北京大学

博士研究生学位论文

题目：基于离散最小二乘重构的椭圆型方程高阶数值方法

姓　　名：唐凤阳

学　　号：1101110028

院　　系：数学科学学院

专　　业：计算数学

研究方向：偏微分方程数值计算

导　　师：李若 教授　明平兵 研究员

二〇一五年六月

版权声明

摘要

最小二乘方法广泛应用于科学实验与工程计算中，特别是曲线拟合、函数逼近、数据处理、最优化、方差分析与回归分析等领域. 在偏微分方程的数值计算中，多项式逼近或重构的思想也已经有不少应用. 本文将主要考虑基于最小二乘重构的高阶异质多尺度有限元方法和高阶间断 Galerkin 方法.

基于最小二乘方法，本文定义了一类重构算子. 它们描述的是在给定的区域和网格上，通过最小二乘的方式来对采样点上的采样值向量进行多项式拟合，或者对区域上的连续函数进行多项式逼近. 本文给出了重构算子的连续性和逼近性的证明，证明过程中涉及到的重构常数的大小则表征着重构过程的稳定性. 为此，本文在拟一致多边形网格的情况下证明了重构常数的有界性，并对其上界进行了估计. 而且对于三角形单元网格的特殊情况，对上述证明和估计进行了改进. 这些结论说明总是可以通过扩大单元模板的规模来控制重构常数, 随后的数值算例验证了上述结论.

在异质多尺度有限元方法的框架下，本文发展了一种基于局部最小二乘重构的高阶方法，提高了异质多尺度有限元方法的计算效率. 其基本思想是：首先在网格的节点上构造单胞问题，并使用有限元方法进行求解，获得宏观尺度问题的有效系数；然后利用顶点上获得的有效系数进行无约束的或带等式约束的最小二乘重构，从而得到宏观问题在整个计算区域上的近似有效系数. 这样一来，该方法就可使得宏观问题高阶求解器的主要计算量仅仅依赖于网格的节点数目，因此计算效率得到了数倍的提升. 随后，本文从理论分析和数值实验两方面共同说明了该方法的有效性和高阶收敛性.

对于经典的二阶椭圆型方程，本文提出了一种基于最小二乘重构的高阶间断 Galerkin 方法. 该方法中的基函数是间断的分片多项式，而且这些基函数在各个单元上的限制是通过局部最小二乘重构来获得的. 由这些基函数张成逼近空间，然后结合椭圆问题在间断 Galerkin 框架下的变分形式即可得到椭圆方程的逼近形式. 上述方法构造的逼近空间的维度和线性系统的规模将始终等于网格中单元的个数. 这样一来只需要提高重构多项式的次数就可得到高阶格式，而不需要增加自由度的个数. 而且该方法对网格单元的几何形状没有任何要求，因此可以应用到任意的多边形单元（多面体单元）网格上. 本文对该方法的稳定性和收敛性进

行了理论分析，并通过一系列的数值算例进行了验证. 值得一提的是，由于每个单元上的局部最小二乘重构都是独立的，因此该方法天然地适合进行并行计算.

关键词：离散最小二乘重构, 异质多尺度方法, 间断 Galerkin 方法, 二阶椭圆型方程, 高阶数值方法

High Order Numerical Methods for Elliptic Equations Based on Discrete Least Squares Reconstructions

Tang Fengyang (Computational Mathematics)
Directed by Prof. Ruo Li Prof. Ping-Bing Ming

Abstract

The least-squares methods are widely used in scientific and engineering computing, such as curve fitting, function approximation, data processing, optimization, variance analysis and regression analysis. There are abundant applications of polynomial approximations or reconstructions in the field of numerical partial differential equations. This thesis will focus on the high order numerical methods for heterogeneous multi-scale methods and discontinuous Galerkin methods by least-squares reconstructions.

The first part of the main content defines the least-squares based reconstruction operators, which map the vectors of sampling value, or continuous functions to piecewise polynomials. The properties of continuity and approximation of the operators are proved. And the stability of the reconstruction operators essentially relies on the uniform boundedness of the reconstruction constants. Under several assumptions the thesis proved that the reconstruction constants can be uniformly bounded over general polygon meshes. Such bounds can be improved for triangular meshes. These results imply that one can always reduce the constants by expanding the element patches. Extensive numerical experiments confirmed these results.

In the second part, the thesis proposes a least-squares based high order heterogeneous multi-scale finite element method. The effective matrix is locally reconstructed by least-squares method by the data retrieved from the solutions of cell

problems posed on the nodes of the triangulation. The method achieves high order accuracy for high order macroscopic solver with essentially the same cost as the linear macroscopic solver. Both two and three dimensional numerical experiments are tested in the thesis. Numerical results demonstrate that the proposed method significantly reduces the cost without loss of accuracy.

For the classical elliptic problems, the thesis proposes a high order discontinuous Galerkin method based on least-squares reconstructions. The basis functions are discontinuous piecewise polynomials, the confinement of which over each element are obtained through the local least-squares reconstruction. The dimension of the approximation space is always equal to the number of the elements in the mesh. High order accuracy is achieved by increasing the degree of the reconstructed polynomials. The method can be easily applied to any polygonal and polyhedral meshes. Both the analysis and the numerical experiments confirm the stability and convergence of the proposed method. Moreover, since the local least-squares reconstructions in each element are carried out separately, the method is suitable for parallel computing.

Keywords: Discrete Least Squares Reconstruction, Heterogeneous Multiscale Method, Discontinuous Galerkin Method, Reconstructed Basis Functions, Second Order Elliptic Equations, High Order Numerical Methods

目录

表格

插图

第一章　绪论

在 1794 年，德国数学家 C.F. Gauss 就将最小二乘法成功地应用于天文观测和大地测量的工作中. 此后二百年，它广泛应用于科学实验与工程计算中. 随着电子计算机的普及与发展，这个古老方法更加显示出其强大的生命力. 它在曲线拟合、函数逼近、数据处理、最优化、方差分析与回归分析中都经常用到.

1.1　最小二乘问题的基本原理

在科学技术的许多领域中，常会遇到下面类型的问题:

设已经通过某种实验观测手段或某种计算方法得到了 m 组数据

$$(\boldsymbol{x}_i, y_i), \quad \boldsymbol{x}_i \in \mathbb{R}^d,\ y_i \in \mathbb{R}, \quad i = 1, \cdots, m, \tag{1.1}$$

其中 d 是采样空间的维度. 实际问题中往往希望能够用一个函数

$$y = f(\boldsymbol{x}; \boldsymbol{b}), \quad \boldsymbol{b} = (b_1, b_2, \cdots, b_n)^{\mathrm{T}} \in \mathbb{R}^n. \tag{1.2}$$

来尽可能准确地描述 y_i 与 $\boldsymbol{x}_i$ 之间的变化关系，其中 n 为正整数且上标 T 表示转置.

这个函数关系往往由实际问题具体确定，或由理论建立，或由经验总结，或由猜测得到，因而(1.2)被称为理论函数曲线或经验公式. 在函数关系中含有 n 个未知参数 $b_1, \cdots, b_n$，如何根据(1.1)来寻求参数的最佳估计 $\hat{b}_1, \ldots, \hat{b}_n$，即寻求最佳的理论曲线 $y = f(\boldsymbol{x}; \hat{b}_1, \cdots, \hat{b}_n)$，这就是通常所说的观测数据的曲线拟合问题. 也可称为观测数据的平滑问题.

一般来说，在实际问题中总有 $m > n$. 此时，方程组

$$\begin{cases} f(\boldsymbol{x}_1;\boldsymbol{b}) = y_1, \\ f(\boldsymbol{x}_2;\boldsymbol{b}) = y_2, \\ \qquad \cdots \\ f(\boldsymbol{x}_m;\boldsymbol{b}) = y_m, \end{cases} \tag{1.3}$$

为超定方程. 一般情况下，(1.3)不存在通常意义下的解，因此又被称为矛盾方程组. 于是，人们转而寻求参数 $\hat{b}$，使得残差向量 $\boldsymbol{r}$ 在某种范数意义下最小，其中

$$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{b}), \tag{1.4}$$

这里

$$\begin{aligned} &\boldsymbol{r} = (r_1, r_2, \cdots, r_m)^{\mathrm{T}},\ \boldsymbol{b} = (b_1, b_2, \cdots, b_n)^{\mathrm{T}}, \\ &\boldsymbol{f}(\boldsymbol{b}) = (f_1(\boldsymbol{b}), f_2(\boldsymbol{b}), \cdots, f_m(\boldsymbol{b}))^{\mathrm{T}},\ f_i(\boldsymbol{b}) = f(\boldsymbol{x}_i;\boldsymbol{b}). \end{aligned}$$

定义欧式范数为

$$\|\boldsymbol{r}\|_2 := \left[\sum_{i=1}^{m} r_i^2\right]^{\frac{1}{2}}.$$

于是所谓的最小二乘方法就是指选择合适的参数，使得残差向量的欧式范数最小. 此时问题转化为：求出参数 $\hat{\boldsymbol{b}}$，使得

$$\hat{\boldsymbol{b}} = \arg\ \min_{\boldsymbol{b}} \sum_{i=1}^{m} [y_i - f(\boldsymbol{x}_i;\boldsymbol{b})]^2, \tag{1.5}$$

其中，当 $f(\boldsymbol{x};\boldsymbol{b})$ 是 $\boldsymbol{b}$ 的线性函数时，(1.5)被称为线性最小二乘问题. 反之，当 $f(\boldsymbol{x};\boldsymbol{b})$ 是 $\boldsymbol{b}$ 的非线性函数时，(1.5)被称为非线性最小二乘问题.

