathcal{L}\left(\bar{u}(x) + \hat{u}'(x)\right) g'(x,y) \mathcal{L}^* \bar{v}(y) dy dx \\ = \int_\Omega f(x)\bar{v}(x)dx - \int_\Omega \int_\Omega f(x) g'(x,y) \mathcal{L}^* \bar{v}(y) dy dx \end{aligned} \tag{66}$$

for all $\bar{v} \in \bar{V}$,

$$\bar{u}_i \leq 0, \text{ if } f \leq 0 \text{ in } \Omega, \ i = 1, \ldots, n_{nd}, \tag{67}$$

$$\bar{u}_i \geq 0, \text{ if } f \geq 0 \text{ in } \Omega, \ i = 1, \ldots, n_{nd}, \tag{68}$$

$$\mu_{i,1} \geq 0, \quad \text{for } i = 1, 2, \ldots, n_{nd} \tag{69}$$

$$\mu_{i,2} \geq 0, \quad \text{for } i = 1, 2, \ldots, n_{nd} \tag{70}$$

$$\mu_{i,1}\bar{u}_i = 0, \quad i = 1, \ldots, n_{nd}, \tag{71}$$

$$\mu_{i,2}\bar{u}_i = 0, \ i = 1, \ldots, n_{nd}. \tag{72}$$

$$\mu_{i,1} = 0 \text{ if } f > 0 \text{ in } \Omega \text{ or } x_i \in \partial\Omega, \tag{73}$$

$$\mu_{i,2} = 0 \text{ if } f < 0 \text{ in } \Omega \text{ or } x_i \in \partial\Omega, \tag{74}$$

and

$$\hat{u}'(x) = \sum_{i=1}^{n_{nd}} \mu_{i,1} B_i(x) - \sum_{i=1}^{n_{nd}} \mu_{i,2} B_i(x), \tag{75}$$

where the functions $B_i \in \bar{V}$ satisfy

$$(B_i, \bar{v})_H = \bar{v}(x_i) = \bar{v}_i \tag{76}$$

for all $\bar{v} \in \bar{V}$.

Recall, the functions $B_i(x)$ in the above formulation can be easily computed by soliving a linear system with a positive-definite mass matrix as illustrated in Appendix B. It is easy to show that Theorems 3 and 4 can be used to infer that the above constrained variational multiscale method has a unique solution $\bar{u}$ and the Lagrange multipliers $\mu_{i,1}, \mu_{i,2}$ are unique. This follows from the structure and convexity of our constraints.

To arrive at a computationally tractable numerical scheme, one must approximate the fine-scale Green's function appearing in the classical and constrained VMS methods. This is traditionally done with a highly-localized approximation as it is obviously more convenient and easy. In the next subsection, it will be shown that with the correct projector, the fine-scale Green's function is indeed attenuated. This is explored in greater detail in [34].

Finally, before proceeding, let us note that while the constrained VMS formulation above seems vastly more complicated than the unconstrained, classical VMS formulation, one can instead post-process the classical VMS solution to arrive at a solution to the constrained formulation. This is done using the procedure detailed in Appendix A.

**Remark 1.** The constrained variational multiscale method can be trivially modified to account for nonhomogeneous Dirichlet boundary conditions. In this case, the set of trail solutions is replaced with a set $\bar{S}$ satisfying nonhomogeneous boundary conditons, and the inequality constraints are modified to enforce appropriate maximum principles.

**Remark 2.** The constrained variational multiscale method can be easily extended to the case of *isogeometric analysis*. In isogeometric analysis, the solution field is represented in terms of B-spline, NURBS, or T-spline basis functions. For such functions, the maximum principle is satisfied if a discrete maximum principle is satisfied by the *control variables*. This is due to the *convex hull property*. As $C^0$ finite elements can be represented in terms of isogeometric *Bézier* basis functions, this provides an avenue for enforcing maximum principle for higher-order $C^0$ elements. For an introduction to NURBS-based isogeometric analysis, see [28], and for an extension of this technique to T-splines, see [3].

## 3.4 The fine-scale Green's function

One can find the fine-scale Green's function by solving

$$\int_\Omega \mathcal{L}^* g'(x,y)\tilde{v}(x)dx = \int_\Omega \delta(x-y)\tilde{v}(x)dx, \;\; \forall \tilde{v} \in \tilde{V} \tag{77}$$

in mixed form for $g'(x,y) \in \tilde{V}$ for each fixed $y \in \Omega$. The structure of the fine-scale Green's function was examined in depth in [34] for a number of choices for finite element spaces $\bar{V}$ and variational projectors $\mathcal{P}$. Here, we review the structure of the fine-scale Green's functions for two chosen projectors.

Let $\Omega = (0,1)^2$, $\kappa = 10^{-3}$, and $\beta = (1/2, 1)/\sqrt{1.25}$. We have numerically approximated the fine-scale Green's functions for the $L^2$-norm and $H^1_0$-seminorm for a piecewise-bilinear finite element space defined on a mesh of 20 x 20 elements. These approximations were obtained by solving (77) in mixed form. Since both $g$ and $g'$ are defined on $(0,1)^2 \times (0,1)^2$, for the purpose of graphical representation we fix $y = y^* = (0.775, 0.775)$ (see Figure 1) and we plot the Green's function $g$ and fine-scale Green's function $g'$ as a function of $x$. Changing this position produces similar results. The numerical approximations to $g$ and $g'$ were computed using Galerkin's method on a fine mesh of 500 x 500 elements, which is able to resolve the solution of the problem

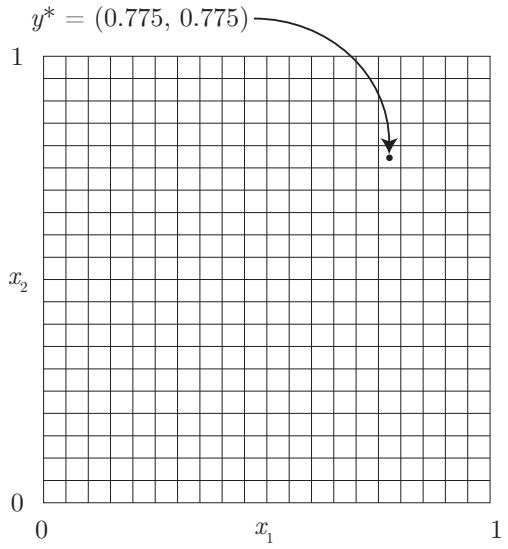



Figure 1: Mesh of the space $\bar{V}$ used for the calculation of the fine-scale Green's functions for the steady advection-diffusion problem and location of $y^* = (0.775, 0.775)$.

sufficiently for our purposes. A contour plot of $x \mapsto g(x, y^*)$ is presented in Figure 2(a). Note that $x \mapsto g(x, y^*)$ is singular when $x = y^*$, and as such, the graph in Figure 2(a) has been truncated at $g = 5$. Roughly speaking, $g$ has support around the upwind tail through $y^*$.

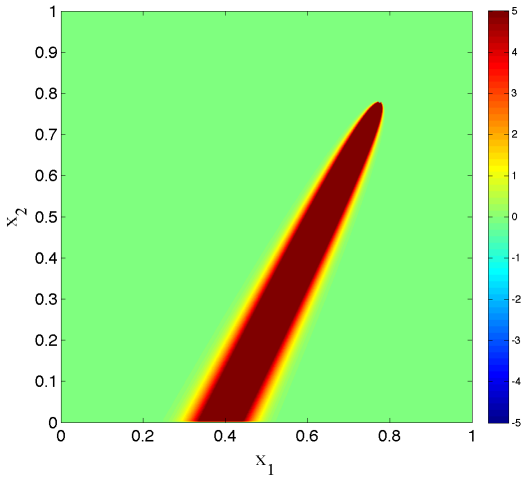
Contour plots of $x \mapsto g'(x, y^*)$ for the $H_0^1$-projector $\mathcal{P}_{H_0^1}$ and the $L^2$-projector $\mathcal{P}_{L^2}$ are presented in Figure 2(b) and 2(c) respectively. The singularity at $x = y^*$ is still truncated. Observe that the fine-scale Green's function for $\mathcal{P}_{H_0^1}$ is more localized about $y^*$ than the fine-scale Green's function for $\mathcal{P}_{L^2}$. In fact, the fine-scale Green's function for $\mathcal{P}_{H_0^1}$ seems to be negligible outside a small neighborhood of elements around $y^*$. This is the same story as was told in [34] and once again suggests that local approximations of $g'$ for $\mathcal{P} = \mathcal{P}_{H_0^1}$ may achieve near $H_0^1$-optimality for the steady advection-diffusion problem.

