arXiv:2304.01067v1 [math.NA] 3 Apr 2023

一种节点界保持有限元方法

GABRIEL R. BARRENECHEA、EMMANUIL GEORGOULIS、TRISTAN PRYER 和 ANDREAS VEESER

Abstract. 本文提出一种非线性有限元方法，其节点值能够保持精确解所满足的已知界。离散问题涉及一个非线性投影算子，该算子将任意节点值映射为保界节点值，并在此投影的值域中求数值解。由于该投影并非单射，因此加入基于互补投影的稳定化项，以恢复适定性。在椭圆问题框架下，离散问题可视为离散障碍问题的一种重构，并通过 Lipschitz 投影纳入不等式约束。

本文以线性和非线性反应扩散问题为例，说明所提出方法的推导过程。我们在适当的范数下建立了近最佳逼近结果。特别地，我们证明在线性情形下，数值解是在所有节点界保持有限元函数中能量范数意义下的最佳逼近。

针对这类问题的一系列数值实验展示了所提出的保界有限元方法的良好表现。

1. Introduction

多年来，保结构数值方法一直是计算偏微分方程（PDE）领域的重要主题。所谓保结构，是指所采用的方法能够生成满足底层精确问题某些期望性质的近似解，例如局部守恒、熵不等式、最大值原理、逐点无散约束或严格对称的应力张量近似，等等。

满足 离散最大值原理（*DMP*）和/或单调性性质的数值方法已在有限元文献中得到广泛研究；参见[11, 12, 22, 27, 8, 7, 1, 2] 以查阅一份（非常）不完整的列表，以及 [3] 以了解近期综述和更多相关参考文献。满足后者性质的方法意味着所得数值解具有正性保持或更一般地说，界保持性质。

然而，至关重要的是，保界似乎是比离散最大值原理（DMP）要求更弱的结构性条件。尽管如此，对于由偏微分方程组描述的许多复杂非线性现象而言，保界数值解对数值稳定性不可或缺。例如，许多此类偏微分方程模型仅在其组成方程的解为正时才有效。这方面的例子包括描述反应物浓度或诱导湍流场的非线性反应扩散系统。此外，相场偏微分方程模型通常具有这样的特征：其解满足逐点的全局最大值和最小值。尽管数值解未能保持相同的界限

对于标量偏微分方程问题而言通常不会造成灾难性后果，但当此类偏微分方程属于更复杂的方程组时，这种影响可能会不断累积。

因此，人们对保界方法产生了兴趣。经过恰当设计，这类方法可能不像当前那些力求满足离散最大值原理或单调性的格式那样“僵硬”。

本文沿着这一思路提出一种新的有限元方法。为此，假设给定一个偏微分方程（及其相应的弱形式）和一个任意阶的拉格朗日有限元空间。此外，为简单起见，假设精确解为正，即零函数是其下界。那么，新方法的形式推导可概括为以下三个步骤：

- 在有限元空间上引入一个投影算子 $v_h \mapsto v_h^+$ ，使投影函数 v_h^+ 的节点值为正。
- 照常将连续问题弱形式中的测试空间替换为有限元空间；但不再寻求以有限元函数作为数值解，而是寻求一个投影函数，其形式为 u_h^+ 。

- 基于互补投影 $v_h \mapsto v_h^-$ 到有限元空间上的稳定化。

请注意，投影算子及所提出的方法都是非线性的。鉴于经典的“戈杜诺夫障碍”原理，后者是意料之中的。所提出的方法的解是在投影的值域中求得的，其中投影充当参数化，而参数区域是有限元空间。投影不是单射这一事实使得第三步中必须进行稳定化。

有趣的是，如果其节点值为正，则经典有限元解与所提出方法的解一致。更准确地说，也更为重要的是，当前方法的解最终证明是精确解到节点值为正的有限元函数所构成的闭凸集上的正交投影。这一观察在本文中起着关键作用，并带来以下两个重要结论：

- 所提出的方法可以看作是对离散障碍问题的一种重新表述。与后者不同，这种重新表述由一个离散变分等式构成，并将关键的不等式约束编码在利普希茨连续投影中。这一差异或许能为通常使用的约束优化技术之外的求解器铺平道路；例如，参见 [16]。在此，我们采用一种基于类理查森迭代的简单方法。

- 适用于若干误差概念的近最佳逼近可作为推论得出。特别地，对于线性反应扩散问题，所提出的方法的解实际上是前述凸子集中精确解在能量范数下的最佳逼近。对于更一般的、具有幂型非线性反应的半线性问题，证明了所提出的方法相对于H^1-半范数与反应项相关拟范数之和是准最优的。

为说明这两个方面，本文给出了一系列数值实验，展示了所提出的方法相较于标准有限元近似的优越性能。

值得一提的是，文献中已有相关方法（尽管与本文方法存在明显差异）。在 [5] 中，连续解的界被作为约束优化问题中的限制条件，并阐述了其与非线性稳定化之间的联系。此外，在 [28] 中，采用涉及混合弱形式的约束优化问题来实现界保持。在 [9] 中，利用正性保持与接触问题之间的联系，提出了一种非线性稳定化方法，以弱形式保证解的正性。此外，截断有限元方法 [26]会在有限元函数计算完成后，于给定时间步对其进行截断，并将截断后的函数作为当前时间步的近似解；另见 [32] ，其中将相关思想应用于 Allen-Cahn 方程。对于稳态情形，文献采用能量论证，证明了将有限元解截断以满足给定界限的方法适用于反应-扩散方程。 [23]在 [24] 中提出了一种守恒恢复策略，并进行了数值测试。最后，在焦耳热问题中，文献 [21]引入对其中一个变量的截断，以正则化粗糙的右端项。

本文其余部分安排如下。引言的剩余部分用于介绍本文的记号和主要假设。第3节给出有限元方法，并证明其稳定性和最优误差估计。随后，在第 4节中，我们将这一框架推广到非线性反应-扩散方程，最后，第 5 节用于介绍数值实验。

2. General setting and the linear model problem

我们将使用Sobole夫空间的标准记号，与例如 [15] 中的记号一致。更确切地说，对于 $D \subseteq \mathbb{R}^d$, $d = 1, 2, 3$，我们将$\|\cdot\|_{0,p,D}$ 记作 $L^p(D)$-范数；当 $p = 2$时，下标 p 通常省略，并且只写作$\|\cdot\|_{0,D}$。此外，对于$s \geq 0, p \in (1, \infty)$，我们将$\|\cdot\|_{s,p,D}$ ($|\cdot|_{s,p,D}$)记作 $W^{s,p}(D)$ 中的范数（半范数）；当 $p = 2$时，我们同样通常省略下标 p。另外，我们用 $H^{-1}(D)$ 表示 $H_0^1(D)$ 的对偶空间，同时将 $L^2(D)$ 与其对偶空间认同。因此，记$\langle\cdot,\cdot\rangle_D$ 为对偶配对，有

$$\langle f, v\rangle_\Omega = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \qquad \forall\, v \in H_0^1(\Omega)\,,$$

只要 $f \in H^{-1}(D)$ 足够正则即可。 $L^2(D)$-内积记为 $(\cdot,\cdot)_D$。对于标量函数和向量值函数，我们不区分内积与对偶配对。

我们所关注的边值问题定义在有界开区域 Ω 上，其边界 $\partial\Omega$ 为多面体利普希茨边界。令 $\mathcal{P}$ 为 Ω 的一个相容的形状正则剖分，剖分单元为单纯形（或四边形/六面体） K。在$\mathcal{P}$上，对于 $k \geq 1$，我们定义有限元空间

$$V_{\mathcal{P}} := \{v_h \in C^0(\Omega) : v_h|_K \in \mathcal{R}(K)\ \forall\, K \in \mathcal{P}\} \cap H_0^1(\Omega)\,, \tag{1}$$

其中 $\mathcal{R}(K) = \begin{cases} \mathbb{P}_k(K), & \text{若 } K \text{是单纯形 } \mathbb{Q}_k \ K \qquad K \qquad (), \\ \text{若是仿射四边形/六面体,} \end{cases}$ (2)

其中 $\mathbb{P}_k(K)$ 表示 K 上的总次数为 k 的多项式，而 $\mathbb{Q}_k(K)$ 表示经仿射映射从参考单元映射而来的、每个变量次数为 $\leq k$ 的多项式。我们用 $\boldsymbol{x}_i$, $i = 1, \ldots, N$ 表示 $\mathcal{P}$的内部节点集合，用 $\phi_1, \ldots, \phi_N$ 表示张成空间 $V_{\mathcal{P}}$的常规拉格朗日基函数集合。