线性最小二乘问题的求解方法一直是人们关注的重点. 当 $f(\boldsymbol{x};\boldsymbol{b})$ 关于 $\boldsymbol{b}$ 为线性时，设

$$f(\boldsymbol{x};\boldsymbol{b}) = \sum_{i=1}^{n} b_i \varphi_i(\boldsymbol{x}).$$

于是矛盾方程组(1.3)转化为线性系统

$$\Psi \boldsymbol{b} = \boldsymbol{y}, \tag{1.6}$$

其中的系数矩阵为

$$\Psi = \begin{pmatrix} \varphi_1(\boldsymbol{x}_1) & \varphi_2(\boldsymbol{x}_1) & \cdots & \varphi_n(\boldsymbol{x}_1) \\ \varphi_1(\boldsymbol{x}_2) & \varphi_2(\boldsymbol{x}_2) & \cdots & \varphi_n(\boldsymbol{x}_2) \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_1(\boldsymbol{x}_N) & \varphi_2(\boldsymbol{x}_N) & \cdots & \varphi_n(\boldsymbol{x}_N) \end{pmatrix}.$$

由于超定方程组(1.6)的解很可能不存在，因此人们往往通过求解方程组

$$\Psi^{\mathrm{T}}\Psi\boldsymbol{b} = \Psi^{\mathrm{T}}\boldsymbol{y} \tag{1.7}$$

来给出(1.6)在最小二乘意义下的解. (1.7)被称为法方程组，上述方法又被称为法方程组方法. 在实际使用中，该方法存在一定的问题: 一方面，得到的法方程组有可能严重病态从而导致不能稳定求解 [106, 91]. 为此，Golub[59] 主张直接从矛盾方程 (1.3)入手，通过 QR 分解的方法来计算问题的解. 后续许多学者对该方法的稳定性进行了探讨 [75, 65, 23]. 另一方面，在实际问题中往往需要求解大型稀疏最小二乘问题，而法方程组中系数矩阵的稀疏性往往会被破坏. 为了克服这一缺点，多采用迭代法对矛盾方程组进行求解，比如超松弛迭代法 (SOR)[89, 104]，共轭梯度法 [21, 64, 102] 等. 关于这方面更具体的讨论可参考 [22].

对于非线性最小二乘问题，最经典的方法是 Gauss-Newton 方法，其基本思想是对(1.2)中的 $f(\boldsymbol{x};\boldsymbol{b})$ 在初值 $\boldsymbol{b}^{(0)}$ 处做线性化来得到一个线性问题，由该问题求出下降方向 $\Delta\boldsymbol{b}$ 来更新 $\boldsymbol{b}^{(0)}$ 得到 $\boldsymbol{b}^{(1)}$，并且不断地重复上述过程来逼近最小二乘问题的解 $\boldsymbol{b}^*$. 但是这样得到的点列 $\{\boldsymbol{b}^{(k)}\}$ 并不一定收敛，它的收敛性非常依赖初值 $\boldsymbol{b}^{(0)}$ 的选取. 为此人们提出多种方法来对此进行改进. Hartley[62] 提出了修正的 Gauss-Newton 方法，该方法用二次插值的线性搜索方法来搜索步长因子 $\alpha^{(k)}$，并用 $\alpha^{(k)}\Delta\boldsymbol{b}^{(k)}$ 来修正 $\boldsymbol{b}^{(k)}$. Levenberg[77] 和 Marquardt[85] 则分别提出了阻尼最小二乘方法，引入阻尼因子 $\lambda^{(k)}$，改进了下降方向的计算方式，避免了下降方向无法求解的问题. Fletcher[57] 在 Marquardt 的方法基础上，引入比例 R 来作为何时增减 $\lambda^{(k)}$ 的一个指标. Meyer 和 Roth[86] 将步长因子 $\alpha^{(k)}$ 和阻尼因子 $\lambda^{(k)}$ 结合到一起，提出了阻尼修正最小二乘法. 该方法结合了前几种方法各自的优点，是一种可靠而有效的方法. 在 Marquardt 方法中，每次为了确定 $\lambda^{(k)}$ 都要计算一系列的线性方程组. A. Jones[71] 提出了螺线法来对上述问题进行改进. 如果在 $f(\boldsymbol{x};\boldsymbol{b})$ 关于 $\boldsymbol{b}$ 的展开式中保留到二次项，就得到了 Newton 法. 虽然该方法在收敛效果上要优于 Gauss-Newton 法，但是其计算过程中需要用到 f 的 Hesse 矩阵，而这在实际问题中很难精确计算. 为此，不少学者考虑用拟 Newton 法 [19, 20, 44] 来求解非线性最小二乘问题，通过迭代得到近似 Hesse 矩阵，避免

了求二阶导数的困难.

1.2　最小二乘方法在偏微分方程数值计算中的应用

在偏微分方程的数值计算中，用多项式逼近或重构的思想已有不少应用. Zienkiewicz 和朱建忠在 [116] 中采用了局部多项式重构来恢复局部的梯度信息. 张智民和 Naga [115] 则应用局部多项式重构来得到数值解的导数的超收敛性. Bochev 和 Gunzburger[24] 提出了最小二乘有限元方法，他们通过在整个区域上对能量泛函求最小二乘意义下的极小值来给出椭圆方程的数值解. ENO/WENO 方法中，需要在每个时间步上给出单元中数值积分点上的函数值，而该值在高维情形下则往往通过在单元模板上对单元平均做最小二乘重构来给出 [116, 68, 82].

本文主要从两个角度来考虑最小二乘方法的应用. 首先，对于异质多尺度方法，通过最小二乘方法来重构出宏观的有效系数矩阵，并由此给出一个高阶的异质多尺度有限元方法. 其次，考虑经典的二阶椭圆方程，采用最小二乘重构可以给出任意多边形网格上的任意阶的基函数，并在间断 Galerkin 方法的框架下求出方程的数值解. 下面将对这两个方面的问题背景和研究现状分别作简单的回顾.

1.2.1　异质多尺度方法

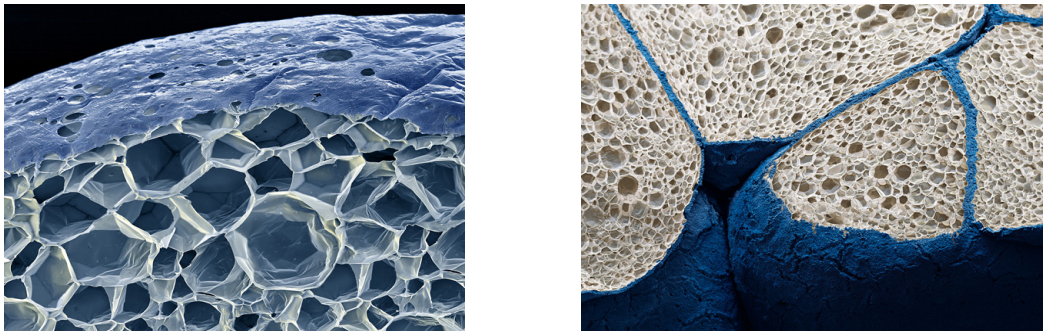
在材料科学、油藏开发、环境保护、地球勘探和农业科学等领域中，许多重要问题的本质都表现为多尺度现象，它们涉及从微观、介观到宏观等不同尺度的耦合与关联. 比如在复合材料领域，经典的连续介质力学将研究对象看作是连续均匀的材料. 然而，均匀材料只是工程结构在宏观尺度上的理想模型，而在微观尺度上往往表现出复杂的结构性和特有的非匀质性，并且这些微观性质会影响材料的宏观性能. 图1.1 [1] 展示了一些复合材料的微观结构，其中图1.1(a)为石墨聚苯板，图1.1(b)为泡沫金属.