## 3.5 Two example problems

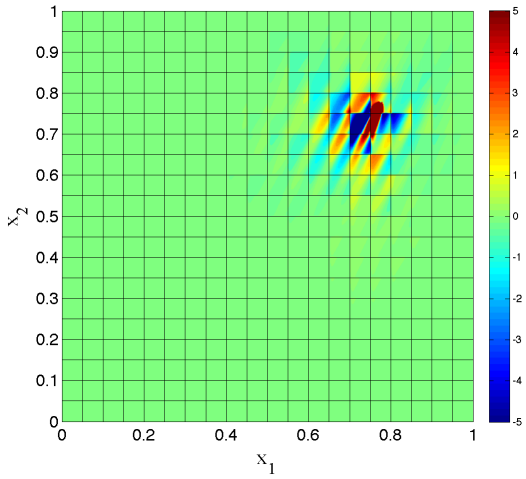
We consider two example problems. The setup for the first example problem is illustrated in Figure 3. For this problem, the scalar diffusivity is set to $\kappa = 10^{-3}$, the velocity field is set to $\beta = (1, 1/2)/\sqrt{1.25}$, and homogeneous boundary conditions are enforced. The forcing function is taken to be

$$f(x_1, x_2) = \begin{cases} 2, & \text{if } x_1 < \min(2x_2, 0.5), \\ 0, & \text{otherwise.} \end{cases} \tag{78}$$

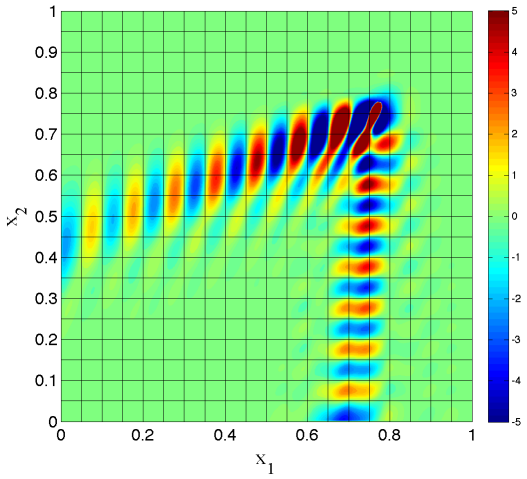
The problem is characterized by an interior layer at $x_1 = 2x_2$ and boundary layers at $x_1, x_2 = 1$. The exact solution is plotted in Figure 4. Since the forcing is every-




(a)




(b)




(c)

Figure 2: Green's function and fine-scale Green's functions for the bilinear mesh.
(a) Contour plot of $x \mapsto g(x, y^*)$.
(b) Contour plot of $x \mapsto g'(x, y^*)$ for $\mathcal{P} = \mathcal{P}_{H_0^1}$ and the associated coarse mesh.
(c) Contour plot of $x \mapsto g'(x, y^*)$ for $\mathcal{P} = \mathcal{P}_{L^2}$ and the associated coarse mesh.

where non-negative, the solution satisfies a non-negativity constraint. We enforce this condition in our constrained VMS method.

We consider a coarse-scale space of bilinear finite element functions corresponding to the mesh in Figure 1. The coarse-scale approximation $\bar{u}$ obtained from the classical VMS method is plotted in Figures 5(a) and (c) for the cases of $H_0^1$- and $L^2$-optimality. Images detailing violations of the non-negativity constraint are included in Figures 5(b) and (d). From these four figures, we see that both approximations suffer from overshoots and undershoots about the interior and boundary layers, and it is obvious that the solution obtained from the $H_0^1$-projector is much better than that obtained from the $L^2$-projector. In fact, the solution obtained from the $H_0^1$-projector almost satisfies the non-negativity constraint. Indeed, the minimum of this solution is approximately $-0.015$. If one further magnifies the blown-up image in Figure 5(d), one finds that the solution obtained from the $L^2$-projector suffers from additional overshoots in the region where the exact solution is zero. These overshoots are a direct result of the $L^2$-projector compensating for undershoots about the interior layer.

In Figure 5(e), we have plotted the coarse-scale approximation obtained from the so-called VMS-$\tau$ method, originally referred to as the unusual stabilized finite element method [23]. In this method, the fine-scale Green's function obtained from the $H_0^1$-projector is replaced by a highly-localized algebraic approximation. Specifically, one writes

$$-\sum_{e=1}^{n_{el}} \tau_e \int_{\Omega_e} r_1(y) r_2(y) dy \approx \int_\Omega \int_\Omega r_1(y) g'(x,y) r_2(x) dx dy \tag{79}$$

for all $r_1, r_2 \in \bar{V}^*$, with

$$\tau_e = \frac{h_e}{2|\beta|} \tag{80}$$

where $h_e$ is the length of element $\Omega_e$ in direction $\beta$. We note this definition of $\tau$ is valid only for the advective limit. For the case of rectangular bilinear finite elements, the VMS-$\tau$ method is identical to SUPG. In Figure 5(f), we have, as in the case of the classical VMS methods, presented an image detailing violations of the non-negativity constraint.

In Figures 6(a)-(d), we have plotted the coarse-scale approximations $\bar{u}$ obtained from the $H_0^1$- and $L^2$-optimal constrained VMS methods enforcing a non-negativity constraint. We see that both of our approximations are everywhere non-negative. In addition, if we magnify the blown-up image in Figure 6(d) further we find that the $L^2$-optimal solution no longer suffers from additional overshoots in the region where the exact solution is zero. However, the solution still suffers from severe overshoots about the boundary layers and minor overshoots about the interior layer. In Figures 6(e)-(f), we have plotted the solution obtained from the VMS-$\tau$ method subject to a non-negativity constraint. This method is identical to the $H_0^1$-optimal constrained VMS method with the exception that the fine-scale Green's function is approximated. The VMS-$\tau$ solution is everywhere non-negative but also suffers from severe overshoots about one of its boundary layers.

To eliminate the remaining overshoots occurring in the constrained $L^2$-optimal and VMS-$\tau$ coarse-scale approximations, one must introduce extra constraints or mecha-

nisms. In the next subsection, we will discuss possible approaches including enforcing a global conservation constraint and applying weak boundary conditions.

The setup for the second example problem is illustrated in Figure 7. For this problem, the scalar diffusivity is again set to $\kappa = 10^{-3}$, the velocity field is again set to $\beta = (1, 1/2)/\sqrt{1.25}$, the forcing is set to zero, and nonhomogeneous boundary conditions are enforced. Specifically, the boundary condition is taken to be 1 along the red lines and 0 along the rest of the boundary. This problem is a standard one in testing the robustness and accuracy of a numerical method for the advection-diffusion problem. The solution to the problem is characterized by boundary layers at $x_1, x_2 = 1$ and an interior layer at $x_1 = 2x_2 + 0.2$.