K 的直径记为 h_K, $h = \max\{h_K : K \in \mathcal{P}\}$，并将网格函数 $\mathfrak{h}$ 定义为一个连续的、逐单元线性函数，其值由局部平均确定。更具体地，对于 $\mathcal{P}$ 的节点 $\boldsymbol{x}_i$，我们定义局部邻域$K_{\boldsymbol{x}_i} = \{K \in \mathcal{P} : \boldsymbol{x}_i \in K\}$，并将 $\mathfrak{h}$ 定义为 $V_{\mathcal{P}}$ 中唯一由节点值给定的元素（其中 $k = 1$）

$$\mathfrak{h}(\boldsymbol{x}_i) = \frac{\sum_{K \in K_{\boldsymbol{x}_i}} h_K}{\# K_{\boldsymbol{x}_i}} . \tag{3}$$

我们回顾逆不等式（例如参见 [15, 引理 12。1]）：对所有 m, $\ell \in \mathbb{N}$, $m \leq \ell$ 以及所有 p, $q \in [1, +\infty]$, 存在常数 C，其与 $K \in \mathcal{P}$ 无关，使得

$$|q|_{\ell,p,K} \leq C h_K^{m-\ell+d\left(\frac{1}{p}-\frac{1}{q}\right)} |q|_{m,q,K} , \tag{4}$$

对于定义在 K 上的每个多项式函数 q。

在空间 $V_{\mathcal{P}}$ 中，我们用 $(\cdot, \cdot)_h$ 表示由下式给出的集总 $L^2(\Omega)$-内积

$$(v_h, w_h)_h := \sum_{i=1}^{N} \mathfrak{h}^d(\boldsymbol{x}_i) v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i) , \tag{5}$$

[1]其对应范数为 $|v_h|_h = (v_h, v_h)_h^{2-}$。

注 2.1。以下结果的证明可见于 [15, 命题 12。5]，该结果对于我们分析第3节所提出的方法非常重要：存在 C, $c > 0$，它们与 $\mathfrak{h}$无关，使得

$$c\,|v_h|_h^2 \leq \|v_h\|_{0,\Omega}^2 \leq C\,|v_h|_h^2 , \tag{6}$$

对于所有 $v_h \in V_{\mathcal{P}}$。

2.1. 线性模型问题。给定 $f \in H^{-1}(\Omega)$，我们考虑以下反应-扩散方程：求 $u : \Omega \to \mathbb{R}$，使得

$$\begin{aligned} -\operatorname{div}\left(\mathcal{D}\nabla u\right) + \mu u &= f \qquad \text{in } \Omega , \\ u &= 0 \qquad \text{on } \partial\Omega , \end{aligned} \tag{7}$$

其中 $\mathcal{D} = \left(d_{ij}\right)_{i,\ j=1}^{d} \in L^\infty(\Omega)^{d\times d}$ 和 $\mu \in L^\infty(\Omega)$ 分别表示扩散张量和反应系数。我们假设 $\mu(\boldsymbol{x}) \geq \mu_0 \geq 0$ 在 Ω 中几乎处处成立，并且扩散张量 $\mathcal{D}$ 在

Ω 中是对称且一致严格正定的，即存在正常数 $d_0 > 0$，使得对于几乎所有 $\boldsymbol{x} \in \Omega$，都有

$$\sum_{i,j=1}^{d} y_i d_{ij}(\boldsymbol{x}) y_j \geq d_0 \sum_{i=1}^{d} y_i^2 \qquad \forall\, (y_1, \ldots, y_d) \in \mathbb{R}^d\,. \tag{8}$$

(7) 的弱形式如下：求 $u \in H_0^1(\Omega)$，使得

$$a(u, v) = \langle f, v\rangle_\Omega \qquad \forall\, v \in H_0^1(\Omega)\,, \tag{9}$$

其中，$a(\cdot,\cdot): H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ 是由下式定义的双线性型

$$a(v, w) := (\mathcal{D}\nabla v, \nabla w)_\Omega + (\mu v, w)_\Omega\,. \tag{10}$$

定义能量范数

$$\|v\|_a := \sqrt{a(v, v)}\,, \tag{11}$$

(9) 的适定性由 Lax-Milgram 引理（参见例如 [15] ）.

注 2.2. 本工作的核心是方程 (9)解的以下性质：由最大值原理和比较原理（参见 [17], 或 [29,推论 4.4]）可证明以下界：对几乎所有 $\boldsymbol{x} \in\Omega$，解u (9)满足

$$-\frac{\|f\|_{0,\infty,\Omega}}{\mu_0} \leq u(\boldsymbol{x}) \leq \frac{\|f\|_{0,\infty,\Omega}}{\mu_0}\,. \tag{12}$$

如果 $f \geq 0$ 属于Ω，则可以将最后一个陈述表述得更加精确。事实上，在这种情况下，对于几乎所有的 $\boldsymbol{x} \in\Omega$，以下不等式成立：

$$0 \leq u(\boldsymbol{x}) \leq \frac{\|f\|_{0,\infty,\Omega}}{\mu_0}\,. \tag{13}$$

上述备注中的结果促使我们引入以下假设。

假设（A1）：我们假设方程（7）的解满足

$$0 \leq u(x) \leq \mathfrak{a} \qquad \text{for almost all } x \in \Omega\,, \tag{14}$$

其中，$\mathfrak{a}$ 是一个已知常数。

备注 2.3。假设（A1）中的下界不必等于零，以下证明的所有结果对于更一般的有界框同样成立，无需进行重大修改。特别地，$\mathfrak{a}$ 可以替换为一个非负连续函数 $\mathfrak{a}(\boldsymbol{x})$。

方程（7）的标准Galerkin有限元方法如下：求使得

$$a(u_h^{\mathrm{FEM}}, v_h) = \langle f, v_h\rangle_\Omega \qquad \forall\, v_h \in V_{\mathcal{P}}\,. \tag{15}$$

众所周知，u_h^{FEM}在以下意义下是最佳逼近：

$$\|u - u_h^{\mathrm{FEM}}\|_a = \inf_{v_h \in V_{\mathcal{P}}} \|u - v_h\|_a. \tag{16}$$

然而，这一良好性质并不能避免在反应项 μ 占主导、扩散项 $\mathcal{D}$较弱时出现以下情况：例如，若 $\|\mathcal{D}\|_{0,\infty,\Omega} \ll \mu$，则 u_h^{FEM}可能会出现伪振荡

（例如，参见 [30]），因此 u_h^{FEM}可能无法满足假设（A1）。事实上，对于（9）所示的反应-扩散方程，必须对网格施加非常严格的条件，才能使其解满足假设（A1）（详情参见 [7]），即使在标量（各向同性）扩散的情况下也是如此。对于一般扩散 $\mathcal{D}$，除了网格尺寸限制外，还需要施加与 $\mathbb{R}^d$ 中的加权内积（$\mathcal{D}\cdot,\cdot$）相关的网格条件（详情参见 [20] ）。

3. The finite element method for the linear problem

本节旨在推导一种方法：一方面，在不对网格施加限制的情况下，基本保持假设(A1)中的界；另一方面，保留经典有限元解良好的近似性质（16）。

为此，我们定义 $V_{\mathcal{P}}$ 的以下闭凸子集：

$$V_{\mathcal{P}}^{+} := \{v_h \in V_{\mathcal{P}} : v_h(\boldsymbol{x}_i) \in [0, \mathfrak{a}] \text{ for all } \boldsymbol{x}_i = 1, \dots, N\}\,, \tag{17}$$

也就是说，该集合由在其自由度处满足界（14）的有限元函数组成。基于这一凸集，每个有限元函数都可分解为$v_h = v_h^+ + v_h^-$ ，其中 $v_h^+ \in V_{\mathcal{P}}^+$ 定义为

$$v_h^+ := \sum_{i=1}^{N} \max\left\{0, \min\{v_h(\boldsymbol{x}_i), \mathfrak{a}\}\right\}\phi_i\,, \tag{18}$$

and

$$v_h^- := v_h - v_h^+\,, \tag{19}$$

是函数在 $V_{\mathcal{P}}^+$ 之外的部分。从现在起，我们将函数 v_h^+ 和 v_h^- 分别称为 v_h 的受约束部分和互补部分。

本文提出的有限元方法如下：求 $u_h \in V_{\mathcal{P}}$，使得 t

$$a_h(u_h; v_h) = \langle f, v_h\rangle_\Omega \qquad \forall\, v_h \in V_{\mathcal{P}}\,, \tag{20}$$

其中，$a_h(\cdot;\cdot)$ 是由下式给出的非线性型

$$a_h(u_h; v_h) := a(u_h^+, v_h) + s(u_h^-, v_h)\,, \tag{21}$$

其中 $a(\cdot,\ \cdot)$ 定义见 (10)，且 $s(\cdot,\ \cdot)$: $C(\Omega) \times C(\Omega) \to \mathbb{R}$ 是由下式定义的稳定化双线性型

$$s(v_h, w_h) = \alpha \sum_{i=1}^{N} \left(\|\mathcal{D}\|_{0,\infty,\omega_i}\mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \|\mu\|_{0,\infty,\omega_i}\mathfrak{h}(\boldsymbol{x}_i)^{d}\right) v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i)\,, \tag{22}$$