这些多尺度现象对应的数学模型是具有快速振荡系数的偏微分方程. 考虑经典的椭圆方程

$$\begin{cases} -\operatorname{div}\left(a^{\varepsilon}(\boldsymbol{x})\nabla u^{\varepsilon}(\boldsymbol{x})\right) = f(\boldsymbol{x}), & \boldsymbol{x} \in \Omega \subset \mathbb{R}^d, \\ u^{\varepsilon}(\boldsymbol{x}) = 0, & \boldsymbol{x} \in \partial\Omega, \end{cases} \tag{1.8}$$

方程中的 Ω 为一凸的多边形区域，边界为 $\partial\Omega$. ∇ 和 div 表示函数的梯度和散度，具体定义见 §2.3节. ε 是一个小量，用来表征系数 a^{ε} 的多尺度性质. 一般把 ϵ 所

[1] 引自网站 http://www.flickr.com/photos/basf/sets/72157624601397168



(a) 石墨聚苯板　　(b) 泡沫金属

图 *1.1:* 复合材料的微观结构

表征的尺度称为微观尺度，在这个尺度上剖分的网格为微观网格. 相对应的，Ω 的尺度被称为宏观尺度，这一尺度下剖分的网格被称为宏观网格.

直接数值求解这一问题是很困难的，原因在于该方程是在 ε 的尺度上快速振荡，因此网格需要剖分到 ε 的尺度才能进行较为精确的求解，从而需要大量的计算机存储量和计算时间. 传统的多尺度方法应运而生，包括多重网格方法 [8](multigrid method)，快速多级方法 [61](fast multipole method)，区域分解方法 [107](domain decomposition method) 等. 这些方法在一定程度上降低了计算困难，但是它们本质上仍是在微观尺度上求解方程.

然而在实际应用中，相比微观尺度上的细微变化，多尺度问题往往更加关注由微观结构所带来的宏观尺度上的性质. 因此，人们研究了各种均匀化方法，试图在宏观尺度上用某种等效解来代替原来的多尺度问题的解. 而这些宏观解可以通过在较粗的宏观网格上对等效方程求解得到，其往往代表了微观解在某种意义下的宏观平均.

在经典的均匀化理论的基础上，国内外学者陆续发展出了一系列的现代多尺度方法. Dorobantu 和 Engquist[45] 研究了小波均匀化技术 (wavelet homogenization techniques). 该方法基于小波的多分辨分析，在微观尺度上得到原问题的半离散方程，然后利用小波映射得到宏观尺度空间的数值均匀化方程，并在宏观尺度上求解该方程从而有效减少计算量. 侯一钊等 [67, 66, 52] 提出了多尺度有限元方法 (multiscale finite element method). 该方法改进了函数空间，将微观尺度的信息引入到基函数中，构造振荡基函数，从而使数值解在宏观尺度上捕捉到微观尺度的效应. Hughes 等 [25, 69, 55, 99, 53] 提出的变分多尺度方法，改进了变分方法. 它主要是用来求解流体力学中的偏微分方程系统. 不同的多尺度方法有着各自的优势，对各种方法更加详细的介绍和比较可参考 [87].

鄂维南和 Engquist[50] 提出了一种关于物理多尺度方法的一般性的算法框

架，即异质多尺度方法 (heterogeneous multiscale method，简称 HMM). 该方法是一种利用问题自身的周期性、自相似性、尺度分离性和其他特殊的性质来设计多尺度算法的一般方法论. HMM 方法通常由两部分构成: 选取宏观求解器和求解局部微观问题来估计缺失的宏观数据，该方法的有效性体现在算法本身的灵活性以及它能用最少的计算开销从微观模型中提取缺失的宏观数据的能力. 文 [4] 是一篇关于异质多尺度方法的综述文献，它总结了异质多尺度方法在常微分方程、动力系统，以及偏微分方程中的有限元方法、有限差分方法等领域的应用.

本文主要考虑应用 HMM 方法求解带振荡系数的椭圆方程(1.8)，并采用有限元方法作为宏观求解器. 该方法又被称为异质多尺度有限元方法 (HMM-FEM). 鄂维南、明平兵和张平文 [51] 把 HMM-FEM 应用到各类椭圆均匀化问题中，包括线性的与非线性的、带确定系数的与带随机系数的. 他们给出了用 HMM-FEM 求解局部周期系数的椭圆均匀化问题的最优半离散误差估计，并对从 HMM-FEM 的解重构局部区域上小尺度信息的策略进行了分析. 而椭圆均匀化问题 HMM 方法的全离散问题最早由 Abdulle[1] 提出，作者采用有限元方法作为微观求解器对单胞问题进行求解，但该文中要求单胞大小等于周期的整数倍且单胞问题满足周期边界条件. 随后，Abdulle 和 Engquist 在 [5] 中采用伪谱方法作为微观求解器，当单胞问题满足周期边界条件时该算法具有准线性的计算复杂度.

从数值计算的角度来讲，在 HMM-FEM 方法中，宏观求解器的缺失数据是等效方程中有效系数位于数值积分点上的值. 在传统的方法中 [51, 87]，是在每个数值积分点周围选取一个单胞，并通过在单胞上求解局部微观问题来给出宏观有效系数的估计. 于是，整个算法的计算复杂度与单胞问题的个数及计算效率呈正相关. 杜锐和明平兵 [48] 采用了两种特殊的数值积分格式，其中的数值积分点位于单元的边界中心或顶点，使得每个单胞问题的计算结果可以被多个单元利用，于是变相地减少了单胞问题的数量，从而达到了提高计算效率的目的，并使整个宏观求解器实现了二阶收敛. 该方法被汪康 [109] 用以计算三维问题，验证了该方法在三维情形下的有效性. Abdulle 和 Bai 在 [3] 中提出了约化基 (reduced basis) 的异质多尺度有限元方法，他们选取少量的采样区域并在其上精确求解单胞问题，然后由这些解来构造约化基空间，并在该空间中给出宏观网格上有效系数矩阵的近似，从而提高了高阶宏观求解器的效率.

1.2.2　间断 Galerkin 方法

1973 年，Reed 和 Hill[95] 首次提出了求解双曲型方程的间断 Galerkin 方法 (discontinuous Galerkin，简称 DG). 随后出现了很多关于该方法在一阶双曲问题方面的研究成果 [76, 70]. 在此基础上，Cockburn 等人发展了 Runge-Kutta DG 方法 (RKDG) [38, 39, 37, 35, 41] 来求解非线性双曲问题，该方法将有限元

方法和有限体积法中的数值通量、近似 Riemann 解子、限制器等优点结合起来，具有稳定性强、精度高等优点. 关于 RKDG 方法更详细的讨论可参考 [42]. 1992 年，Richter[97] 改进了原始的 DG 方法，并将其应用到线性对流扩散方程中. 他证明了如果对流项占优，那么采用 k 次多项式为基函数就可得到最佳收敛阶为 $k+1/2$. 同时，由于 DG 方法在求解对流扩散方程时不需要设计复杂的策略就能保持稳定性 [56, 72]，这使得 DG 方法在该问题上受到广泛的关注. 相关的文献综述可以参考 [36].

DG 方法在椭圆问题上的早期应用集中体现在内罚方法 (interior penalty，简称 IP) 的发展上. 在 19 世纪六七十年代，Lions[80] 和 Nitsche[90] 分别提出可将一维椭圆方程中的 Dirichlet 边界条件以惩罚项的方式加入到变分形式中，而不用体现在有限元空间中. 基于此，人们开始考虑用间断分片多项式空间来求解椭圆问题，并引入惩罚项来控制数值解在单元边界上的跳跃. Babuška 和 Zlámal[12] 在用非协调元求解四阶问题时，将试探函数空间取为连续但分片可微的分片三次多项式空间，并通过对边界上的导数跳跃施加惩罚项来得到数值解. Wheeler[110] 将 Nitsche 的方法自然地推广到二维椭圆方程中. Arnold[9] 细致地分析了该方法在线性和非线性的椭圆问题、抛物问题中的应用. Douglas 和 Dupont[46] 在二阶椭圆问题和抛物问题中对连续基函数的导数的跳跃进行惩罚，并得到了介于 C^0 和 C^1 之间的连续性.