The exact solution to this problem does not lie in $H^1(\Omega)$ since the boundary condition is discontinuous. In addition, standard $C^0$ finite elements cannot represent the given boundary condition. As such, we consider modified "ramp" boundary conditions for which a set of bilinear finite element functions corresponding to a uniform 20 x 20 mesh can replicate. The effective boundary conditions are then

$$u(x_1, x_2) = \begin{cases} 0, & \text{if } x_1 = 1, x_2 = 1, \text{ or } x_1 \geq 0.2 \text{ and } x_2 = 0, \\ 1, & \text{if } x_1 = 0 \text{ and } x_2 \leq 0.95 = 0 \text{ or } x_1 \leq 0.15 \text{ and } x_2 = 0, \\ 4 - 20x_1, & \text{if } 0.15 \leq x_1 \leq 0.2 \text{ and } x_2 = 0, \\ 20 - 20x_2, & \text{if } x_1 = 0 \text{ and } x_2 \geq 0.95. \end{cases} \tag{81}$$

The exact solution for these "ramp" boundary conditions is given in Figure (8). Since the forcing is zero, the solution inside the domain is bounded above by 1 and below by 0. We enforce our numerical approximation to lie within these bounds in our constrained VMS method.

As with the first example, we consider a coarse-scale space of bilinear finite element functions corresponding to a uniform 20 x 20 mesh. The coarse-scale approximation $\bar{u}$ obtained from the classical VMS method is plotted in Figures 9(a) and (c) for the cases of $H_0^1$- and $L^2$-optimality. We see that both approximations violate the maximum principle about the interior and boundary layers, and the approximation induced by the $H_0^1$-projector is again much better than that obtained from the $L^2$-projector. The coarse-scale approximation $\bar{u}$ obtained from the constrained VMS method is plotted in Figures 9(b) and (d). We see that these approximations are far superior to the approximations obtained from classical VMS in an aesthetic sense (i.e., in the elusive "eyeball" norm), that these approximations are nearly monotone, and that the $H_0^1$- and $L^2$-optimal solutions are nearly indistinguishable. (There is a small dip in the $H_0^1$-optimal solution near the point $x = (0.15, 0.05)$ that is not discernible in Figure 9(b). This is why we use the modifier "nearly.")

In Figure 9(e), we have plotted the coarse-scale approximation obtained from the unconstrained VMS-$\tau$ method, and in Figure 9(f), we have plotted the solution obtained from the VMS-$\tau$ method subject to a constraint enforcing the discrete maximum principle. We see that the constrained approximation is monotone and superior to the unconstrained approximation in the eyeball norm. The constrained approximation, however, does suffer from a ragged profile in the direction tangential to the boundary along the layer located at $x_1 = 1$. This is reminiscent of the "stair-casing" behavior that plagues flux-corrected transport schemes in computational fluid dynamics [44].

### 3.6 Conservation considerations and weak boundary conditions

As illustrated by Figures 6(c) and (e), the direct enforcement of a discrete maximum principle in VMS may not, by itself, result in a satisfactory coarse-scale approximation. The constraint may not ensure a monotone solution free of unphysical oscillations. Indeed, finding a proper set of constraints that guarantees monotonicity is still an open research area, and a precise mathematical definition of monotonicity does not yet exist. In [21], the authors attempted to obtain a monotone solution to the steady advection-diffusion equation utilizing a total variation constraint. The constraint, however, involves a constant whose exact value is not known in the steady case. A total variation bound for the unsteady case, however, is explicitly known and is the basis of a forthcoming paper [20].

To handle the remaining overshoots in Figures 6(c) and (e), we must introduce additional mechanisms. To begin, note that the addition of a non-negativity constraint in our VMS method results in the distribution of sources throughout our domain. The presence of these sources inadvertently results in violation of conservation in the sense that the coarse-scale solution $\bar{u}$ of the constrained VMS method satisfies

$$\int_\Omega \bar{u} dx \geq \int_\Omega \mathcal{P}u dx \tag{82}$$

rather than

$$\int_\Omega \bar{u} dx = \int_\Omega \mathcal{P}u dx. \tag{83}$$

That is, the constrained VMS method does not globally conserve $\bar{u}$ in constrast with the classical VMS method. It is desirable to recover equality to arrive at a *conservation* principle. This can be done through the enforcement of an additional constraint, namely, (83), in our constrained VMS method. The enforcement of this constraint tends to compensate for the non-negativity constraint. Since violations of the non-negativity constraint are often accompanied by additional overshoots, it may be hoped that the conservation constraint has the effect of removing these overshoots.

One finds that enforcing a conservation principle does lessen the undesirable overshoots associated with Figures 6(c) and (e) but does not remove them entirely because the projected solution $\mathcal{P}u$ suffers from overshoots about the boundary layer without accompanying undershoots. To remove the overshoots associated with the boundary layer, one can employ weak boundary conditions [5, 6]. In previous works, this has been found to be a very promising approach to dealing with sharp boundary layers associated with prescribed Dirichlet boundary conditions. In Figure 10, we have plotted the coarse-scale approximation obtained from a VMS-$\tau$ method utilizing weakly enforced boundary conditions and direct enforcement of the non-negativity and conservation constraints. The conservation constraint is satisfied in the sense that

$$\int_\Omega \bar{u} dx = \int_\Omega u_{\text{VMS}-\tau} dx \tag{84}$$

where $u_{\text{VMS}-\tau}$ is the unconstrained VMS-$\tau$ solution. From Figure 10, we see that the constrained approximation with weakly enforced boundary conditions is monotone and

superior in quality to the VMS-$\tau$ solutions illustrated in Figures 5(e) and 6(e) in terms of the eyeball norm. This suggests an effective four-part recipe for solving the steady advection-diffusion equation: (i) utilization of a VMS-inspired stabilized method, (ii) weakly enforced boundary conditions, (iii) direct enforcement of the discrete maximum principle, and (iv) direct enforcement of a conservation constraint.

# 4 Conclusions

In this paper, we derived a new theoretical framework for the enforcement of equality and inequality constraints in variational multiscale analysis. The theory is first presented in an abstract operator format. The decomposition of the exact solution into its coarse-scale and fine-scale components is defined through a constrained optimization problem. In the case of convex constraints, this problem is found to be well-posed. An exact expression for the fine-scales is derived in terms of variational derivatives of the constraints, Lagrange multipliers, and a fine-scale Green's function. Using this framework, we developed a variational multiscale method for the steady advection-diffusion problem enforcing a discrete maximum principle. Numerical results illustrated the promise of such a multiscale technology. When coupled with weak boundary conditions and the direct enforcement of an additional conservation constraint, the method was found to produce accurate and monotone solutions.

This work raises important questions regarding solution quality: what exactly are the constraints that one wants to enforce on a numerical solution, and to what degree is one willing to trade off solution accuracy and computational ease in order to satisfy them? Once one's goals are delineated, the present variational multiscale formulation enables the development of a numerical method that achieves them. We believe this framework can be utilized in a rigorous way to develop numerical methods with predefined attributes and provide satisfactory solutions in the critical, but elusive, eyeball norm.

# Acknowledgements

J.A. Evans was partially supported by the Department of Energy Computational Science Graduate Fellowship, provided under grant number DE-FG02-97ER25308. T.J.R. Hughes was partially supported by the Office of Naval Research under Contract Number N00014-08-0992, by the Boeing Company under Contract Number 20010, and by the Department of Energy [National Nuclear Security Administration] under Award Number DE-FC52-08NA28615. This support is gratefully acknowledged.

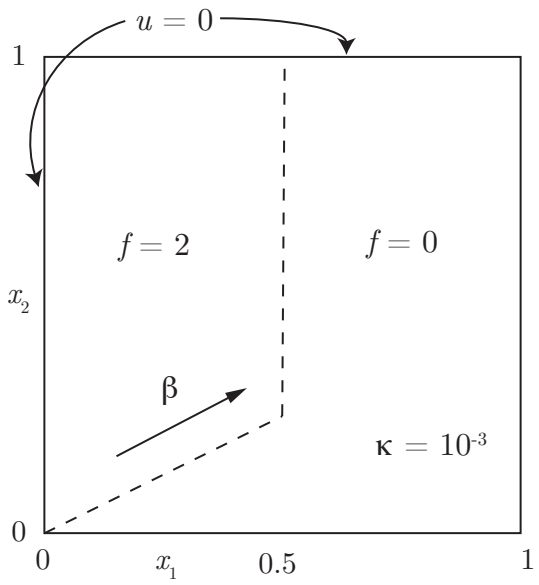



Figure 3: Problem description for the first steady advection-diffusion example problem.