其中 $\alpha > 0$是一个无量纲常数，其精确取值将在定理 3.2中确定，且 $\omega_i :$ $= \cup_{K\in\mathcal{P}: K\cap K_{\boldsymbol{x}_i}\neq\emptyset} K$ 表示一个扩展补丁。其定义将在下文建立 (24) 时用到。

将稳定化范数定义为

$$\|v_h\|_s := \sqrt{s(v_h, v_h)}\,, \tag{23}$$

并使用（6），存在一个 $C_{\text{equiv}} > 0$，其仅取决于形状正则性常数，使得

$$\|v_h\|_a^2 \le \frac{C_{\text{equiv}}}{\alpha} \|v_h\|_s^2 \qquad \forall\, v_h \in V_{\mathcal{P}}\,, \tag{24}$$

其中 $\alpha > 0$ 是定义（22） $s(\cdot,\ \cdot)$ 中出现的稳定化参数。

3.1. **适定性与相容性**。

本节将分析所提出的方法（20）的存在性、唯一性和稳定性结果。我们先给出稳定化形式 $s(\cdot,\ \cdot)$ 的以下单调性结果。

引理 3.1。双线性型 $s(\cdot,\cdot)$ 满足以下不等式：

$$s(v_h^- - w_h^-, v_h^+ - w_h^+) \ge 0 \qquad \forall\, v_h, w_h \in V_{\mathcal{P}}\,, \tag{25}$$

$$s(v_h^-, w_h - v_h^+) \le 0 \qquad \forall\, v_h \in V_{\mathcal{P}}, w_h \in V_{\mathcal{P}}^+\,. \tag{26}$$

证明。设 $\boldsymbol{x}_i \in \mathcal{P}$ 是任意一个内部节点。若 $v_h(\boldsymbol{x}_i) \ge w_h(\boldsymbol{x}_i)$，则 $v_h^+(\boldsymbol{x}_i) \ge w_h^+(\boldsymbol{x}_i)$。此外，若 $v_h(\boldsymbol{x}_i) \ge w_h(\boldsymbol{x}_i) > \mathfrak{a}$，则 $v_h^-(\boldsymbol{x}_i) = v_h(\boldsymbol{x}_i) - \mathfrak{a} \ge w_h(\boldsymbol{x}_i) - \mathfrak{a} = w_h^-(\boldsymbol{x}_i)$。若 $v_h(\boldsymbol{x}_i) > \mathfrak{a}$ 且 $w_h(\boldsymbol{x}_i) \le \mathfrak{a}$，则 $v_h^-(\boldsymbol{x}_i) = v_h(\boldsymbol{x}_i) - \mathfrak{a} \ge 0 \ge w_h^-(\boldsymbol{x}_i)$。若 $v_h(\boldsymbol{x}_i) \le \mathfrak{a}$，则或者 $v_h^-(\boldsymbol{x}_i) = 0 \ge w_h^-(\boldsymbol{x}_i)$，或者 $v_h^-(\boldsymbol{x}_i) = v_h(\boldsymbol{x}_i) \ge w_h(\boldsymbol{x}_i) = w_h^-(\boldsymbol{x}_i)$。因此

$$\begin{aligned} & s(v_h^- - w_h^-, v_h^+ - w_h^+) \\ &= \alpha \sum_{i=1}^{N} \left(\|\mathcal{D}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \|\mu\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^d \right) (v_h^- - w_h^-)(\boldsymbol{x}_i)(v_h^+ - w_h^+)(\boldsymbol{x}_i) \\ &\ge 0, \end{aligned} \tag{27}$$

这证明了（25）。取$w_h \in V_{\mathcal{P}}^+$ 可得 $w_h^- = 0$，于是（26）由（25）得出。□

定理 3.2（适定性）。设 $T\colon V_{\mathcal{P}} \to [V_{\mathcal{P}}]'$ 为如下定义的映射

$$[Tv_h, w_h] = a(v_h^+, w_h) + s(v_h^-, w_h)\,. \tag{28}$$

那么，T 是连续的；若无量纲参数 α 的选取满足$\alpha \ge C_{\text{equiv}}$，则它也是强单调的，因为此时 T 满足：存在一个与 h无关的 $\beta > 0$，使得

$$[Tv_h - Tw_h, v_h - w_h] \ge \beta \|v_h - w_h\|_a^2\,, \tag{29}$$

对所有v_h，$w_h \in V_{\mathcal{P}}$。因此，（20）有唯一解 $u_h \in V_{\mathcal{P}}$。

证明。我们先定义网格依赖范数 $\|\cdot\|_{\mathcal{P}}$ 如下：

$$\|v_h\|_{\mathcal{P}} := \left\{ \|v_h\|_a^2 + \|v_h\|_s^2 \right\}^{\frac{1}{2}}\,. \tag{30}$$

那么，对于所有 v_h，w_h，$z_h \in V_{\mathcal{P}}$ 利用柯西-施瓦茨不等式和（24），可得

$$\begin{aligned} [Tv_h - Tw_h, z_h] &= a\big(v_h^+ - w_h^+, z_h\big) + s\big(v_h^- - w_h^-, z_h\big) \\ &\le \left(\|v_h^+ - w_h^+\|_a^2 + \|v_h^- - w_h^-\|_s^2 \right)^{\frac{1}{2}} \|z_h\|_{\mathcal{P}}\,, \end{aligned} \tag{31}$$

这证明了 T 的连续性。为证明 T 的单调性，令 v_h， $w_h \in V_{\mathcal{P}}$。利用柯西-施瓦茨不等式、杨氏不等式和（24），可得

$$
\begin{aligned}
&[Tv_h - Tw_h, v_h - w_h] = a(v_h^+ - w_h^+, v_h - w_h) + s(v_h^- - w_h^-, v_h - w_h) \\
&= \|v_h^+ - w_h^+\|_a^2 + a(v_h^+ - w_h^+, v_h^- - w_h^-) + \|v_h^- - w_h^-\|_s^2 + s(v_h^- - w_h^-, v_h^+ - w_h^+) \\
&\geq \frac{1}{2}\|v_h^+ - w_h^+\|_a^2 + \left(1 - \frac{C_{\text{equiv}}}{2\alpha}\right)\|v_h^- - w_h^-\|_s^2 + s(v_h^- - w_h^-, v_h^+ - w_h^+) \\
&\geq \frac{1}{2}\|v_h^+ - w_h^+\|_a^2 + \left(1 - \frac{C_{\text{equiv}}}{2\alpha}\right)\|v_h^- - w_h^-\|_s^2 ,
\end{aligned}
\tag{32}
$$

其中，在最后一个不等式中，我们用到了引理 3.1 中的（25）。接下来，利用 $\alpha \geq C_{\text{equiv}}$ 和（24），可得

$$
\left(1 - \frac{C_{\text{equiv}}}{2\alpha}\right)\|v_h^- - w_h^-\|_s^2 \geq \frac{\alpha}{2C_{\text{equiv}}}\|v_h^- - w_h^-\|_a^2 . \tag{33}
$$

因此，将（33）代入（32），并注意到由于 $a(\cdot, \cdot)$ 是对称且椭圆的，故有 $\|v_h\|_a \leq \sqrt{2}\left(\|v_h^+\|_a + \|v_h^-\|_a\right)$，即可得到（29）。

最后，利用单调算子理论中的经典结果（例如，参见 [29, 定理 10。49]），可得存在性和唯一性。□

对于下面的观察，值得回顾的是，对于 Galerkin 方法，一致性可以表述为将精确解映射到 Galerkin 解的算子具有不变性。

引理 3.3（相容性）。在假设（A1）下，该方法（20）具有如下不变性：如果精确解 u 属于 $V_{\mathcal{P}}$，则 $u_h^+ = u_h = u$。

证明。由于 $u \in V_{\mathcal{P}}$，函数 u^+ 和 u^- 均有定义，并且根据假设（A1），有 $u^+ = u$ 和 $u^- = 0$。因此，我们得到

$$
a_h(u; v_h) = a(u^+, v_h) = a(u, v_h) = \langle f, v_h \rangle_\Omega ,
$$

对所有 $v_h \in V_{\mathcal{P}}$。由定理 3.2可知，（20）的解是唯一的，因此我们得到 $u_h = u$。利用 $u_h^- = u^- = 0$，我们得出 $u_h^+ = u_h$。□

注 3.4。只要 $k = 1$，该方法（20）实际上在整个区域 Ω内都具有保界性，而不仅限于节点处。此外，如果 $\mu > 0$，它还满足离散最大值原理。也就是说，如果 $f \geq 0$，那么 u_h^+ 不可能取得内部负极小值，而会在边界处取得其最小值。如果 $\mu = 0$，则从构造上看，该方法似乎无法保证 u_h^+ 不会取得内部极小值。不过，迄今为止，我们所有的数值实验都未能产生这样的情形。