在二十世纪末，对于椭圆问题不少学者独立于 IP 的框架提出了新的 DG 方法. 1997 年，Bassi 和 Rebay[13] 由 RKDG 的思想出发，提出了一种新的方法来求解可压的 Navier-Stokes 方程. 区别于以往方法的是，他们同时对速度和通量采用了间断 Galerkin 逼近. 随后，在 [27, 28] 中 Bassi 等人又讨论了上述方法的一些变形. 在 1998 年，Cockburn 和舒其望 [40] 由 Bassi 和 Rebay 的方法出发，提出了局部间断 Galerkin 方法 (local discontinuous Galerkin，简称 LDG) 来求解非线性对流扩散问题. 他们通过引入辅助变量，将带有二阶导数项的原方程转化为一个一阶方程组. 对这个方程组用标准的 DG 方法进行离散，并恰当地选取数值通量来消去辅助变量，最终得到一个只依赖于原始未知变量的稳定的离散格式. Cockburn 和 Dawson[34] 将 LDG 方法推广到一般性的带有二阶项的问题中. Castillo[31] 等在一维空间对 hp-自适应的 LDG 方法进行了分析. 同一时期，还有 Baumann 和 Oden[14, 15] 对扩散问题提出了另一种 DG 方法，该方法中的双线性泛函甚至在惩罚系数为 0 时仍具有强制性，但却不具有对称性.

对于上述两大类的 DG 方法，Arnold 等人 [11] 提出了一个统一的分析框架. 他们通过引入辅助变量得到了椭圆方程在 DG 框架下的变分形式，并通过选取不同的数值通量得到了上述所有的 DG 格式. 他们还给出了所涉及的双线性算子的有界性、强制性以及收敛性的理论分析. 文 [30] 中则比较了不同的 DG 方法在实

际应用中的数值表现.

1.3　本文的内容安排

本文共分六章，除本章外，其余五章的结构如下所述. 第二章将介绍本文涉及的一些基础知识，包括线性最小二乘问题、Sobolev 空间的基础知识、一些符号的定义和周期函数的基本性质. 第三章首先定义了本文中所用到的网格和单元模板，然后给出了网格上的重构算子的定义和性质，并重点讨论了重构算子的稳定性，最后给出了多项式重构的数值算例. 第四章提出了一种基于最小二乘重构的高阶异质多尺度方法，给出了该方法的具体描述和收敛性分析，并通过数值算例对该方法的数值表现进行了测试. 第五章首先介绍了 DG 方法在求解椭圆问题时的变分形式，然后给出了一种基于重构基函数的 DG 方法，并对该方法进行了理论分析和数值实验. 第六章简要的总结了本文的内容并对进一步的研究进行了展望.

第二章　准备知识

本章将介绍本文所涉及的基础知识，主要包括线性最小二乘问题及其求解方法，和 Sobolev 空间及一些符号的定义.

2.1　线性最小二乘问题

本文中所涉及的最小二乘问题仅限于线性最小二乘问题. 本节将从离散范数的基本定义出发，介绍线性最小二乘问题的概念，并给出两种解法：法方程组方法和 QR 方法. 最后简要介绍带等式约束的最小二乘问题.

2.1.1　离散范数的基本定义

假设 N 为给定整数，并且有一个点集

$$\mathcal{I} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}, \quad \boldsymbol{x}_i \in \mathbb{R}^d.$$

那么对于定义在点集 $\mathcal{I}$ 上的离散函数 $f(\boldsymbol{x})$ 和 $g(\boldsymbol{x})$，作出如下定义

1. 离散内积：

$$(f, g)_{\mathcal{I}} := \sum_{i=1}^{N} f(\boldsymbol{x}_i) g(\boldsymbol{x}_i), \quad \boldsymbol{x}_i \in \mathcal{I}. \tag{2.1}$$

2. 离散范数 [1]：

$$\|f\|_{\mathcal{I}} := (f, f)_{\mathcal{I}}^{1/2} = \left(\sum_{i=1}^{N} f^2(\boldsymbol{x}_i) \right)^{\frac{1}{2}}, \quad \boldsymbol{x}_i \in \mathcal{I}. \tag{2.2}$$

[1] 范数定义中的正定性要求在这里为

$$\|f\|_{\mathcal{I}} \geqslant 0 \text{ 且 } \|f\|_{\mathcal{I}} = 0 \Rightarrow f(\boldsymbol{x}_i) = 0, \ \forall \boldsymbol{x}_i \in \mathcal{I}.$$

3. 正交：如果

$$(f,g)_{\mathcal{I}}=0,$$

则称 $f(\boldsymbol{x})$ 和 $g(\boldsymbol{x})$ 关于点集 $\mathcal{I}$ 正交.

4. 函数组的线性相关与线性无关：设有连续函数组 $\{\varphi_j(\boldsymbol{x})\}_{j=0}^n$，如果

$$\begin{aligned}&\sum_{j=1}^n c_j\varphi_j(\boldsymbol{x}_i)=0,\quad \forall \boldsymbol{x}_i\in\mathcal{I},\\ \Rightarrow &c_1=c_2=\cdots=c_n=0,\end{aligned}$$

那么称函数组 $\{\varphi_j(\boldsymbol{x})\}_{j=0}^n$ 在点集 $\mathcal{I}$ 上线性无关. 否则，称 $\{\varphi_j(\boldsymbol{x})\}_{j=0}^n$ 在点集 $\mathcal{I}$ 上线性相关.

2.1.2　最小二乘问题与法方程组方法

给定一个点集 $\mathcal{I}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$ 和一个向量 $\boldsymbol{y}=(y_1,y_2\ldots,y_N)^{\mathrm{T}}\in\mathbb{R}^N$，那么线性最小二乘问题是指：

对于给定的 $\mathcal{I}$ 和 $\boldsymbol{y}$，找出 $u(\boldsymbol{x})\in\mathcal{V}$，使得

$$u(\boldsymbol{x})=\arg\min_{v(\boldsymbol{x})\in\mathcal{V}}\sum_{\boldsymbol{x}_i\in\mathcal{I}}|v(\boldsymbol{x}_i)-y_i|^2, \tag{2.3}$$

其中 $\mathcal{V}$ 是一个 n 维线性空间，并且具有基函数 $\{\varphi_i(\boldsymbol{x})\}_{i=1}^n$. 在下文中，总是假设基函数组 $\{\varphi_i(\boldsymbol{x})\}_{i=1}^n$ 在点集 $\mathcal{I}$ 上线性无关. 常用的函数系 $\{\varphi_i(\boldsymbol{x})\}_{i=1}^n$ 有幂函数、三角函数和指数函数等. 如果将函数系取为 $\{\varphi^0(\boldsymbol{x}),\{\varphi^1(\boldsymbol{x})\},\cdots,\{\varphi^m(\boldsymbol{x})\}\}$，其中

$$\varphi^0(\boldsymbol{x})\equiv 1,\quad \{\varphi^k(\boldsymbol{x})\}=\left\{x_1^{i_1}x_2^{i_2}\cdots x_d^{i_d}\middle|\sum_{s=1}^d i_s=k\right\},\ k=1,2,\cdots,m,$$

那么此时函数空间 $\mathcal{V}$ 为多项式空间 $\mathbb{P}_m$，其中的函数为次数不超过 m 的多项式.

对于 $\mathcal{V}$ 中的函数 $u(\boldsymbol{x})$，存在向量 $\boldsymbol{u}=(u_1,u_2,\cdots,u_n)^{\mathrm{T}}\in\mathbb{R}^n$，使得

$$u(\boldsymbol{x})=\sum_{i=1}^n u_i\varphi_i(\boldsymbol{x}).$$

因此求解 $u(\boldsymbol{x})$ 等价于求解 $\boldsymbol{u}$. 由上一章的分析可知，线性最小二乘问题中的矛盾方程组(1.3)将简化为线性系统(1.6). 于是在最小二乘意义下求解该矛盾方程，就意味着求解如下问题:

对于给定的 $\mathcal{I}$ 和 $\boldsymbol{y}$，找出 $\boldsymbol{u} \in \mathbb{R}^n$，使得

$$\boldsymbol{u} = \arg\min_{\boldsymbol{v}\in\mathbb{R}^n} \|\Psi\boldsymbol{v} - \boldsymbol{y}\|_2^2. \tag{2.4}$$

在方程(1.6)的两边同时乘以 Ψ^{T}，就得到了法方程组(1.7). 对于最小二乘解与法方程组的解，有如下定理：

定理 2.1. [117, 定理 3.1.4] $\boldsymbol{u}$ 为最小二乘问题(2.4)的解的充分必要条件是 $\boldsymbol{u}$ 为法方程组 (1.7)的解.