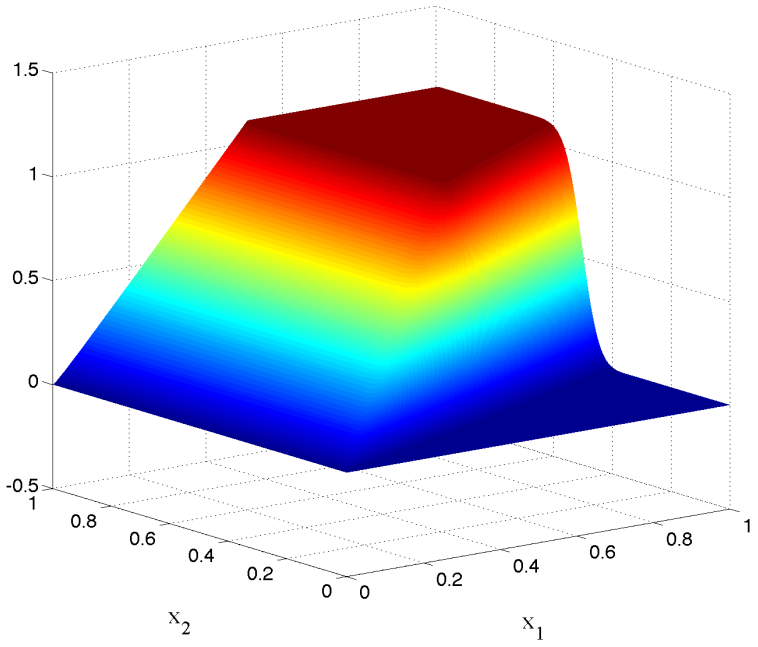



Figure 4: Exact solution of the example problem illustrated in Figure 3.

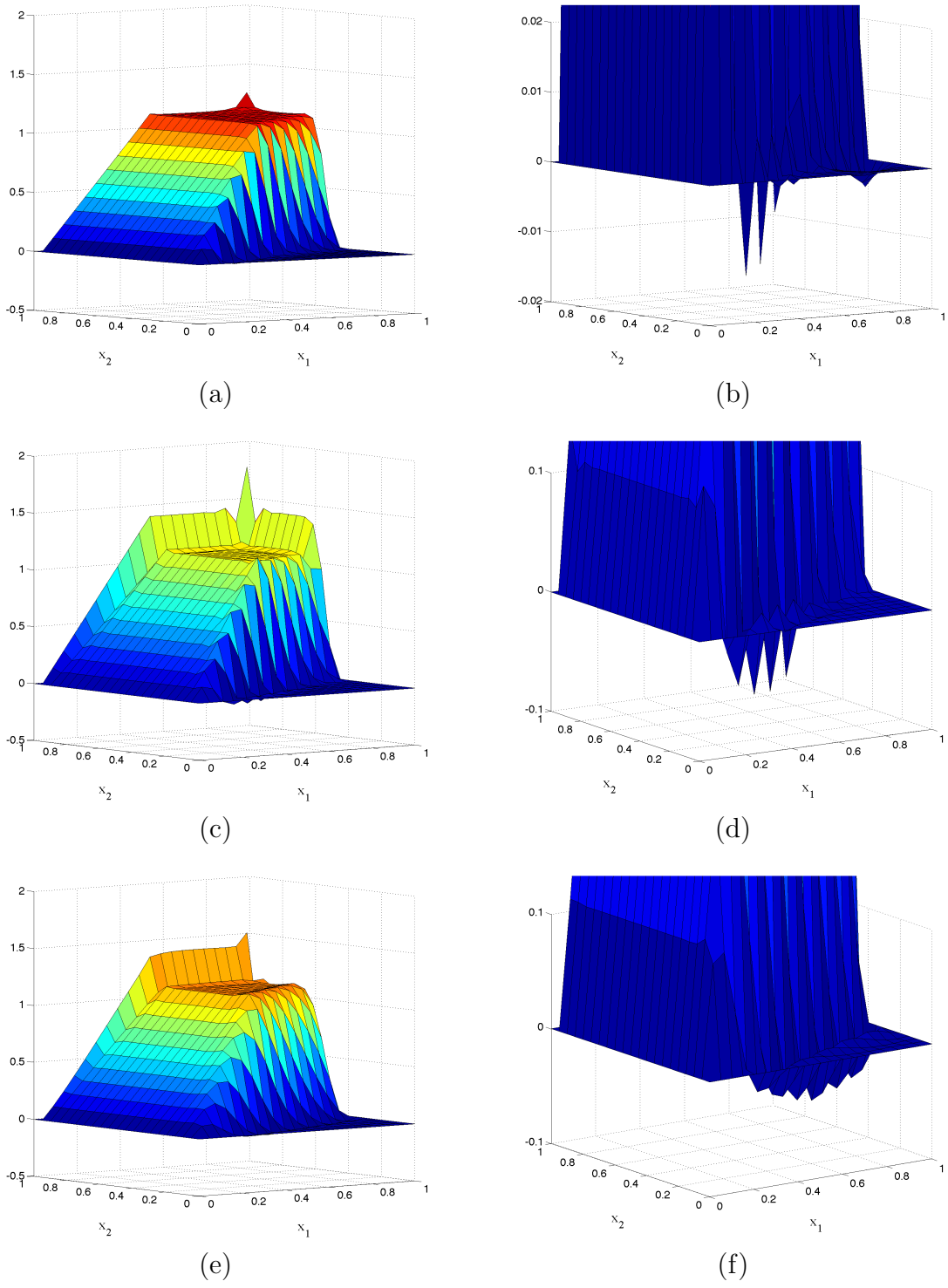



Figure 5: Numerical approximations for the example problem illustrated in Figure 3.
(a), (b) Classical VMS solution for $\mathcal{P} = \mathcal{P}_{H_0^1}$.
(c), (d) Classical VMS solution for $\mathcal{P} = \mathcal{P}_{L^2}$.
(e), (f) VMS-$\tau$ solution.