3.2. 受约束部分的刻画与误差估计。

该方法（20）的一个显著特点是 u_h^+ 可被刻画为定义在闭凸集 $V_{\mathcal{P}}^+$ 上的变分不等式的唯一解。这将在下一个结果中证明。

定理 3.5 （约束部分的刻画）。 设 $u_h \in V_{\mathcal{P}}$ 为 （20）的唯一解。那么， u_h^+ 满足以下变分不等式： $u_h^+ \in V_{\mathcal{P}}^+$ 并满足

$$a(u_h^+, v_h - u_h^+) \geq \langle f, v_h - u_h^+ \rangle_\Omega \qquad \forall\, v_h \in V_{\mathcal{P}}^+ . \tag{34}$$

证明。 利用（20）并取 $v_h \in V_{\mathcal{P}}$、 v_h^- 和 u_h^+ 作为测试函数，我们得到以下等式

$$\begin{aligned} a(u_h^+, v_h) + s(u_h^-, v_h) &= \langle f, v_h \rangle_\Omega , \\ a(u_h^+, v_h^-) + s(u_h^-, v_h^-) &= \langle f, v_h^- \rangle_\Omega , \\ a(u_h^+, u_h^+) + s(u_h^-, u_h^+) &= \langle f, u_h^+ \rangle_\Omega . \end{aligned}$$

将第一个方程分别减去第二个和第三个方程，并利用 $v_h^+ = v_h - v_h^-$ 可得

$$a(u_h^+, v_h^+ - u_h^+) + s(u_h^-, v_h^+ - u_h^+) = \langle f, v_h^+ - u_h^+ \rangle_\Omega , \tag{35}$$

对所有 $v_h^+ \in V_{\mathcal{P}}^+$ 成立。最后，在引理 3.1 中利用（26）可得 $u_h^+ \in V_{\mathcal{P}}^+$ 满足（34），从而完成证明。 □

注 3.6。 根据 *Stampacchia* 定理，问题 （34）可以直接证明具有唯一解。这表明，稳定化方法 （20）与变分不等式 （34）事实上是等价的。

将该方法表述为变分不等式后，便可用标准方法证明最佳逼近误差估计。

定理 3.7 (抽象误差分析**)**。 设 u 是 （(7) 的解，且 $u_h \in V_{\mathcal{P}}$是 （(20)的唯一解。则，

$$\|u - u_h^+\|_a = \min_{v_h \in V_{\mathcal{P}}^+} \|u - v_h\|_a . \tag{36}$$

此外，设 u_h^{FEM}是（标准）有限元方法 (15)的解。那么，负部 u_h^- 满足以下误差估计

$$s(u_h^-, u_h^-)^{\frac{1}{2}} \leq \sqrt{\frac{C_{\mathrm{equiv}}}{\alpha}} \min \left\{ \|u - u_h^+\|_a, \|u_h^{\mathrm{FEM}} - u_h^+\|_a \right\} . \tag{37}$$

证明。 由于 u_h^+ 是（34）的解，因此它特别满足

$$a(u_h^+ - u, v_h - u_h^+) \geq 0 \qquad \forall\, v_h \in V_{\mathcal{P}}^+ . \tag{38}$$

因此，u_h^+ 是 u 在 $V_{\mathcal{P}}^+$ 中相对于由 $a(\cdot, \cdot)$ 诱导的范数的最佳逼近，从而证明了（ u_h^+ 在（36）中陈述的拟最优性。

（37）的证明类似。事实上，利用柯西-施瓦茨不等式和（24），可得

$$s(u_h^-, u_h^-) = a(u - u_h^+, u_h^-) \leq \sqrt{\frac{C_{\mathrm{equiv}}}{\alpha}} \|u - u_h^+\|_a \sqrt{s(u_h^-, u_h^-)} . \tag{39}$$

或者，利用 $a(u - u_h^{\mathrm{FEM}}, u_h^-) = 0$，我们也可以得到

$$s(u_h^-, u_h^-) = a(u_h^{\mathrm{FEM}} - u_h^+, u_h^-) \leq \sqrt{\frac{C_{\mathrm{equiv}}}{\alpha}} \|u_h^{\mathrm{FEM}} - u_h^+\|_a \sqrt{s(u_h^-, u_h^-)} , \tag{40}$$

这就证明了（37）。 □

注 3.8（互补部分的收敛）。 我们可以从两个方面理解最后一个结果。首先， u_h^- 至少以与 u_h^+ 收敛到 u的速度相同的速度趋于零。此外，（37）表明，在某些情况下，这种收敛远快于 u_h^+ 的收敛。更确切地说，考虑分片线性有限元函数的情形，如果网格满足标准伽辽金方法满足离散最大值原理的条件（例如，参见 [7]），那么 $u_h^{\mathrm{FEM}} \in V_{\mathcal{P}}^+$，这意味着 $u_h = u_h^{\mathrm{FEM}}$也是（20）的解。因此，由于（37），对于某些网格及其规则细化，有 $u_h^- = 0$。

注 3.9（受约束部分的最佳逼近）。 鉴于受约束部分 u_h^+ 的最佳逼近结果（36），不存在能量范数下节点值保界的更优有限元函数。因此，（36）是所提出方法对应于经典最佳逼近结果（16）的结果。

我们以简要讨论数值解的概念来结束本节。由于我们主要关注解 u_h 中属于 $V_{\mathcal{P}}^+$ 的部分，即 u_h^+，因此在本文余下部分中，我们将后者视为数值解。“中间”解 u_h 主要作为一种工具，用于将变分不等式替换为整个向量空间V_h上的等式。

4. A problem with nonlinear reaction

为展示所提出的保界方法在一般性方面的潜力，我们现在将 (20) 推广到带有单调非线性项的半线性问题。具体而言，对于 $p \in (1, \infty)$，我们考虑如下问题：求 u 使得

$$
\begin{aligned}
-\operatorname{div}\left(\mathcal{D}\,\nabla u\right) + |u|^{p-2}\,u &= f \qquad \text{in } \Omega\,, \\
u &= 0 \qquad \text{on } \partial\Omega\,,
\end{aligned}
\tag{41}
$$

其中 $\mathcal{D}$ 满足与上文相同的假设。为避免技术上的枝节，我们仅考虑 $p \geq 2$。这类方程有时称为 Lane-Emden-Fowler 方程，并与具有临界指数的问题[13]有关。此外，这类方程还出现在黏性流体的边界层理论 [31],等应用领域。

该问题的弱形式为：求 $u \in \mathcal{X} := H_0^1(\Omega) \cap L^p(\Omega)$，使得

$$
a(u, v) + b(u; u, v) = \langle f, v\rangle_\Omega \qquad \forall\, v \in \mathcal{X}\,,
\tag{42}
$$

其中 $a(\cdot,\ \cdot)$ 由以下公式给出 (10)（本例中 $\mu = 0$），且半线性形式 b 由以下公式给出

$$
b(w; u, v) := (|w|^{p-2}\,u, v)_\Omega\,.
\tag{43}
$$

空间 $\mathcal{X}$ 配备范数

$$
\|v\|_{\mathcal{X}} := |v|_{1,\Omega} + \|v\|_{0,p,\Omega}\,,
\tag{44}
$$

从而使其成为一个自反巴拿赫空间。因此，利用单调算子理论（参见，例如， [29, 第 10]），可以证明该问题存在唯一解。

这类问题的误差分析早在 [18, 19]（在 p-拉普拉斯算子的背景下）就已在多项研究中展开。其中，对于指数 p的某些取值，所得估计并非最优。因此，后续研究（参见，例如，[4]）采用了拟范数的概念，以获得最优误差估计。本文将采用这一方法，因此我们先回顾拟范数的定义。

定义 4.1（拟范数）。设 V 是一个实向量空间。若映射 $\|\cdot\|_{(q)}: V \to \mathbb{R}$ 满足以下条件，则称其为V 中的一个拟范数 $\|\cdot\|_{(q)}$ ：

$$\|v\|_{(q)} \geq 0 \text{ and } \|v\|_{(q)} = 0 \iff v = 0, \tag{45}$$

对所有 $v \in V$ 均成立。不过，通常的三角不等式被以下条件取代：

$$\|v + w\|_{(q)} \leq C\Big(\|v\|_{(q)} + \|w\|_{(q)}\Big), \tag{46}$$

对所有 v、$w \in V$ 均成立，其中 C 可能取决于 $\|\cdot\|_{(q)}$ 的定义，以及元素 v 和 w 本身。

注 4.2。严格来说，我们还应要求该拟范数具有齐次性，即对于所有 $v \in V$ 和所有 $\alpha \in \mathbb{R}$，均有 $\|\alpha v\|_{(q)} = |\alpha| \|v\|_{(q)}$ 。我们将用于度量误差的映射不满足这一性质，但这不会影响下文给出的误差估计。

在下面的分析中，我们将使用 $L^p(\Omega)$ 中的以下拟范数：对于给定的 $w \in L^p(\Omega)$，我们定义

$$\|v\|^2_{(w,p)} := \int_\Omega |v|^2 (|w| + |v|)^{p-2} \, \mathrm{d}x \,, \tag{47}$$