证明. “⇒” 定义

$$Q(\boldsymbol{v}) := \sum_{i=1}^{N}\left[y_i - \sum_{k=1}^{n} v_k\varphi_k(\boldsymbol{x}_i)\right]^2.$$

设 $\boldsymbol{u}$ 是最小二乘问题(2.4)的解，那么 $\boldsymbol{u}$ 也是 $Q(\boldsymbol{v})$ 的极小值点，于是有

$$\frac{\partial Q}{\partial v_k}(\boldsymbol{u}) = \sum_{i=1}^{N}\left[y_i - \sum_{j=1}^{n} u_j\varphi_j(\boldsymbol{x}_i)\right](-\varphi_k(\boldsymbol{x}_i)) = 0, \quad k = 1, 2, \cdots, n.$$

从而得到方程组

$$\sum_{j=1}^{n}\left(\sum_{i=1}^{N}\varphi_k(\boldsymbol{x}_i)\varphi_j(\boldsymbol{x}_i)\right)u_j = \sum_{i=1}^{N}\varphi_k(\boldsymbol{x}_i)y_i, \quad k = 1, 2, \cdots, n. \tag{2.5}$$

而事实上方程组(2.5)写成矩阵形式就得到式(1.7)，于是 $\boldsymbol{u}$ 也是法方程组(1.7)的解.

“⇐” 设 $\boldsymbol{u}$ 是法方程组(1.7)的解，则对于 $\forall \boldsymbol{v} \in \mathbb{R}^n$ 有

$$\|\Psi(\boldsymbol{u}+\boldsymbol{v}) - \boldsymbol{y}\|_2^2 = \|\Psi\boldsymbol{u} - \boldsymbol{y}\|_2^2 + \|\Psi\boldsymbol{v}\|_2^2 + 2(\Psi\boldsymbol{u} - \boldsymbol{y})^{\mathrm{T}}\Psi\boldsymbol{v}.$$

由 $\Psi^{\mathrm{T}}(\Psi\boldsymbol{u} - \boldsymbol{y}) = 0$ 可以得到 $(\Psi\boldsymbol{u} - \boldsymbol{y})^{\mathrm{T}}\Psi\boldsymbol{v} = 0$，于是

$$\|\Psi(\boldsymbol{u}+\boldsymbol{v}) - \boldsymbol{y}\|_2^2 = \|\Psi\boldsymbol{u} - \boldsymbol{y}\|_2^2 + \|\Psi\boldsymbol{v}\|_2^2 \geqslant \|\Psi\boldsymbol{u} - \boldsymbol{y}\|_2^2.$$

可知 $\boldsymbol{u}$ 也是最小二乘问题(2.4)的解. □

将法方程组(1.7)中的系数矩阵定义为 $\Phi := \Psi^{\mathrm{T}}\Psi$，那么 Φ 的具体形式为：

$$\Phi = \begin{pmatrix} (\varphi_1,\varphi_1)_{\mathcal{I}} & (\varphi_1,\varphi_2)_{\mathcal{I}} & \cdots & (\varphi_1,\varphi_n)_{\mathcal{I}} \\ (\varphi_2,\varphi_1)_{\mathcal{I}} & (\varphi_2,\varphi_2)_{\mathcal{I}} & \cdots & (\varphi_2,\varphi_n)_{\mathcal{I}} \\ & \cdots & \cdots & \\ (\varphi_n,\varphi_1)_{\mathcal{I}} & (\varphi_n,\varphi_2)_{\mathcal{I}} & \cdots & (\varphi_n,\varphi_n)_{\mathcal{I}} \end{pmatrix}.$$

对于法方程组(1.7)的解的存在唯一性有如下定理.

定理 2.2. *如果 $\mathcal{V}$ 的基函数 $\{\varphi_i(\boldsymbol{x})\}_{i=1}^n$ 在点集 $\mathcal{I}$ 上线性无关，那么方程组(1.7)有且仅有唯一解.*

证明. 为了证明方程组 (1.7) 存在唯一解，那么只需证明其对应的齐次方程组

$$\Phi\boldsymbol{u} = \boldsymbol{0} \tag{2.6}$$

只有零解. 为此，采用反证法，假如(2.6)存在一个非零解：

$$\hat{\boldsymbol{u}} = (\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_n)^{\mathrm{T}},$$

记

$$\hat{u}(\boldsymbol{x}) := \sum_{i=1}^{n} \hat{u}_i \varphi_i(\boldsymbol{x}),$$

那么

$$\begin{aligned} \|\hat{u}\|_{\mathcal{I}}^2 &= \sum_{i=1}^{N} \left(\sum_{j=1}^{n} \hat{u}_j \varphi_j(\boldsymbol{x}_i) \right) \left(\sum_{k=1}^{n} \hat{u}_k \varphi_k(\boldsymbol{x}_i) \right) \\ &= \sum_{j=1}^{n} \sum_{k=1}^{n} \left(\sum_{i=1}^{N} \varphi_j(\boldsymbol{x}_i)\varphi_k(\boldsymbol{x}_i) \right) \hat{u}_j \hat{u}_k \\ &= \sum_{j=1}^{n} \sum_{k=1}^{n} (\varphi_j, \varphi_k)_{\mathcal{I}} \hat{u}_j \hat{u}_k \\ &= \hat{\boldsymbol{u}}^{\mathrm{T}} \Phi \hat{\boldsymbol{u}} = 0. \end{aligned}$$

从而

$$\hat{u}(\boldsymbol{x}_i) = 0, \quad \forall \boldsymbol{x}_i \in \mathcal{I}.$$

即

$$\hat{u}_1\varphi_1(\boldsymbol{x}_i) + \hat{u}_2\varphi_2(\boldsymbol{x}_i) + \cdots + \hat{u}_n\varphi_n(\boldsymbol{x}_i) = 0, \quad \forall \boldsymbol{x}_i \in \mathcal{I}.$$

由于 $\hat{\boldsymbol{u}}$ 非零，因此上式与函数系 $\{\varphi_i(\boldsymbol{x})\}_{i=1}^n$ 在点集 $\mathcal{I}$ 上线性无关相矛盾. 因此法方程组(1.7)存在唯一解. □

推论 2.3. 如果矩阵 Ψ 的各列线性无关，那么 Φ 非奇异，此时 $\boldsymbol{u} = (\Psi^{\mathrm{T}}\Psi)^{-1}\Psi^{\mathrm{T}}\boldsymbol{y}$ 为方程组(1.7)的唯一解；如果 Φ 奇异，那么一定有 $\mathrm{rank}(\Psi) < n$.

于是法方程组的解(1.7)可以表示为

$$\boldsymbol{u} = (\Psi^{\mathrm{T}}\Psi)^{-1}\Psi^{\mathrm{T}}\boldsymbol{y}.$$

若定义

$$\Psi^{\dagger} := (\Psi^{\mathrm{T}}\Psi)^{-1}\Psi^{\mathrm{T}}.$$

则最小二乘解 $\boldsymbol{u}$ 可以表示为

$$\boldsymbol{u} = \Psi^{\dagger}\boldsymbol{y}. \tag{2.7}$$

$\Psi^{\dagger}$ 也被称为 Ψ 的 Moors-Penrose 广义逆. 定义 $\mathrm{cond}(A) = \|A\|_2\|A^{\dagger}\|_2$ 为矩阵 A 的条件数，它的大小表征着最小二乘问题(2.4) 的敏感性. 若 $\mathrm{cond}(A)$ 很大，则称最小二乘问题是病态的. 否则，称最小二乘问题是良态的.

对于法方程组(1.7)，在问题规模不是很大的时候，可以用列主元 Gauss 消去法求解. 同时，注意到如果矩阵 Ψ 列满秩，那么 Φ 为对称正定的矩阵，因此也可以考虑用 Cholesky 分解来求解法方程组.

2.1.3 最小二乘问题的正交化解法

利用法方程组求解线性最小二乘问题是一种通用的古老方法. 但对于法方程组的系数 Φ，其条件数满足 (详见 [117, 定理 3.1.6]) $\mathrm{cond}(\Phi) = \mathrm{cond}(\Psi)^2$. 这意味着法方程组系数矩阵 Φ 会将矛盾方程系数矩阵 Ψ 中的病态以平方的倍数放大，从而导致法方程组求解的不稳定.