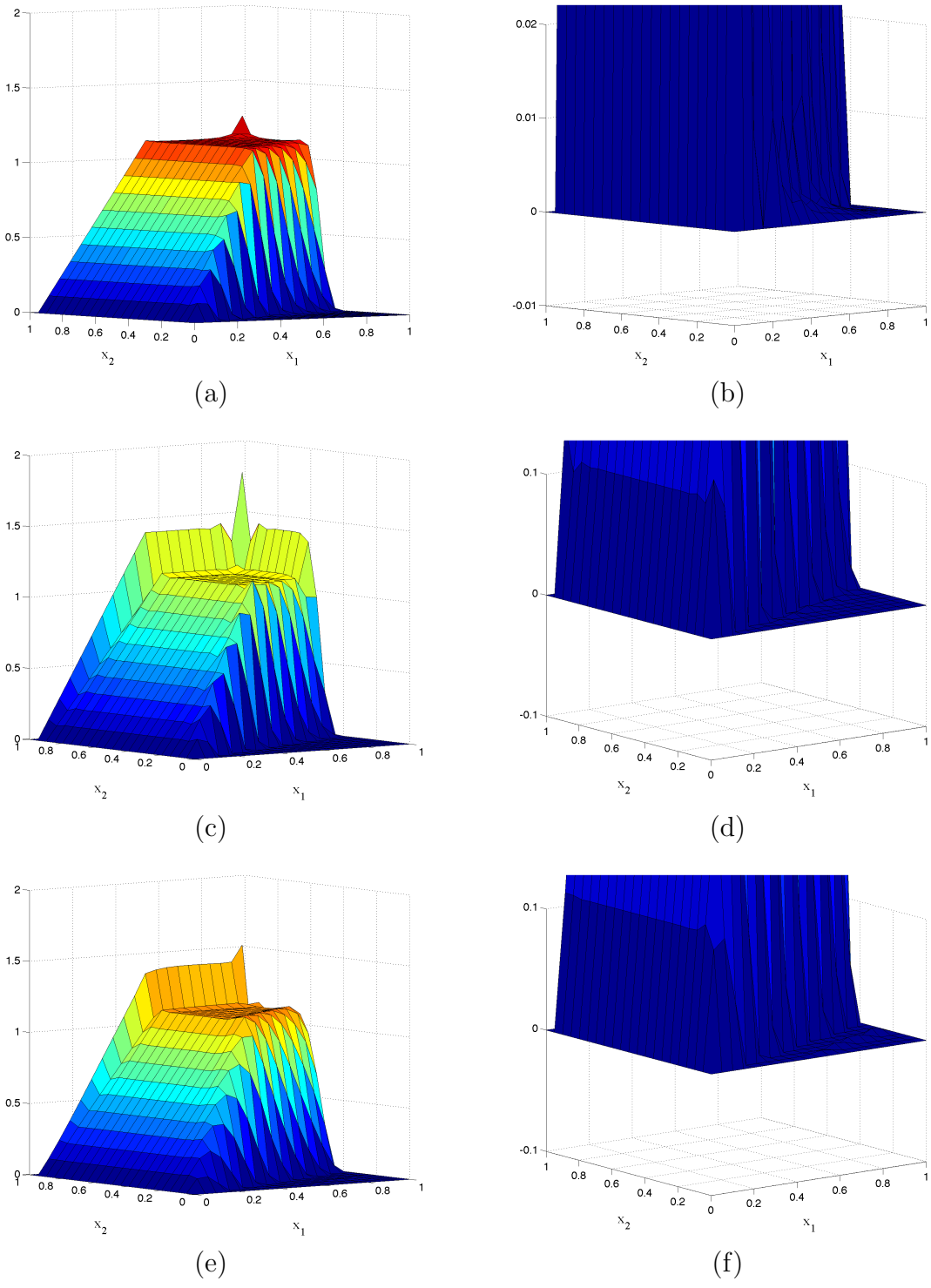



Figure 6: Numerical approximations for the example problem illustrated in Figure 3.
(a), (b) Constrained VMS solution for $H_0^1$-optimality.
(c), (d) Constrained VMS solution for $L^2$-optimality.
(e), (f) VMS-$\tau$ solution with direct enforcement of non-negativity constraint.

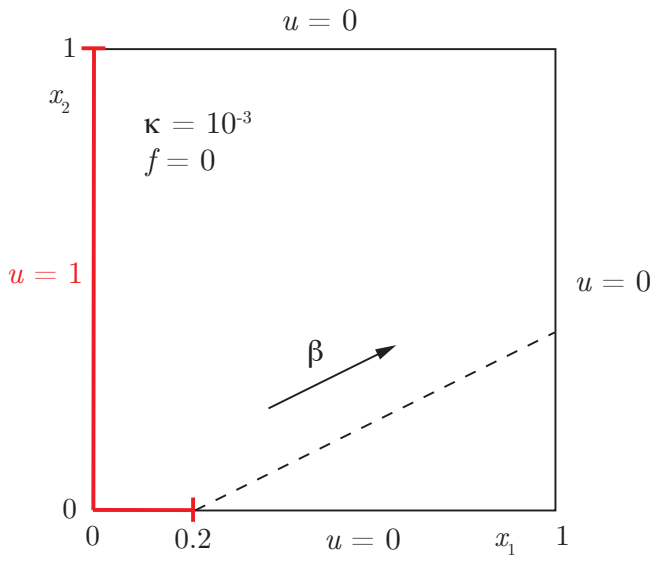



Figure 7: Problem description for the second steady advection-diffusion example problem.

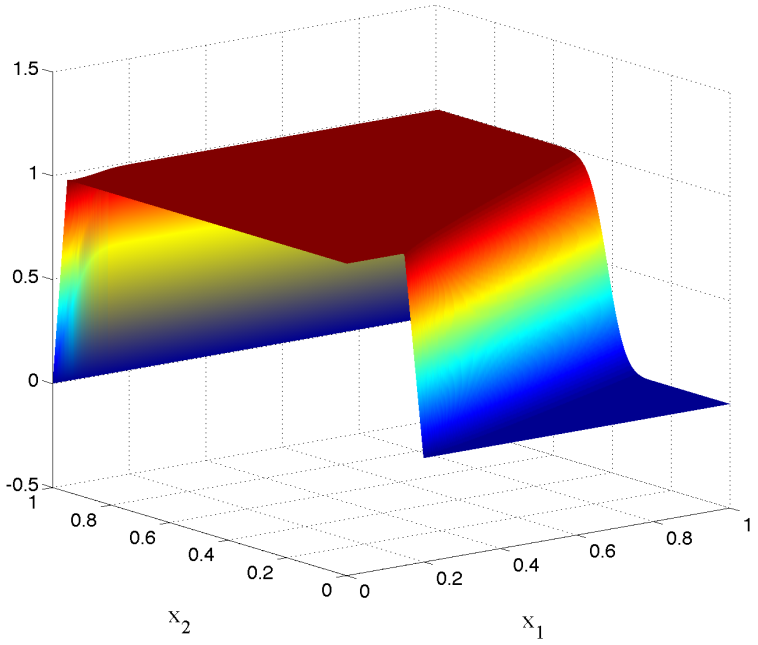
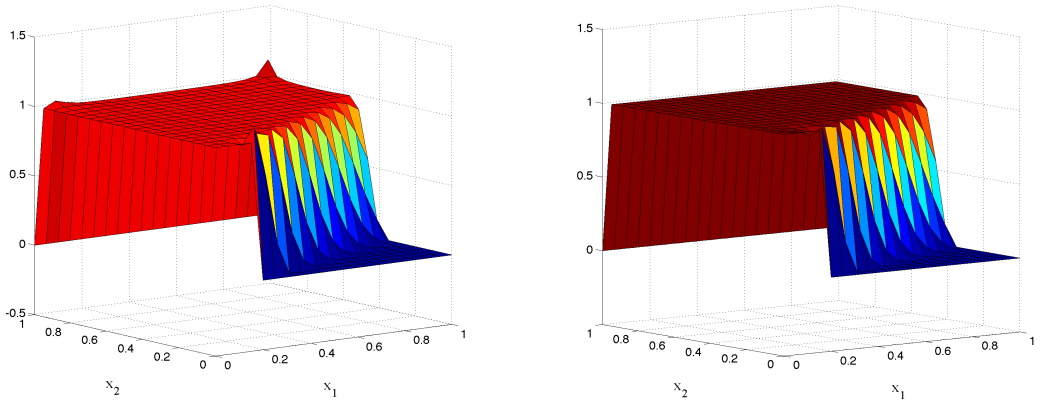


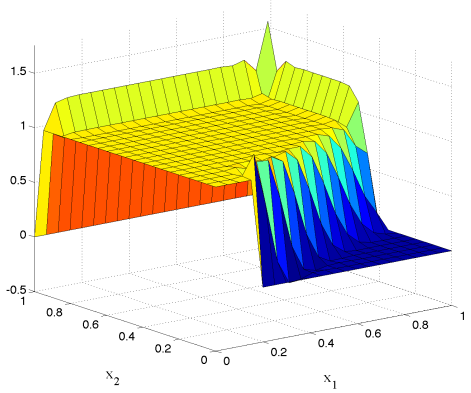

Figure 8: Exact solution of the example problem illustrated in Figure 7 using "ramp" boundary conditions.