对所有 $v \in L^p(\Omega)$ 均成立。它具有以下性质。

引理4.3。 映射 $\|\cdot\|_{(w,p)}$ 是 $L^p(\Omega)$ 中的拟范数。此外，对于所有v、$w \in L^p(\Omega)$，以下等价关系成立

$$\|v\|^p_{0,p,\Omega} \leq \|v\|^2_{(w,p)} \leq \||v| + |w|\|^{p-2}_{0,p,\Omega} \, \|v\|^2_{0,p,\Omega} \,. \tag{48}$$

证明。 关于 $\|\cdot\|_{(w,p)}$ 是拟范数的证明，与 [14,命题2。1] 的证明完全类似。 [14,等式（10）] 的证明中采用完全相同的步骤即可证明等价关系（48）。 [25, （另见命题3。1]）□

此外，对于非线性型 $b(\cdot;\cdot, \cdot)$，还可以证明以下单调性和连续性结果。

引理 4.4. 非线性型 $b(\cdot; \cdot, \cdot)$ 关于拟范数 (47)是强单调的。更确切地说，存在常数 $C_C > 0$ 使得

$$b(u; u, u - v) - b(v; v, u - v) \geq C_C \|u - v\|^2_{(u,p)} \,. \tag{49}$$

此外，对于任意 $\theta \in (0, 1]$ ， 下式成立

$$|b(u; u, w) - b(v; v, w)| \leq C_B\Big(\theta \|u - v\|^2_{(u,p)} + \theta^{-p+1} \|w\|^2_{(u,p)}\Big) \qquad \forall\, u, v, w \in L^p(\Omega) \,, \tag{50}$$

其中， C_B 仅取决于 p。

证明。 在 [4, 引理 2 中，1] 证明了以下界

$$\left(|x|^{p-2}x-|y|^{p-2}y\right)(x-y)\geq C_2\big(|x|+|y|\big)^{p-2}|x-y|^2\,, \tag{51}$$

$$\left||x|^{p-2}x-|y|^{p-2}y\right|\leq C_1\big(|x|+|y|\big)^{p-2}|x-y|\,, \tag{52}$$

对所有 $x,\ y\in\mathbb{R}$。接下来，式（49）由式（51）并注意到 $|x|+|y|\geq\big(|x|+|x-y|\big)/2$ 即可推出。为证明式（50），我们将式（52）与 [25, 引理 2。2]中的结果结合起来。□

4.1. **有限元方法**。 我们考虑的有限元方法是对式（20）的自然推广，即：求 $u_h\in V_{\mathcal{P}}$ 使得

$$a\big(u_h^+,v_h\big)+b\big(u_h^+;u_h^+,v_h\big)+s\big(u_h^-,v_h\big)=\langle f,v_h\rangle_\Omega\qquad\forall\, v_h\in V_{\mathcal{P}}\,. \tag{53}$$

这里，稳定化项定义为

$$s(v_h,w_h):=\alpha\sum_{i=1}^{N}\|\mathcal{D}\|_{0,\infty,\omega_i}\mathfrak{h}(\boldsymbol{x}_i)^{d-2}\,v_h(\boldsymbol{x}_i)w_h(\boldsymbol{x}_i)\,, \tag{54}$$

其中，再次有 $\alpha>0$ 为一个无量纲常数。按照与定理 3.2 证明中非常相似的步骤，可以证明方法（53）存在唯一解 $u_h\in V_{\mathcal{P}}$。我们注意到，至少当 p 取值适度偏大时，在稳定化（54）中加入线性化反应项似乎没有可观察到的数值优势，因此为简洁起见，我们省略该项。

与线性情形一样，(53) 可以与一个变分不等式联系起来，以下结果（其证明与定理 3.5 的证明完全类似）表明了这一点。

定理 4.5. 设u_h 为 *(53)* 的唯一解。那么， u_h^+ 满足以下变分不等式

$$a\big(u_h^+,v_h-u_h^+\big)+b\big(u_h^+;u_h^+,v_h-u_h^+\big)\geq\big\langle f,v_h-u_h^+\big\rangle_\Omega\qquad\forall\, v_h\in V_{\mathcal{P}}^+. \tag{55}$$

以下结果是使用拟范数而非问题所诱导的范数的主要原因。事实上，由上一个结果可以得到一个 C´ea 型估计，但与 [10], 中所述类似，对于某些 p 值，这将导致次优估计。因此，在下一个结果中，我们给出关于拟范数(47)的最优估计。

定理 4.6. 设 u 满足 *(41)*，且 $u_h\in V_{\mathcal{P}}$ 为 (53) 的解。那么

$$\big\|u-u_h^+\big\|_a^2+\big\|u-u_h^+\big\|_{(u,p)}^2\leq C\inf_{v_h\in V_{\mathcal{P}}^+}\Big(\|u-v_h\|_a^2+\|u-v_h\|_{(u,p)}^2\Big)\,. \tag{56}$$

证明。 利用 b 的强制性 (49)，再结合定理 4.5，得到

$$\begin{aligned}C_C\Big(\big\|u-u_h^+\big\|_a^2+\big\|u-u_h^+\big\|_{(u,p)}^2\Big)&\leq a\big(u-u_h^+,u-u_h^+\big)+b\big(u;u,u-u_h^+\big)-b\big(u_h^+;u_h^+,u-u_h^+\big)\\&\leq a\big(u-u_h^+,u-v_h\big)+b(u;u,u-v_h)-b\big(u_h^+;u_h^+,u-v_h\big)\,,\end{aligned} \tag{57}$$

对每个 $v_h\in V_{\mathcal{P}}^+$。接下来，利用杨氏不等式和（50），可得对所有 $\theta\in(0,1]$,

$$\begin{aligned}C_C\Big(\big\|u-u_h^+\big\|_a^2+\big\|u-u_h^+\big\|_{(u,p)}^2\Big)\leq{}&\frac{C_C}{4}\big\|u-u_h^+\big\|_a^2+C_C\|u-v_h\|_a^2\\&+C_B\Big(\theta\big\|u-u_h^+\big\|_{(u,p)}^2+\theta^{-p+1}\|u-v_h\|_{(u,p)}^2\Big)\,.\end{aligned} \tag{58}$$

取 $\theta = \min\left(\frac{C_C}{4C_B}, \frac{1}{2}\right)$，并重新整理该不等式，即可得到所需结果。
Choo

□

5. Numerical tests

在这些测试中，我们详细介绍实现的一些方面，并通过考察二维中的对称问题（7）及其非线性对应问题（41）展示该方法。我们再次提醒，所有提及数值解的地方，指的都是函数 $u_h^+ \in V_{\mathcal{P}}^+$，而非函数 u_h。

为将问题线性化，我们对（20）提出如下类似 Richardson 方法的迭代近似：给定 u^0 和 $\omega \in (0,1]$, 对每个 $n = 0, 1, \ldots\ldots$ 求 u^{n+1} 使得

$$a(u^{n+1}, v) = a(u^n, v) + \omega\left(\langle f, v\rangle_\Omega - a((u^n)^+, v) - s((u^n)^-, v)\right). \tag{59}$$

我们通过 Galerkin 近似初始化（59）的有限元近似，即令 $u_h^0 \in V_{\mathcal{P}}$ 满足：对所有 $v_h \in V_{\mathcal{P}}$

$$a(u_h^0, v_h) = \langle f, v_h\rangle_\Omega. \tag{60}$$

于是，对于每个 $n = 0, \ldots\ldots, N-1$，迭代（59）的近似变为：求 $u_h^{n+1} \in V_{\mathcal{P}}$ 使得对所有 $v_h \in V_{\mathcal{P}}$

$$a(u_h^{n+1}, v_h) = a(u_h^n, v_h) + \omega\left(\langle f, v_h\rangle_\Omega - a((u_h^n)^+, v_h) - s((u_h^n)^-, v_h)\right). \tag{61}$$

在每个实验中，我们在稳定化中取 $\alpha = 1$。使用 Eigen 库中的 LU 分解求解式（61）所产生的线性系统。当 $\|u_h^{n+1} - u_h^n\|_{0,\Omega} \le 10^{-12}$ 时，终止线性化。