为此，可以考虑用正交化的方法对矛盾方程组(1.6)直接进行求解. 应用 Householder 变换 ([59, 60, 117]) 把系数矩阵 Ψ 正交三角化，使得

$$Q\Psi = \begin{pmatrix} R \\ O \end{pmatrix},$$

其中 R 为 n 阶上三角矩阵，O 为 $(N-n)\times n$ 阶的零矩阵，Q 是一个 m 阶正交矩阵. 把 N 维向量 $Q\boldsymbol{y}$ 相应地分块成 n 维向量 $\boldsymbol{z}$ 和 $N-n$ 维向量 $\boldsymbol{w}$，即

$$Q\boldsymbol{y} = \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{w} \end{pmatrix}.$$

于是

$$Q\boldsymbol{r} = Qy - Q\Psi\boldsymbol{u} = \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{w} \end{pmatrix} - \begin{pmatrix} R \\ O \end{pmatrix} \boldsymbol{u}$$
$$= \begin{pmatrix} \boldsymbol{z} - R\boldsymbol{u} \\ \boldsymbol{w} \end{pmatrix}.$$

由于 Q 是正交矩阵，所以

$$\|\boldsymbol{r}\|_2^2 = \|Q\boldsymbol{r}\|_2^2 = \|\boldsymbol{z} - R\boldsymbol{u}\|_2^2 + \|\boldsymbol{w}\|_2^2.$$

如果选取 $\boldsymbol{u}$ 使得

$$R\boldsymbol{u} = \boldsymbol{z}, \tag{2.8}$$

那么 $\|\boldsymbol{r}\|_2$ 可以取到最小值 $\|\boldsymbol{w}\|_2$. 因此方程(2.8)的解就是最小二乘问题(2.4)的解. 由于 R 是上三角矩阵，因此该方程非常容易求解.

Householder 方法并不是实现 QR 分解的唯一方法，常用的方法还有 Givens 变换 ([112, 58] 等) 或 Gram-Schmidt 正交化方法 ([101] 等).

QR 方法是求解线性最小二乘问题的一个非常稳定的方法. 相比于法方程组方法，它避免了 $\Psi^{\mathrm{T}}\Psi$ 所带来的条件数放大而导致的求解不稳定. 不过由于 QR 方法中的约化大约需要 $\frac{2}{3}n^3$ 次乘法，而如果用 Gauss 消去法解法方程组(1.7)，大约只需要 $\frac{1}{3}n^3$ 次乘法. 在本文的计算中，三角分解法和正交化方法都会被用到. 一般来说，如果 Ψ 不是病态的，会优先考虑用法方程组求解最小二乘问题；如果 Ψ 有可能病态，并且数值方法对最小二乘问题解的精度要求较高，那么采用 QR 方法则会比较稳妥.

2.1.4　带等式约束的最小二乘问题

在有些问题中，$u_1, u_2, \cdots, u_n$ 之间并不是完全独立的，而是满足一定的约束条件. 因此，本节将考虑带有等式约束的线性最小二乘问题.

找出 $\boldsymbol{u} \in \mathbb{R}^n$，使得它是如下问题的解

$$\min_{\boldsymbol{v}\in\mathbb{R}^n} \|\Psi\boldsymbol{v} - \boldsymbol{y}\|_2^2, \tag{2.9}$$

并且服从 n_c 个约束条件

$$C\boldsymbol{v} = \boldsymbol{d}. \tag{2.10}$$

其中 C 是 $n_c \times n$ 的矩阵，$\boldsymbol{d}$ 是长度为 n_c 的向量，并且总是假设 $n > n_c = \mathrm{rank}(C)$.

将带等式约束的线性最小二乘问题简称为 LSE 问题,其含义是在满足方程 (2.10)的所有解中，$\|\Psi\boldsymbol{u}-\boldsymbol{y}\|_2$ 的值为最小. 显然，LSE 问题有解的充分必要条件是约束方程组(2.10)是相容的. 在下文中总是假定这个条件成立.

求解 LSE 问题的方法很多，本文将采用 Lagrange 乘子法. 现将对此方法做一简要的介绍. 引入 Lagrange 乘子 $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_{n_c})^{\mathrm{T}}$，得到 Lagrange 目标函数

$$L(\boldsymbol{v},\boldsymbol{\lambda})=\|\Psi\boldsymbol{v}-\boldsymbol{y}\|_2^2+2\boldsymbol{\lambda}^{\mathrm{T}}(C\boldsymbol{v}-\boldsymbol{d}).$$

用 $L(\boldsymbol{v},\boldsymbol{\lambda})$ 分别对 $\boldsymbol{v}$ 和 $\boldsymbol{\lambda}$ 求偏导并令其为 0，整理后得到

$$\begin{pmatrix} \Psi^{\mathrm{T}}\Psi & C^{\mathrm{T}} \\ C & 0 \end{pmatrix}\begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{\lambda} \end{pmatrix}=\begin{pmatrix} \Psi^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{d} \end{pmatrix}. \tag{2.11}$$

上述方程也被称为 LSE 问题的法方程组. 当 $\mathrm{rank}(\Psi)=n$ 且 $\mathrm{rank}(C)=n_c$ 时，方程组(2.11)存在唯一解. 设该解为 $(\boldsymbol{u},\boldsymbol{\lambda}^*)^{\mathrm{T}}$，那么 $\boldsymbol{u}$ 就是原 LSE 问题的解.

2.2 Sobolev 空间

下面以区域 Ω 为例，给出主要 Sobolev 空间的定义 [7].

$C^m(\Omega)$ 表示区域 Ω 上的 m 次连续可微函数组成的集合，其中 m 为非负整数；$C^\infty(\Omega)$ 表示区域 Ω 上的无穷次连续可微函数的集合；$C_0^\infty(\Omega)$ 表示区域 Ω 上具有紧支集的无穷次连续可微函数组成的集合.

令 m 为非负整数，$\alpha\in(0,1]$，则 $C^{m,\alpha}(\Omega)$ 表示 $C^m(\Omega)$ 中那些 m 阶导数为 α 次 Hölder 连续的函数的全体. 特别地，设 $u\in C^{0,\alpha}(\Omega)$，则存在常数 $M>0$，使得

$$|u(\boldsymbol{x})-u(\boldsymbol{y})|\leqslant M|\boldsymbol{x}-\boldsymbol{y}|^\alpha\quad\forall\boldsymbol{x},\boldsymbol{y}\in\Omega,$$

其中 $|\cdot|$ 表示标量的绝对值或 $\mathbb{R}^d$ 中向量的欧式范数. 当 $\alpha=1$ 时，称 u 是 Lipschitz 连续的.

设 f 是 Ω 上的实值 Lebesque 可测函数，记

$$\int_\Omega f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

为 Lebesque 积分. 现引入记号

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(\boldsymbol{x})|^p \,\mathrm{d}\boldsymbol{x}\right)^{1/p}, \quad 1 \leqslant p < \infty,$$
$$\|f\|_{L^\infty(\Omega)} = \operatorname{ess\,sup}_{\boldsymbol{x}\in\Omega} |f(\boldsymbol{x})|, \quad p = \infty.$$

定义空间

$$L^p(\Omega) = \left\{ f \;\middle|\; \|f\|_{L^p(\Omega)} < \infty \right\}, \quad 1 \leqslant p \leqslant \infty.$$

当 $p = 2$ 时，$L^2(\Omega)$ 是 Hilbert 空间，其范数为 $\|f\|_{L^2(\Omega)}$(简记为 $\|f\|_0$)，对应的内积定义为 $(f, g)_{L^2(\Omega)} = \int_\Omega f(\boldsymbol{x})g(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$.

Sobolev 空间 $W^{m,p}(\Omega)$(m 为非负整数，$1 \leqslant p \leqslant \infty$) 定义为

$$W^{m,p}(\Omega) = \{u \;|\; D^\alpha u \in L^p(\Omega),\ |\alpha| \leqslant m\},$$

其中 $D^\alpha u$ 为 u 的广义导数，其范数为

$$\|u\|_{W^{m,p}(\Omega)} = \left(\sum_{|\alpha|\leqslant m} \int_\Omega |D^\alpha u|^p \,\mathrm{d}\boldsymbol{x}\right)^{1/p}, \quad 1 \leqslant p < \infty,$$
$$\|u\|_{W^{m,\infty}(\Omega)} = \max_{|\alpha|\leqslant m} \|D^\alpha u\|_{L^\infty(\Omega)}, \quad p = \infty,$$

相应的半范数为

$$|u|_{W^{m,p}(\Omega)} = \left(\sum_{|\alpha|=m} \int_\Omega |D^\alpha u|^p \,\mathrm{d}\boldsymbol{x}\right)^{1/p}, \quad 1 \leqslant p < \infty,$$
$$|u|_{W^{m,\infty}(\Omega)} = \max_{|\alpha|=m} \|D^\alpha u\|_{L^\infty(\Omega)}, \quad p = \infty,$$

当 $p = 2$ 时，记 $H^m(\Omega) = W^{m,2}(\Omega)$，它们都是 Hilbert 空间，其范数为 $\|u\|_{H^m(\Omega)}$ (简记为 $\|u\|_m$)，对应的内积定义为 $(u, v)_m = \sum\limits_{|\alpha|<m} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}$. 注意到当 $m = 0$ 时，$W^{0,p}(\Omega) = L^p(\Omega)$.