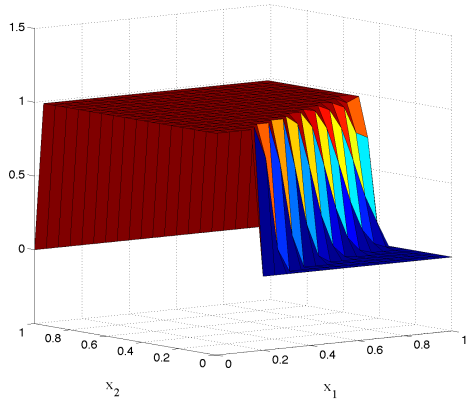

(a)



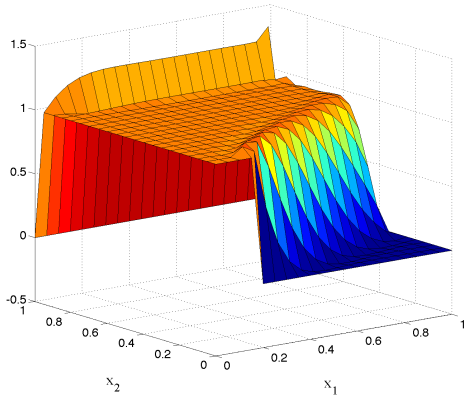

(b)



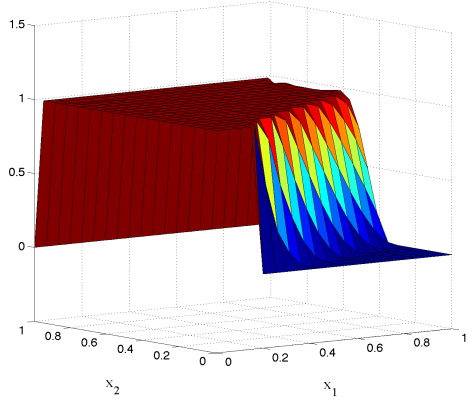

(c)




(d)

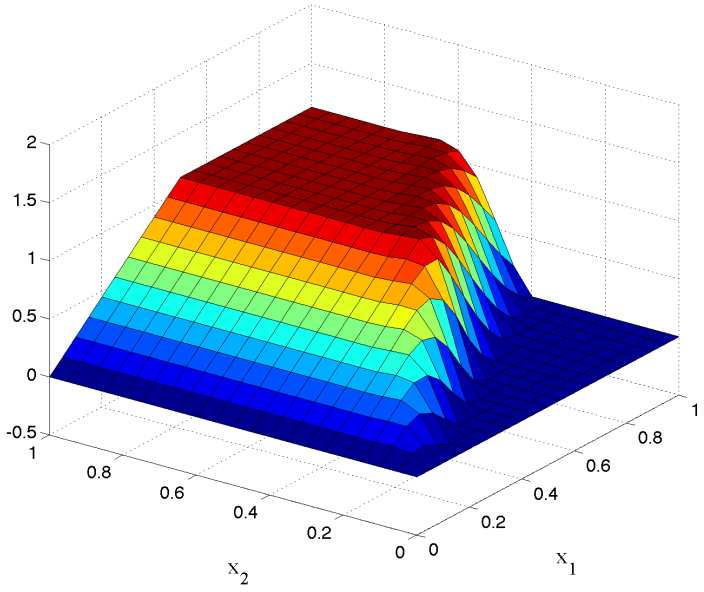



(e)

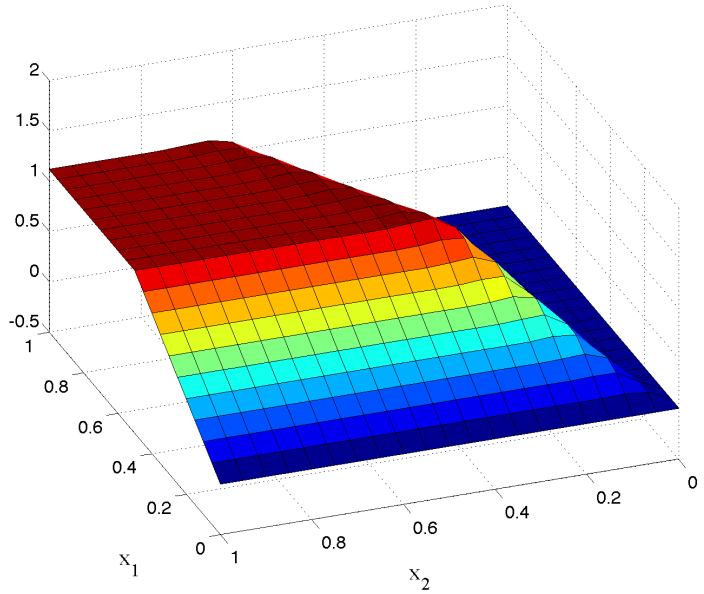



(f)

Figure 9: Numerical approximations for the example problem illustrated in Figure 7.
(a) Classical and (b) Constrained VMS solutions for $H_0^1$-optimality.
(c) Classical and (d) Constrained VMS solutions for $L^2$-optimality.
(e) VMS-$\tau$ solution.
(f) VMS-$\tau$ solution with direct enforcement of maximum principles.




(a)




(b)

Figure 10: VMS-$\tau$ solution for the example problem described in Figure 3 with weakly enforced boundary conditions and direct enforcement of non-negativity and conservation.
(a) Surface plot highlighting quality of solution.
(b) Surface plot illustrating weak enforcement of boundary conditions.

# References


[1] I. Akkerman, Y. Bazilevs, V.M. Calo, T.J.R. Hughes, and S. Hulshoff. The role of continuity in residual-based variational multiscale modeling of turbulence. *Computational Mechanics*, 41:371–378, 2008.

[2] K. Baba and M. Tabata. On a conservative upwind finite element scheme for convective diffusion equations. *RAIRO, Analyse Numérique*, 15:3–35, 1981.

[3] Y. Bazilevs, V.M. Calo, J.A. Cottrell, J.A. Evans, T.J.R. Hughes, S. Lipton, M.A. Scott, and T.W. Sederberg. Isogeometric analysis using T-splines. *Computer Methods in Applied Mechanics and Engineering*, 2009. In press.

[4] Y. Bazilevs, V.M. Calo, J.A. Cottrell, T.J.R. Hughes, A. Reali, and G. Scovazzi. Variational multiscale residual-based turbulence modeling for large eddy simulation of incompressible flows. *Computer Methods in Applied Mechanics and Engineering*, 197:173–201, 2007.

[5] Y. Bazilevs and T.J.R. Hughes. Weak imposition of Dirichlet boundary conditions in fluid mechanics. *Computers and Fluids*, 36:12–26, 2007.

[6] Y. Bazilevs, C. Michler, V.M. Calo, and T.J.R. Hughes. Weak Dirichlet boundary conditions for wall-bounded turbulent flows. *Computer Methods in Applied Mechanics and Engineering*, 196:4853–4862, 2007.

[7] Y. Bazilevs, C. Michler, V.M. Calo, and T.J.R. Hughes. Isogeometric variational multiscale modeling of wall-bounded turbulent flows with weakly-enforced boundary conditions on unstretched meshes. *Computer Methods in Applied Mechanics and Engineering*, 2009. doi:10.1016/j.cma.2008.11.020.

[8] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[9] F. Brezzi and A. Russo. Choosing bubbles for advection-diffusion problems. *Mathematical Models and Methods in the Applied Sciences*, 4:571–587, 1994.

[10] A.N. Brooks and T.J.R. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 32:199–259, 1982.

[11] E. Burman and A. Ern. Discrete maximum principles for Galerkin approximations of the Laplace operator on arbitrary meshes. *Comptes Rendus Mathematique*, 338:641–646, 2004.

[12] P.G. Ciarlet. Discrete maximum principle for finite difference operators. *Aequationes Mathematicae*, 4:338–352, 1970.

[13] P.G. Ciarlet and P.A. Raviart. Maximum principle and uniform convergence for the finite element method. *Computer Methods in Applied Mechanics and Engineering*, 2:17–31, 1973.

[14] R. Codina, J. Principe, O. Guasch, and S. Badia. Time dependent subscales in the stabilized finite element approximation of incompressible flow problems. *Computer Methods in Applied Mechanics and Engineering*, 196:2413–2430, 2007.

[15] J.A. Cottrell. *Isogeometric analysis and numerical modeling of the fine scales within the variational multiscale method*. PhD thesis, The University of Texas at Austin, 2007.

[16] D. Demarco, V.M. Calo, J.A. Cottrell, and T.J.R. Hughes. Monotone variational multiscale methods. In *9th US Congress on Computational Mechanics*, 2007.