5.1. 光滑解在规则网格上的收敛。我们首先考虑$\mathcal{D} = \epsilon\mathcal{I}$，其中 $\mathcal{I}$ 表示 2×2 单位矩阵，令 $\epsilon = 10^{-5}$、$\mu = 1$，并设置 f，使得函数

$$u(x,y) = \sin(\pi x)\sin(\pi y) \tag{62}$$

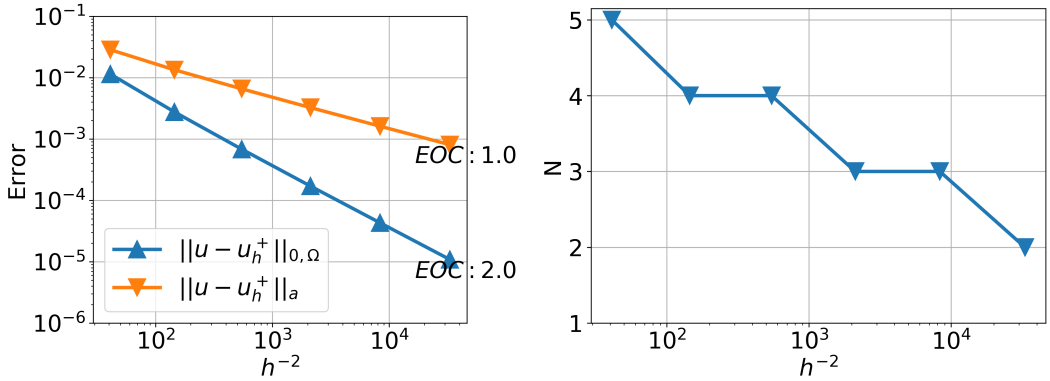
在单位正方形上求解问题（7）。

在一系列均匀细化的交叉网格上，分段线性 $k = 1$和二次 $k = 2$单元的收敛结果分别见图1 和2。与第3.2节的误差估计一致，该方法对于 $k = 1$ 和 $k = 2$，在函数逼近意义下均以最优收敛率收敛。还请注意，线性化的迭代次数随 h增加而减少。此处及下文中，EOC 表示估计收敛阶。

5.2. **在钝角网格上对光滑解的收敛性。**我们再次考虑 $\mathcal{D} = \epsilon\mathcal{I}$，其中 $\epsilon = 10^{-5}$，$\mu = 1$。(7) 在 $(-1,\ 1)\times(0,\ 1)$ 上的解析解取为

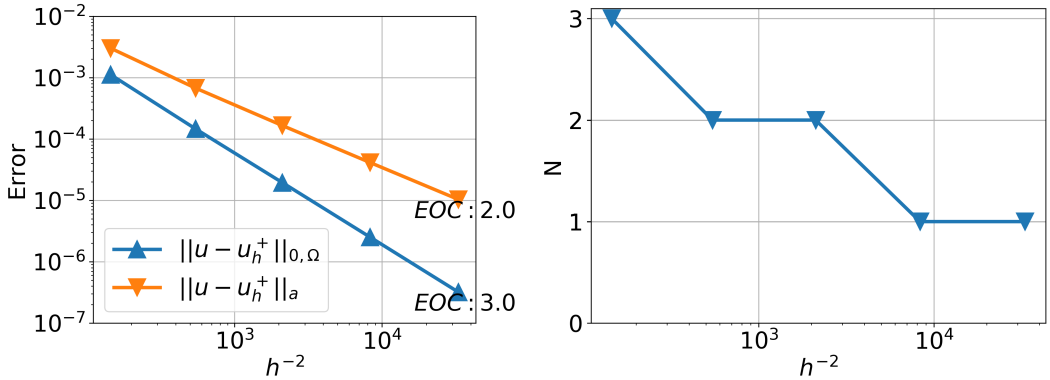
$$u(x,y) = \sin(\pi(x+1)/2)\sin(\pi y)\,, \tag{63}$$

并据此计算 f。注意，$u(\boldsymbol{x}) \in [0,1]$ 对所有 $\boldsymbol{x} \in \Omega$。我们在一个含钝角单元的三角剖分上提出该问题，如 [6]图3所示。这在 [6] 中被用作三角剖分的一个例子；对于这类三角剖分，即使是



(A) 在 L^2 范数和能量范数下的收敛。 (B) 达到收敛所需的迭代次数随网格尺寸的变化。

FIGURE 1. 我们测试由（61）给出的分片线性协调逼近在一系列同步细化的交叉网格上的收敛性。Richardson迭代在 $\omega = 1$时收敛。

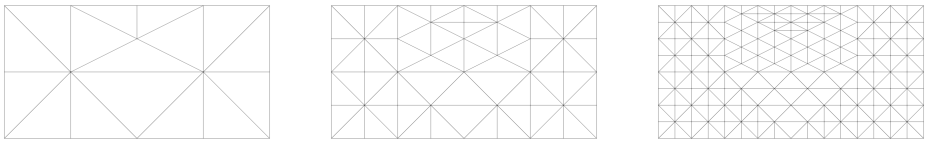


(A) 在 L^2 范数和能量范数下的收敛。 (B) 达到收敛所需的迭代次数随网格尺寸的变化。

FIGURE 2. 我们在一系列同步加密的交叉网格上检验分片二次协调近似解（61）。Richardson迭代在 $\omega = 1$ 上收敛。

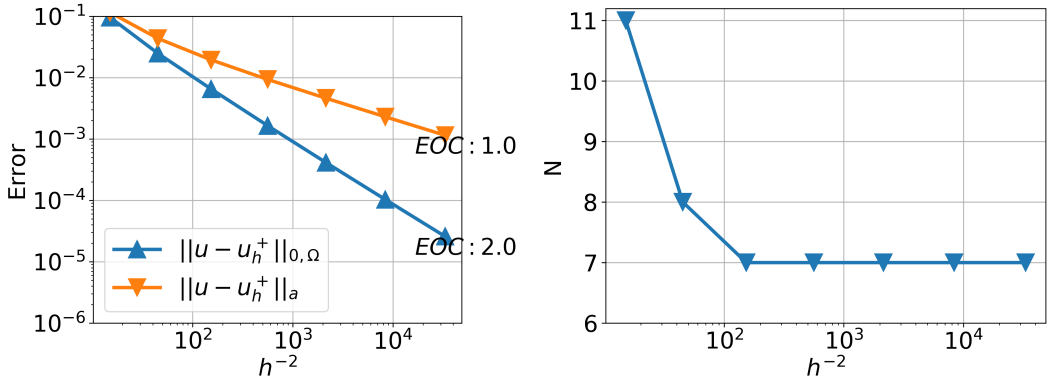
泊松方程；因此，这对有限元方法构成了挑战，因为即使 $\epsilon \gg 1$，其解通常也不满足离散最大值原理。

该网格上分片线性元的收敛结果见图 4。该方法具有最优收敛性，高阶元也观察到类似结果。请注意，线性化需要更多次迭代才能达到收敛，这是意料之中的，因为无论网格多么精细，伽辽金解很可能永远不会满足问题的界限（尽管迭代次数仍然很少）。



(A) 迭代 1。 (B) 迭代 2。 (C) 迭代 3。

FIGURE 3. 对含有一个钝角三角形的网格进行三次细化，来自 [6]。结果形成了一层钝角三角形。



(A) 在 L^2 范数和能量范数下的收敛性。 (B) 达到收敛所需的迭代次数随网格尺寸的变化。

FIGURE 4. 我们在一系列含有钝角单元的同步细化网格上，测试分片线性协调近似（61）的收敛性，如图3所示。这里，Richardson线性化在 $\omega = 1$时收敛。

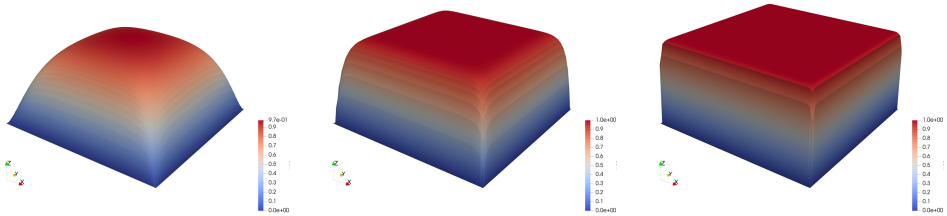
5.3. **边界层的分辨**。考虑问题

$$\begin{aligned} -\epsilon\Delta u + u &= 1 \text{ in } \Omega\,, \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{64}$$

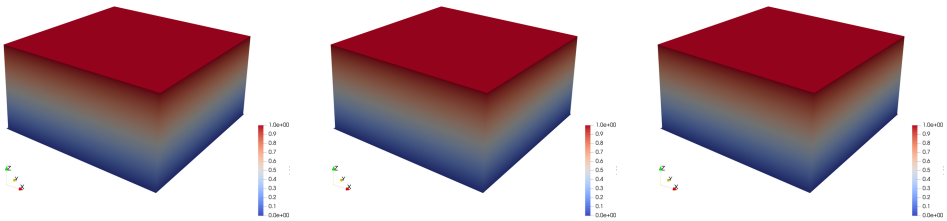
我们在交叉网格上固定 $h \approx 0.02$ ，并改变 $\epsilon \in [10^{-2}, 10^{-7}]$。当$\epsilon$ 特别小时，Richardson迭代需要进行阻尼才能收敛。当 $\epsilon > 10^{-5}$时，我们采用 $\omega = 1$ ，并在4次迭代内实现收敛。当 $\epsilon \leq 10^{-5}$时，$\omega = \frac{1}{2}$ 即可实现收敛，且每种情况下的迭代次数均少于46次。最具挑战性的情况对应于 ϵ的最小值。不同 ϵ 取值对应的计算解如图5所示。

5.4.**具有不连续狄利克雷条件的边界层的解析**。考虑问题

$$\begin{aligned} -\epsilon\Delta u + u &= 0 \text{ in } \Omega\,, \\ u &= g_D \text{ on } \partial\Omega, \end{aligned} \tag{65}$$



(A) $\epsilon = 10^{-2}$. (B) $\epsilon = 10^{-3}$. (C) $\epsilon = 10^{-4}$.