令 $W_0^{m,p}(\Omega)$ 是 $C_0^\infty(\Omega)$ 在范数 $\|\cdot\|_{W^{m,p}(\Omega)}$ 意义下的完备化空间. 当 $p = 2$ 时，记 $H_0^m(\Omega) = W_0^{m,2}(\Omega)$. 引入 $H_0^1(\Omega)$ 的对偶空间

$$H^{-1}(\Omega) = \left(H_0^1(\Omega)\right)',$$

其范数为

$$\|F\|_{H^{-1}(\Omega)} = \sup_{H_0^1(\Omega)\setminus\{0\}} \frac{|\langle F, u\rangle|}{\|u\|_{H_0^1(\Omega)}}.$$

其中 $\langle\cdot,\cdot\rangle$ 是 $H^{-1}(\Omega)$ 和 $H_0^1(\Omega)$ 之间的对偶对.

2.3 特殊记号

区域 Ω 的闭包记为 $\overline{\Omega}$，内点集记为 $\underline{\Omega}$.

函数 $u(\boldsymbol{x})$ 的梯度记作

$$\nabla u = \operatorname{grad} u := \left(\frac{\partial u}{\partial x_1}, \cdots, \frac{\partial u}{\partial x_d}\right)^{\mathrm{T}}.$$

如果函数 $u = u(\boldsymbol{x}, \boldsymbol{y})$，那么 $\nabla_{\boldsymbol{y}} u$ 表示函数 u 关于分量 $\boldsymbol{y}$ 的梯度.

向量函数 $v(\boldsymbol{x}) = (v_1(\boldsymbol{x}), v_2(\boldsymbol{x}), \cdots, v_d(\boldsymbol{x}))^{\mathrm{T}}$ 的散度记作

$$\operatorname{div} v = \nabla \cdot v := \sum_{i=1}^{d} \frac{\partial v_i}{\partial x_i}.$$

$(\cdot,\cdot)$ 表示 $\mathbb{R}^d$ 上的内积，并且诱导出的范数为 $|\cdot|$.

设 $\beta > \alpha > 0$，定义函数空间

$$\mathcal{M}(\alpha,\beta,\Omega) := \Big\{\mathcal{B} \in [L^\infty(\Omega)]^{d^2} \mid (\mathcal{B}(\boldsymbol{x})\boldsymbol{\xi},\boldsymbol{\xi}) \geq \alpha|\boldsymbol{\xi}|^2, |\mathcal{B}(\boldsymbol{x})\boldsymbol{\xi}| \leq \beta|\boldsymbol{\xi}|, \\ \forall \boldsymbol{\xi} \in \mathbb{R}^d, \quad \forall \boldsymbol{x} \in \Omega \ a.e.\Big\}.$$

设 $u(\boldsymbol{x}) \in L^1(\Omega)$，则函数 u 在区域 Ω 上的积分平均值记为

$$\langle u\rangle_\Omega = \fint_\Omega u(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} := \frac{1}{|\Omega|}\int_\Omega u(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

其中 $|\Omega|$ 表示区域 Ω 的面积或体积. 同理, 可以将此记号推广到其他区域.

2.4 周期函数

定义 2.4. 记 $Y := (-1/2, 1/2)^d$. 设 $f(\boldsymbol{x})$ 是定义在 $\mathbb{R}^d$ 上的函数，若

$$f(\boldsymbol{x} + k\boldsymbol{e}_i) = f(\boldsymbol{x}), \quad \forall k \in Z,\ \forall i \in \{1, 2, \cdots, d\},\ \forall \boldsymbol{x} \in \mathbb{R}^d,$$

其中 $\boldsymbol{e}_i$ 是 $\mathbb{R}^d$ 的标准基，则称 $f(\boldsymbol{x})$ 是 Y-周期函数.

定理 2.5. [33, 注 2.10] 令 $1 \leqslant p < \infty$，设 f 是 $L^p(Y)$ 上的 Y-周期函数，$f_\varepsilon(\boldsymbol{x}) = f(\boldsymbol{x}/\varepsilon)$，则当 ε 足够小时，对于任意至少包含 Y 的一个平移集的开区间 I，存在只与 d 相关的常数 C，使得

$$\|f_\varepsilon\|^p_{L^p(I)} \leqslant C\frac{|I|}{|Y|}\|f\|^p_{L^p(Y)}.$$

令 $C^\infty_{\mathrm{per}}(Y)$ 表示 $C^\infty(\mathbb{R}^d)$ 中 Y-周期函数的集合，$H^1_{\mathrm{per}}(Y)$ 表示 $C^\infty_{\mathrm{per}}(Y)$ 在范数 $\|\cdot\|_{H^1(\Omega)}$ 意义下的闭包. 若 $u \in H^1_{\mathrm{per}}(Y)$，则 u 在 Y 的对应面上有相同的迹.

定义 2.6. 商空间 $\mathcal{W}_{\mathrm{per}}(Y) = H^1_{\mathrm{per}}(Y)/\mathbb{R}$ 用如下的等价关系来定义

$$u \simeq v \Leftrightarrow u - v \in \mathbb{R}, \quad \forall u, v \in H^1_{\mathrm{per}}(Y).$$

用 $\dot{u}$ 表示函数 u 的等价关系类.
$\mathcal{W}_{\mathrm{per}}(Y)$ 上的范数定义为

$$\|\dot{u}\|_{\mathcal{W}_{\mathrm{per}}(Y)} = \|\nabla u\|_{L^2(Y)}, \quad \forall u \in \dot{u},\ \dot{u} \in \mathcal{W}_{\mathrm{per}}(Y).$$

它的对偶空间为

$$\Big(\mathcal{W}_{\mathrm{per}}(Y)\Big)' := \left\{F \in \Big(H^1_{\mathrm{per}}(Y)\Big)' \ \middle|\ F(c) = 0,\ \forall c \in \mathbb{R}\right\}.$$

考虑周期边界条件的椭圆问题

$$\begin{cases} -\operatorname{div}(A\nabla u) = f, \quad \boldsymbol{x} \in Y, \\ \quad u\text{为}Y\text{-周期函数}. \end{cases} \tag{2.12}$$

定理 2.7. [33, 定理 4.26] 设矩阵 $A \in \mathcal{M}(\alpha, \beta, Y)$ 的每一个分量都是 Y-周期函数，$f \in \Big(\mathcal{W}_{\mathrm{per}}(Y)\Big)'$，则问题(2.12)有唯一解并且

$$\|\dot{u}\|_{\mathcal{W}_{\mathrm{per}}(Y)} \leqslant \frac{1}{\alpha}\|f\|_{\Big(\mathcal{W}_{\mathrm{per}}(Y)\Big)'}.$$

其中

$$\|f\|_{\Big(\mathcal{W}_{\mathrm{per}}(Y)\Big)'} := \sup_{\mathcal{W}_{\mathrm{per}}(Y)\setminus\{0\}} \frac{|\langle f, \dot{v}\rangle|}{\|\dot{v}\|_{\mathcal{W}_{\mathrm{per}}(Y)}}.$$

第三章　重构算子

在很多数值方法中，经常需要利用网格中局部的采样点上的信息来重构单元上的函数. 本章将重点讨论最小二乘重构算子及其性质. §3.1节将给出所使用的网格的定义；§3.2节将给出单元模板的定义；在 §3.3节中，对多项式重构算子进行了定义；§3.4节考虑重构算子的稳定性，着重分析了其插值不等式所涉及的常数；§3.5节给出了几个多项式重构的数值算例；§3.6节对本章作了小结.

3.1　网格的定义

首先给出本文中所使用的网格的定义.