[17] S. Deng, K. Ito, and Z. Li. Three-dimensional elliptic solvers for interface problems and applications. *Journal of Computational Physics*, 184:215–243, 2003.

[18] A. Draganescu, T.F. Dupont, and L.R. Scott. Failure of the discrete maximum principle for an elliptic finite element problem. *Mathematics of Computation*, 74:1–23, 2005.

[19] J. Droniou and R. Eymard. A mixed finite volume scheme for anisotropic diffusion problems on any grid. *Numerische Mathematik*, 105:35–71, 2006.

[20] J.A. Evans and T.J.R. Hughes. Monotonicity, total variation, and the variational multiscale method for evolution equations. In preparation.

[21] J.A. Evans, T.J.R. Hughes, and G. Sangalli. Discontinuity capturing and the variational multiscale method. In *8th World Congress on Computational Mechanics*, 2008.

[22] L.C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, Rhode Island, 1998.

[23] L.P. Franca, S.L. Frey, and T.J.R. Hughes. Stabilized finite element methods: I. Application to the advective-diffusive model. *Computer Methods in Applied Mechanics and Engineering*, 95:253–276, 1992.

[24] J.L. Guermond. A finite element technique for solving first-order PDEs in $L_p$. *SIAM Journal of Numerical Analysis*, 42:714–737, 2005.

[25] W. Höhn and H.D. Mittelmann. Some remarks on the discrete maximum principle for finite elements of higher-order. *Computing*, 27:145–154, 1981.

[26] J. Holmen, T.J.R. Hughes, A.A. Oberai, and G.N. Wells. Sensitivity of the scale partition for variational multiscale large-eddy simulation of channel flows. *Physics of Fluids*, 16:824–827, 2004.

[27] T.J.R. Hughes. Multiscale phenomena: Greens functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Computer Methods in Applied Mechanics and Engineering*, 127:387–401, 1995.

[28] T.J.R. Hughes, J.A. Cottrell, and Y. Bazilevs. Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. *Computer Methods in Applied Mechanics and Engineering*, 194:4135–4195, 2005.

[29] T.J.R. Hughes, G.R. Feijoo, L. Mazzei, and J.-B. Quincy. The variational multiscale method – a paradigm for computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, 166:3–24, 1998.

[30] T.J.R. Hughes, L.P. Franca, and G. Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The Galerkin least squares method for advective-diffusive equations. *Computer Methods in Applied Mechanics and Engineering*, 73:173–189, 1989.

[31] T.J.R. Hughes, L. Mazzei, and K.E. Jansen. Large eddy simulation and the variational multiscale method. *Computing and Visualization in Science*, 3:47–59, 2000.

[32] T.J.R. Hughes, L. Mazzei, A.A. Oberai, and A.A. Wray. The multiscale formulation of large eddy simulation: Decay of homogeneous isotropic turbulence. *Physics of Fluids*, 13:505–512, 2001.

[33] T.J.R. Hughes, A.A. Oberai, and L. Mazzei. Large eddy simulation of turbulent channel flows by the variational multiscale method. *Physics of Fluids*, 13:1784–1799, 2001.

[34] T.J.R. Hughes and G. Sangalli. Variational multiscale analysis: The fine-scale Green's function, projection, optimization, localization, and stabilized methods. *SIAM Journal on Numerical Analysis*, 45:539–557, 2007.

[35] T.J.R. Hughes, G.N. Wells, and A.A. Wray. Energy transfers and spectral eddy viscosity in large-eddy simulations of homogeneous isotropic turbulence: Comparison of dynamic Smagorinsky and multiscale models over a range of discretizations. *Physics of Fluids*, 16:4044–4052, 2004.

[36] T. Ikeda. *Maximum Principle in Finite Element Models for Convection-Diffusion Phenomena*. North-Holland, Amsterdam, NL, 1980.

[37] H. Kanayama. Discrete models for salinity distribution in a bay: Conservation laws and maximum principle. *Theoretical and Applied Mechanics*, 28:559–579, 1980.

[38] B. Koobus and C. Farhat. A variational multiscale method for the large eddy simulation of compressible turbulent flows on unstructured meshes – applications to vortex shedding. *Computer Methods in Applied Mechanics and Engineering*, 193:1367–1383, 2004.

[39] K. Lipnikov, M. Shashkov, D. Svyatskiy, and Y. Vassilevski. Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. *Journal of Computational Physics*, 227:492–512, 2007.

[40] R. Liska and M. Shashkov. Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems. *Communications in Computational Physics*, 3:852–877, 2008.

[41] A. Mizukami and T.J.R. Hughes. A Petrov-Galerkin finite element method for convection-dominated flows: An accurate upwinding technique for satisfying the maximum principle. *Computer Methods in Applied Mechanics and Engineering*, 50:181–193, 1985.

[42] M.J. Mlacnik and L.J. Durlofsky. Unstructured grid optimization for improved monotonicity of discrete solutions of elliptic equations with highly anisotropic coefficients. *Journal of Computational Physics*, 216:337–361, 2006.

[43] J.M. Nordbotten, I. Aavatsmark, and G.T. Eigestad. Monotonicity of control volume methods. *Numerische Mathematik*, 106, 2007.

[44] E.S. Oran and J.P. Boris. *Numerical Simulation of Reactive Flow*. Cambridge University Press, Cambridge, UK, 2001.

[45] C. Le Potier. Finite volume monotone scheme for highly anisotropic diffusion operators on unstructured triangular meshes. *Comptes Rendus Mathematique*, 341:787–792, 2005.

[46] S. Ramakrishnan and S.S. Collis. Turbulence control simulation using the variational multiscale method. *AIAA Journal*, 42:745–753, 2004.

[47] M. Tabata. A finite element approximation corresponding to the upwind finite differencing. *Memoirs of Numerical Mathematics*, 4:47–63, 1977.

[48] P. Šolína and T. Vejchodský. A weak discrete maximum principle for hp-FEM. *Journal of Computational and Applied Mathematics*, 209:54–65, 2007.

# Appendix A: Postprocessing

Theorem 5 of Section 2 suggests that one may postprocess a classical VMS solution to obtain the constrained VMS solution. In particular, given $\mathcal{P}u \in \bar{V}$, one solves the following formulation for $\bar{u} \in \bar{V}$.

Find $\bar{u} \in \bar{K}$, $\lambda_i \in \mathbb{R}$ $(i = 1, 2, \ldots, n_{eq})$, and $\mu_i \in \mathbb{R}$ $(i = 1, 2, \ldots, n_{ineq})$ such that

$$(\bar{u}, \bar{v})_H = (\mathcal{P}u, \bar{v})_H - \sum_{i=1}^{n_{eq}} \lambda_i \; {}_{\bar{V}^*}\langle Df_i, \bar{v}\rangle_{\bar{V}} - \sum_{i=1}^{n_{ineq}} \mu_i \; {}_{\bar{V}^*}\langle Dg_i(\bar{u}), \bar{v}\rangle_{\bar{V}}, \tag{85}$$

$$f_i(\bar{u}) = 0, \quad \text{for } i = 1, 2, \ldots, n_{eq}, \tag{86}$$

$$g_i(\bar{u}) \leq 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq}, \tag{87}$$

$$\mu_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq}, \tag{88}$$

$$\mu_i g_i(\bar{u}) = 0, \quad \text{for } i = 1, 2, \ldots, n_{ineq}. \tag{89}$$

The above equations may be solved using a variety of algorithms from constrained optimization. We describe here how to use the active set method to arrive at a solution. Suppose the space $\bar{V}$ has the representation

$$\bar{V} = \left\{ \bar{v} \in \bar{V} : \bar{v} = \sum_{i=1}^{n_{dof}} \bar{v}_i N_i, \;\; \bar{v}_i \in \mathbb{R} \right\}. \tag{90}$$