(D) $\epsilon = 10^{-5}$。 (E) $\epsilon = 10^{-6}$。 (F) $\epsilon = 10^{-7}$。

FIGURE 5. 近似解的高程（64），固定 h 且 ϵ取不同值时。我们注意到，即使 ϵ取值特别小时，也没有振荡。

其中 $g_D = 1$在 $[0, \frac{1}{2}] \times \{0\}$， $g_D = 0$在（12, 1] $\times \{0\}$上，并以逆时针方向周期性地遵循相同模式。我们在交叉网格上固定 $h \approx 0.02$，并改变 $\epsilon \in [10^{-2}, 10^{-7}]$。对于特别小的 ϵ，Richardson迭代需要阻尼才能收敛。对于 $\epsilon > 10^{-5}$，我们使用了 $\omega = 1$，并在5次迭代内实现收敛。当 $\epsilon \leq 10^{-5}$时，则 $\omega = 0.5$ 提供了一个收敛算法，且在每种情况下迭代次数均少于40次。图6展示了一些解。

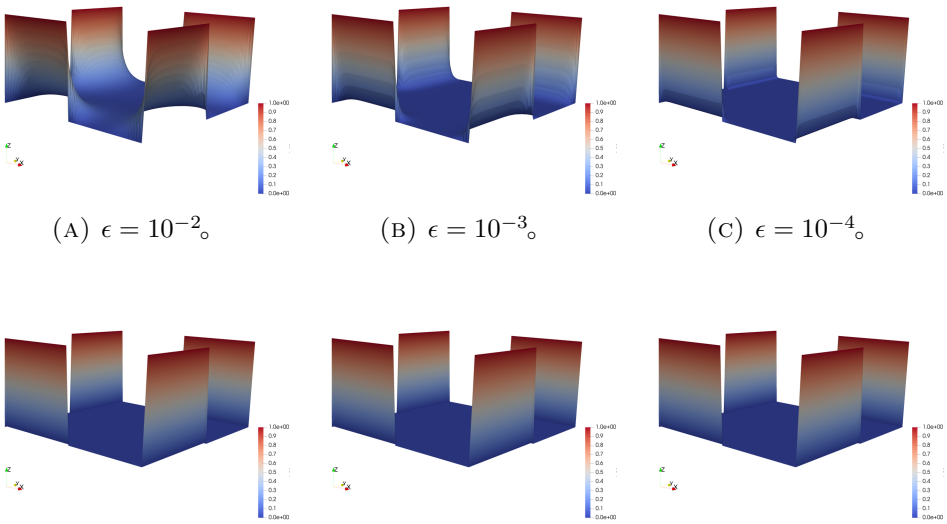
5.5. **具有内部层的解**。考虑问题

$$\begin{aligned} -\epsilon \Delta u + u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{66}$$

with

$$f = \begin{cases} \frac{1}{2} \text{ in } [\frac{1}{4}, \frac{3}{4}]^2 \\ 1 \text{ otherwise.} \end{cases} \tag{67}$$

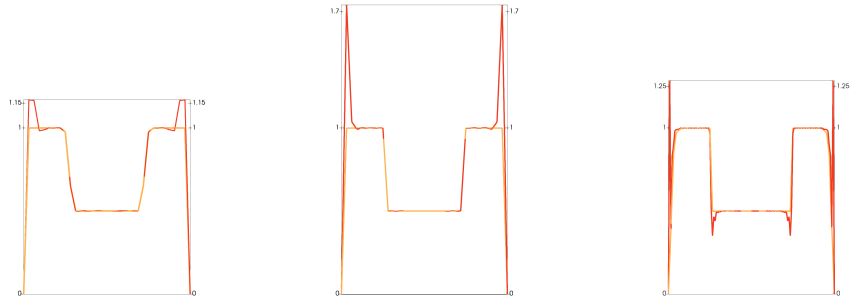
在这种情况下，预计解会在区域内部达到局部极小值。我们固定 $h \approx 0.02$，并考察 $\epsilon = \{10^{-4}, 10^{-7}\}$. 时的解。图 61）给出的近似解与标准有限元解的比较见图7。注意，普通伽辽金解在边界层附近存在振荡，当 $\epsilon \ll 1$时，振荡变得极为剧烈，而本方法完全消除了这些振荡。此外，当 $k = 2$时，内部层附近存在明显的下冲，而本方法也完全消除了这些下冲。



(A) $\epsilon = 10^{-2}$。 (B) $\epsilon = 10^{-3}$。 (C) $\epsilon = 10^{-4}$。

(D) $\epsilon = 10^{-5}$。 (E) $\epsilon = 10^{-6}$。 (F) $\epsilon = 10^{-7}$。

FIGURE 6. 固定65时，对 h 并改变 ϵ得到的近似解高程。即使 ϵ特别小，也没有明显的振荡。



(A) $\epsilon = 10^{-4}$, $k = 1$。 (B) $\epsilon = 10^{-7}$, $k = 1$。 (C) $\epsilon = 10^{-7}$, $k = 2$。

FIGURE 7. 沿 $x = y$ 平面截取的横截面。有限元解以红色表示，在边界层处呈现振荡；而式 (61) 给出的近似解以绿色表示，在边界处没有振荡。面板 (A) 和 (B) 使用$k = 1$计算，而 (C) 展示了使用 $k = 2$ 得到的近似解。

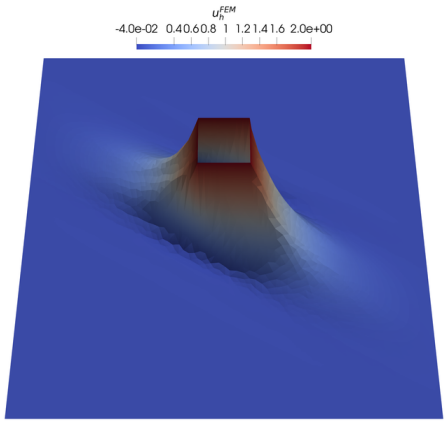
5.6. 含非线性反应的各向异性扩散。考虑区域 $\Omega = \Omega_1 \backslash \Omega_2$，其中 $\Omega_1 = (0,1)^2, \Omega_2 = [\frac{4}{9}, \frac{5}{9}]^2$，以及如下问题

$$-\operatorname{div}(\epsilon \mathcal{D} \nabla u) + u^3 = f \text{ in } \Omega, \quad u = \begin{cases} 0 \text{ on } \partial\Omega_1 \\ 2 \text{ on } \partial\Omega_2 \end{cases}, \tag{68}$$

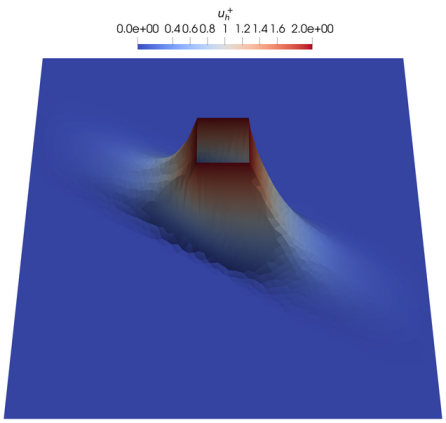
与 $\epsilon = 10^{-5}$,

$$\mathcal{D} = \mathcal{R}\mathcal{A}\mathcal{R}^T, \quad \mathcal{R} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \quad \mathcal{A} = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix}, \tag{69}$$

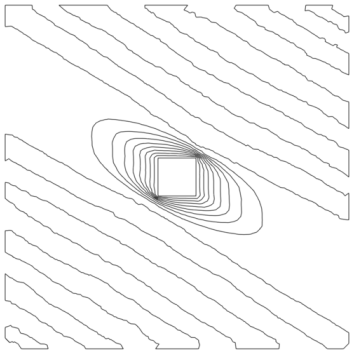
以及$\theta = -\frac{\pi}{6}$。这是对具有各向异性扩散系数和非线性反应项的（41）的一种颇具挑战性的实现，其中 $p = 4$，问题定义在一个非凸区域上。在图8 中，我们展示了该问题的有限元解、保界解，以及凸显有限元解振荡特性的等高线图。