Define the matrix $\mathbf{M} = [M_{ij}] \in \mathbb{R}^{n_{dof} \times n_{dof}}$ such that

$$M_{ij} = (N_i, N_j)_H, \tag{91}$$

the matrices $\mathbf{F} = [F_{ij}] \in \mathbb{R}^{n_{eq} \times n_{dof}}$ and $\mathbf{G}(\bar{\mathbf{u}}) = [G_{ij}(\bar{u})] \in \mathbb{R}^{n_{ineq} \times n_{dof}}$ such that

$$F_{ij} = {}_{\bar{V}^*}\langle Df_i, N_j\rangle_{\bar{V}}, \tag{92}$$

$$G_{ij}(\bar{u}) = {}_{\bar{V}^*}\langle Dg_i(\bar{u}), N_j\rangle_{\bar{V}}, \tag{93}$$

the vectors $\bar{\mathbf{u}} = \{\bar{u}_i\}^T$, $\mathbf{u} = \{u_i\}^T$ where

$$\bar{u} = \sum_{i=1}^{n_{dof}} \bar{u}_i N_i, \tag{94}$$

$$\mathcal{P}u = \sum_{i=1}^{n_{dof}} u_i N_i, \tag{95}$$

and the vectors $\boldsymbol{\lambda} = \{\lambda_i\}^T$, $\boldsymbol{\mu} = \{\mu_i\}^T$, and $\mathbf{g}(\bar{\mathbf{u}}) = \{g_i(\bar{u})\}^T$. If finite element basis functions are employed to represent the coarse-scales, then the above matrices and vectors can be constructed using a standard assembly procedure. Note that by construction, the matrix $\mathbf{M}$ is symmetric and positive-definite[1].

[1] A caution is in order. $\mathbf{M}$ is defined with respect to the scalar product on $H$, that is, $(\cdot, \cdot)_H$. This means that $\mathbf{M}$ will *only* be the classical mass matrix when $H = L^2$.

Our problem has the following algebraic form.

Find $\bar{\mathbf{u}}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\mu}$ such that

$$\mathbf{M}\bar{\mathbf{u}} = \mathbf{M}\mathbf{u} - \boldsymbol{\lambda}^T\mathbf{F} - \boldsymbol{\mu}^T\mathbf{G}, \tag{96}$$

$$\mathbf{F}\bar{\mathbf{u}} = 0, \tag{97}$$

$$\mathbf{g}(\bar{\mathbf{u}}) \leq 0, \tag{98}$$

$$\boldsymbol{\mu} \geq 0, \tag{99}$$

$$\boldsymbol{\mu}^T\mathbf{g}(\bar{\mathbf{u}}) = 0. \tag{100}$$

To solve the above problem, one can partition inequality constraints into two groups: those that are to be treated as active and those that are to be treated as inactive. The inequality constraints treated as inactive are effectively ignored. The idea of active set methods is to define at each step of an algorithm a set of constraints, termed the *working set*, that is to be treated as the active set. At each step, the working set constraints are treated as equality constraints and the other inequality constraints are ignored. A standard algebraic solver can then be utilized to solve the resulting system. At the end of this solve, the working set is modified to ensure that all of the inequality constraints are satisfied and to ensure non-negativity of the inequality Lagrange multipliers. This procedure is illustrated below.

```
1  Set the working set W to be the empty set;
2  Set Converged = 0;
3  while Converged = 0 do
4      Solve the algebraic problem

           Mū   = Mu − λ^T F − μG
           Fū   = 0
           g_i(ū) = 0   if i ∈ W
           μ_i  = 0   if i ∉ W

       for ū, λ, μ;
5      Set Converged = 1;
6      foreach i = 1, 2, ..., n_ineq do
7          if i ∉ W and g_i(ū) ≥ 0 then
8              Set Converged = 0;
9              Insert i into W;
10         else if i ∈ W and μ_i < 0 then
11             Set Converged = 0;
12             Remove i from W;
13         end
14     end
15 end
```

**Algorithm 1**: Postprocessing active set algorithm

The algebraic solve appearing in the postprocessing algorithm can be solved using either a direct or iterative solver in the case of linear inequality constraints or a Newton-Raphson or other nonlinear procedure in the case of nonlinear inequality constraints. We do not detail this here. The active set algorithm is guaranteed to converge, but theoretical estimates only guarantee that this occurs with an exponential number of steps. Our experience, however, has shown that this algorithm converges in a very small number of steps for practical applications (between 3 and 4 steps for the examples presented herein).

# Appendix B: Computing the fine-scale component $\hat{u}'$ and its variational derivatives

If one wishes to directly solve the full form of the constrained variational multiscale formulation, i.e., formulation (‡) from Section 2, one must compute the fine-scale quantity $\hat{u}'$ and various variational derivatives.

Recall that we can write

$$\hat{u}' = \sum_{i=1}^{n_{eq}} \lambda_i b_i + \sum_{i=1}^{n_{ineq}} \mu_i c_i(\bar{u}). \tag{101}$$

where $b_i$ and $c_i$ are defined by (20) and (21) respectively. Assume the coarse-scale space $\bar{V}$ has the representation of (90). Suppose further that $b_i$ and $c_i$ have the representations

$$b_i = \sum_{j=1}^{n_{dof}} b_{ij} N_j, \tag{102}$$

$$c_i(\bar{v}) = \sum_{j=1}^{n_{dof}} c_{ij}(\bar{v}) N_j. \tag{103}$$

Defining the vectors $\mathbf{b}_i = \{b_{ij}\}^T$ and $\mathbf{c}_i(\bar{\mathbf{v}}) = \{c_{ij}(\bar{v})\}^T$ and using notation introduced in Appendix A, we have that

$$\mathbf{b}_i = \mathbf{M}^{-1}\mathbf{F}_i^T, \tag{104}$$

$$\mathbf{c}_i(\bar{\mathbf{u}}) = \mathbf{M}^{-1}\mathbf{G}_i(\bar{\mathbf{u}})^T, \tag{105}$$

where $\mathbf{F}_i$ and $\mathbf{G}_i(\bar{\mathbf{u}})$ are the $i^{\text{th}}$ rows of matrices $\mathbf{F}$ and $\mathbf{G}(\bar{\mathbf{u}})$ respectively. Since $\mathbf{M}$ is a symmetric, positive-definite matrix, the above expressions can be evaluated using a conjugate gradient linear solver. Hence, the quantity $\hat{u}'$ and its variational derivatives with respect to the Lagrange multipliers can be evaluated fairly simply.

In the case of nonlinear inequality constraints, the variational multiscale formulation consists of a set of nonlinear algebraic relations. If a Newton-Raphson solver is to be employed to solve this nonlinear system, one must then also compute the variational derivative of $\hat{u}'$ with respect to the coarse-scale solution. For example, for a finite element implementation, one computes the variational derivative of $\hat{u}'$ with respect to the nodal variables. One writes

$$\begin{aligned} \frac{\partial \hat{u}'}{\partial \bar{u}_k} &= \sum_{i=1}^{n_{ineq}} \mu_i \frac{\partial c_i(\bar{u})}{\partial \bar{u}_k} \\ &= \sum_{i=1}^{n_{ineq}} \sum_{j=1}^{n_{dof}} \mu_i \frac{\partial c_{ij}(\bar{u})}{\partial \bar{u}_k} N_j. \end{aligned} \tag{106}$$

The variational derivatives of the various components $c_{ij}$ with respect to the coarse-scale variables $\bar{u}_k$ satisfy

$$\frac{\partial \mathbf{c}_i(\bar{\mathbf{u}})}{\partial \bar{u}_k} = \mathbf{M}^{-1} \frac{\partial \mathbf{G}_i(\bar{\mathbf{u}})^T}{\partial \bar{u}_k} \tag{107}$$

for each $k = 1, 2, \ldots, n_{dof}$. Thus, for each constraint, one must form $n_{dof}$ right-hand-sides and solve $n_{dof}$ linear systems. This is wholly impracticable. Consequently, it is advised that a nonlinear solver which does not require the exact variational derivatives be used. For example, these derivatives can be approximated using a secant technique. This results in an approximate tangent matrix and a Quasi-Newton method.