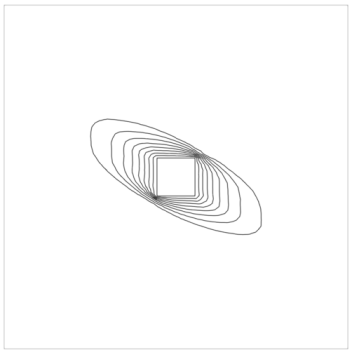
(A) 有限元解。



(B) 保界解。



(C) 有限元解的等值线图。



(D) 保界解的等值线图。

FIGURE 8. 问题（68）的 $h \approx 0.02$ 德劳内网格上的近似解。请注意，与保界方法相比，有限元近似解呈现出振荡行为。

6. Concluding remarks

我们提出了一种经济而简单的方法，用于对有限元解的取值范围施加硬界。该方法通过定义一种非线性稳定化伽辽金方法来实现，其设计目标是将解正交投影到满足硬界的物理可容许解所构成的闭凸集上。为了突出核心思想，我们将本文的讨论限定在线性和单调半线性反应扩散方程上。不过，我们强调，该框架具有足够的通用性，可以处理其他模型问题，也可以依照本文提出的方法构造其他已知有限元方法的保界变体。特别是，将所提出的方法推广至对流主导问题既有现实意义，而且我们认为也切实可行。此外，也可以考虑将其推广至定义在一般多边形/多面体网格上的间断伽辽金方法。这些以及其他课题目前仍在研究中，将另文讨论。

Acknowledgements

GRB 的研究工作由 Leverhulme Trust 通过研究员奖学金 RF-2019-510 资助。AV 的研究工作得到意大利 GNCS 和 MIUR PRIN 2017 NA-FROM-PDEs 的支持。

References

[1] G. R. Barrenechea、E. Burman 和 F. Karakatsani。用于对流扩散方程有限元近似的基于边的非线性扩散及其与代数通量修正格式的关系。 *Numer. Math.*，135(2):521–545，2017年。[2] G. R. Barrenechea、V. John 和 P. Knobloch。满足离散最大值原理并在线性保持方面适用于一般网格的代数通量修正格式。 *Math. Models Methods Appl. Sci.* 27 3 525 548, 2017 , ( ): – 。

[3] G. R. Barrenechea、V. John 和 P. Knobloch。用于对流扩散方程有限元近似的基于边的非线性扩散及其与代数通量修正格式的关系。 *SIAM Review*，2023年。即将发表。[4] J. W. Barrett 和 W. Liu。p-拉普拉斯算子的有限元近似。 *Mathematics of Computation* 61 523 537, 1993 , : – 。

[5] P. Bochev、D. Ridzal、M. D'Elia、M. Perego 和 K. Peterson。面向标量对流方程的基于优化、保持性质的有限元方法及其与代数通量修正的联系。 *Computer Methods in Applied Mechanics and Engineering* 367 112982, , : 。[6] 2020 J. Brandts、S. Korotov、M. K˘r ´ *SIAM* ı˘zek 和 *Review* 51 2 317 335, 2009 J. ˘Solc。关于非钝角单纯形剖分。 , ( ): – 。[7] J. H. Brandts、S. Korotov 和 M. K˘r´ı˘zek。反应扩散问题的线性单纯形有限元近似的离散最大值原理。 *Linear Algebra Appl.*，429 10 2344 2357，2008年 ( ): – 。[8] E. Burman 和 A. Ern。对流扩散反应方程的稳定化伽辽金近似：离散最大值原理与收敛性。 *Math. Comp.*，74(252):1637–1652（电子版），2005年。[9] E. Burman 和 A. Ern。在线性元近似中弱施加正性的非线性一致罚方法，用于输运方程。 *Computer Methods in Applied Mechanics and Engineering* 320 122 132, 2017 , : – 。

[10] P. Ciarlet。 *The Finite Element Method for Elliptic Problems*。Classics in Applied Mathematics。工业与应用数学学会，2002年。[11] P. G. Ciarlet。有限差分算子的离散最大值原理。 *Aequationes Math.*，4 338 352，1970年 : – 。

[12] P. G. Ciarlet 和 P.-A. Raviart。有限元方法的最大值原理与一致收敛。 *Comput. Methods Appl. Mech. Engrg.*，2:17–31，1973年。[13] P. Cl´ement、D. G. de Figueiredo 和 E. Mitidieri。具有临界指数的拟线性椭圆方程。 *Topological Methods in Nonlinear Analysis*，7(1):133–170，1996年。[14] C. Ebmeyer 和 W. Liu。p-Laplacian问题的分片线性有限元近似的拟范数插值误差估计。 *Numerische Mathematik*，100(2):233–258，2005年。[15] A. Ern 和 J.-L. Guermond。 *Finite Elements I. Approximation and Interpolation*， *Texts in Applied Mathematics*的第72卷。Springer-Verlag，2021年。[16] J. A. Evans、T. J. Hughes 和 G. Sangalli。变分多尺度方法中的约束施加与最大值原理。 *Computer Methods in Applied Mechanics and Engineering* 199 1 61 76, 2009 ，( ): – 。

[17] D. Gilbarg 和 N. S. Trudinger。 *Elliptic partial differential equations of second order*。Springer-Verlag，2001年。1998年版重印。[18] R. Glowinski 和 A. Marrocco。关于通过 ´el´ements finis d'ordre 1进行近似，以及通过罚函数-对偶方法求解一类非线性Dirichlet问题。 ，Ser. A : – 。[19] R. Glowinski 和 A. Marrocco。关于通过 ´el´ements finis d'ordre un进行近似，以及通过p´enalisation-dualit´e求解一类非线性Dirichlet问题。 *R.A.I.R.O.*，9(R-2):41–76，1975年8月。[20] W. Huang。二维各向异性扩散问题线性有限元近似的离散最大值原理与Delaunay型网格条件。 *Numer. Math.* ，( ): – 。 *Theory Methods Appl.* 4 3 319 334, 2011

[21] M. Jensen 和 A. Malquist。具有混合边界条件的焦耳热问题的有限元收敛。 *BIT Numerical Mathematics*，53:475–496，2013年。[22] F. Kikuchi。对流扩散方程有限元近似中的离散最大值原理与人工黏性。 *Institute of Space and Aeronautical Science University of* , , 。 *Tokyo* 550 1977

[23] C. Kreuzer。关于为何通过简单的后验截断施加离散最大值原理是一个好主意的注记。 *Numerical Methods for Partial Differential Equations* 30 3 994 1002, ，( ): – 。 2014

[24] R. Lisa 和 M. Shashkov。对二阶椭圆问题线性有限元解施加离散最大值原理。 *Communications in Computational Physics*，3(4):852–877，2008年。[25] W. Liu。弹塑性力学中双材料问题所产生的非线性椭圆方程的有限元近似。 *Numerische Mathematik*，86(2):491–506，2000年。[26] C. Lu、W. Huang 和 E. S. V. Vleck。用于数值计算抛物型偏微分方程非负解的截断法及其在各向异性扩散和润滑型方程中的应用。 *J. Comput. Phys.*，242:24–36，2013年。[27] A. Mizukami 和 T. J. R. Hughes。用于对流占优流的Petrov-Galerkin有限元方法：满足最大值原理的精确迎风技术。 *Comput. Methods Appl. Mech. Engrg.* 50 2 181 193, 1985 ，( ): – 。

[28] M. Mudunuru 和 K. Nakshatrala。有限元方法下对流–扩散–反应方程中最大值原理的施加与逐单元物种守恒的实现。 *Journal of Computational Physics*，305:448–493，2016年。[29] M. Renardy 和 R. C. Rogers。 *An introduction to partial differential equations*，*Texts in Applied Mathematics*的第13卷。Springer-Verlag，纽约，第二版，2004年。[30] H.-G. Roos、M. Stynes 和 L. Tobiska。 *Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion–Reaction and Flow Problems. 2nd ed.* Springer-Verlag，柏林，2008年。[31] J. S. Wong。关于广义Emden–Fowler方程。 *Siam Review*，17(2):339–360，1975年。[32] J. Yang、Z. Yuan 和 Z. Zhou。带有截断后处理的Allen–Cahn方程任意高阶最大界保持格式。 *Journal of Scientific Computing*，90:文章编号76，2022年。

(G.R.B.) Department of Mathematics and Statistics University of Strathclyde 26 , , , , Richmond Street Glasgow G1 1XH Scotland.
gabriel.barrenechea@strath.ac.uk

(E.G.) 1. School of Mathematical & Computer Sciences, Heriot Watt University, United Kingdom School of Applied Mathematical and , Physical Sciences National Technical University of Athens Greece and 3. IACM-
2. Department of Mathematics, , „FORTH, GreeceE.Georgoulis@hw.ac.uk

(T.P.) Department of Mathematical Sciences, University of Bath, Claverton down, Bath BA2 7AY, UKtmp38@bath.ac.uk

(A.V.) Dipartimento di Matematica, Università degli Studi di Milano, Via C. Saldini50, 20133 Milano, Italy andreas.veeser@unimi.it