定义 3.1. 将区域 Ω 分成有限个子集 K，称每个子集为一个单元，记这些子集的集合为 $\mathcal{T}_h$. 称 $\mathcal{T}_h$ 是区域 Ω 的一个剖分，如果下面的性质成立：

1. $\bigcup\limits_{K\in\mathcal{T}_h} K = \overline{\Omega}$.
2. 对任意两个不同的 K_i 和 K_j, $\underline{K_i} \bigcap \underline{K_j} = \emptyset$.
3. $\mathcal{T}_h$ 中的单元 K 是闭合的，且其内点集非空. 而其边界 ∂K 是 Lipschitz 连续的.

在本文中，主要涉及的单元形状为多边形 (多面体).

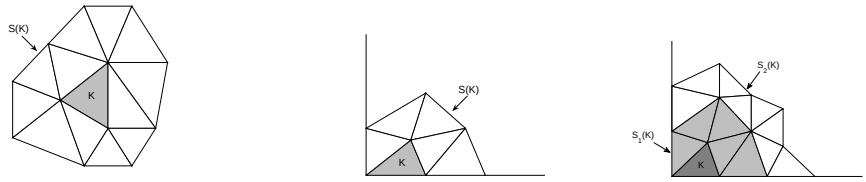
定义 3.2. 设剖分 $\mathcal{T}_h$ 是对 Ω 的一族剖分，对 $\forall K \in \mathcal{T}_h$，记 h_K 为单元 K 的直径，ρ_K 为包含在 K 内的最大内切球的直径，记 $h := \max\limits_{K\in\mathcal{T}_h} h_K$，若存在常数 C，使得

$$\frac{h_K}{\rho_K} \leqslant C, \quad \forall K \in \mathcal{T}_h,$$

则称剖分是正则的. 如果剖分族不仅是正则的，而且存在常数 ν，使

$$\frac{h}{\rho_K} \leqslant \nu, \tag{3.1}$$

则称剖分是拟一致的.



(a) K 在区域的内部　　(b) K 在区域的边界　　(c) $S_1(K)$ 与 $S_2(K)$

图 *3.1:* 单元模板 $S(K)$ 的例子.

需要注意的是，(3.1)式和标准的逆假设的定义 (见 [32, p. 140]) 并不完全相同，但两者对于形状规则的网格来说是等价的，(3.1)式的定义更便于下文的叙述.

3.2　单元模板的选取

接下来，对于每个单元 $K \in \mathcal{T}_h$，往往希望能通过该单元以及其周边单元的信息来重构 K 上函数. 为此，首先需要构造单元 K 的模板 $S(K)$，来确定提取信息的局部区域. 为了实际应用的需要，本文中假设 $S(K)$ 中的单元距离 K 足够近，使得 $S(K)$ 的直径保持在 $O(h_K)$.

本文将采用一种递归的方式来定义单元 K 的模板，其构造方式如下：

$$S_0(K) = K, \qquad S_t(K) = \bigcup_{\substack{\widetilde{K}\in\mathcal{T}_h,\, \widetilde{K}\cap K'\neq\emptyset,\\ K'\subset S_{t-1}(K)}} \widetilde{K}, \tag{3.2}$$

其中 $t \in \mathbb{N}$ 表示模板的层数. 为简便起见, 在不引起歧义的情况下，将省去下标 t.

由式(3.2)定义的单元模板会从单元本身开始向四周延伸. 图3.1展示了几个单元模板的例子, 其中图3.1(a)中是一个区域内部单元的模板示意图，该模板包含了所有与 $\overline{K}$ 相交非空的单元. 图3.1(b)中的单元则是位于区域的边界上，该单元的模板只能向区域内延伸. 图3.1(c)展示了单元位于边界，且模板扩展了两层的例子. 在后文中可以看出，这种方式定义的模板单元更便于下一节中定义的重构算子的稳定性的证明.

3.3 重构算子的定义

假设对于 $\mathcal{T}_h$ 中的每个单元 K，都有采样点集

$$I_K = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{l_K}\},$$

并且 $l_K := \#I_K$，其中 $\#$ 表示集合的基数. 同时，存在与 K 无关的常数 M 使得

$$\max_{K\in\mathcal{T}_h} l_K \leqslant M. \tag{3.3}$$

注. (3.3)式可以看做一个技术性的假设，它实质上是要求每个单元内的采样点的个数不会随着网格的加密而无限制的增长. 本文中所讨论的情形中(3.3)式总是成立.

对整个区域 Ω，其采样点集为

$$I_\Omega = \bigcup_{K\in\mathcal{T}_h} I_K,$$

并且 $l := \#I_\Omega$. 设向量 $\boldsymbol{v} \in \mathbb{R}^l$ 为任取的一组采样点上的采样值

$$\boldsymbol{v} = (v_1, v_2, \cdots, v_l),$$

并且 $v_i \in \boldsymbol{v}$ 与 $\boldsymbol{x}_i \in I_\Omega$ 为一一对应的关系.

首先考虑函数拟合问题. 对于任意单元 $K \in \mathcal{T}_h$，考虑利用其周围网格上采样点的信息来重构出 K 上的函数. 为此，构造单元模板 $S(K)$，并记 $S(K)$ 上的采样点集为

$$\mathcal{I}_K := I_{S(K)} = \bigcup_{K'\in S(K)} I_{K'},$$

并且 $n_K := \#\mathcal{I}_K$. 定义 $\mathcal{S}_K\boldsymbol{v}$ 为 $\boldsymbol{v}$ 对应于 $S(K)$ 的采样点的子向量，于是有 $\mathcal{S}_K\boldsymbol{v} \in \mathbb{R}^{n_K}$. 记 $\mathbb{P}_m(S(K))$ 为定义在 $S(K)$ 上的 m 次多项式空间. 如果能求得函数 $\tilde{q}(\boldsymbol{x}; \boldsymbol{v}_K) \in \mathbb{P}_m(S(K))$，使其满足如下的最小二乘问题

$$\tilde{q}(\boldsymbol{x}; \boldsymbol{v}) = \arg\min_{p(\boldsymbol{x})\in\mathbb{P}_m} \sum_{\boldsymbol{x}_i\in\mathcal{I}_K} \left|p(\boldsymbol{x}_i) - (\mathcal{S}_K\boldsymbol{v})_i\right|^2, \tag{3.4}$$

则称 $\tilde{q}(\boldsymbol{x}; \boldsymbol{v})$ 是向量 $\boldsymbol{v}$ 在 $S(K)$ 上的m 次重构多项式.

在 $S(K)$ 及其点集 $\mathcal{I}_K$ 上，定义如下的局部 m 阶重构算子 $\widetilde{\mathcal{R}}_m^K$：

$$\begin{aligned}\widetilde{\mathcal{R}}_m^K &: \mathbb{R}^{n_K} \mapsto \mathbb{P}^m(S(K)),\\ &\widetilde{\mathcal{R}}_m^K(\mathcal{S}_K\boldsymbol{v}) \to \tilde{q}(\boldsymbol{x};\boldsymbol{v}).\end{aligned} \tag{3.5}$$

可以看出，重构算子 $\widetilde{\mathcal{R}}_m^K$ 是把 $\mathbb{R}^{n_K}$ 中的一个向量映射成了 $\mathbb{P}_m(S(K))$ 中的一个函数. 由 $\widetilde{\mathcal{R}}_m^K$ 出发，可以定义全局的 m 阶重构算子 $\widetilde{\mathcal{R}}_m$，使得 $\widetilde{\mathcal{R}}_m\boldsymbol{v}$ 在 K 上的限制恰好就是 $\widetilde{\mathcal{R}}_m^K(\mathcal{S}_K\boldsymbol{v})\big|_K$，即

$$(\widetilde{\mathcal{R}}_m\boldsymbol{v})\big|_K := \left(\widetilde{\mathcal{R}}_m^K(\mathcal{S}_K\boldsymbol{v})\right)\big|_K. \tag{3.6}$$

于是，$\widetilde{\mathcal{R}}_m$ 将 I_Ω 上的向量 $\boldsymbol{v}$ 映射成了 Ω 上的分片 m 阶多项式. 由(2.7)可知，最小二乘问题(3.4)的解对 $\boldsymbol{v}$ 是线性依赖的，因此 $\widetilde{\mathcal{R}}_m^K$ 和 $\widetilde{\mathcal{R}}_m$ 都是线性算子.

另一方面，考虑函数逼近的问题. 给定函数 $g(\boldsymbol{x}) \in C^0(\Omega)$，那么希望能够利用单元 K 周围的局部信息来构造出 $g(\boldsymbol{x})$ 的逼近函数. 仍然采用上述定义的模板 $S(K)$ 和点集 $\mathcal{I}_K$，并记 $g_K(\boldsymbol{x}) := g(\boldsymbol{x})|_{S(K)} \in C^0(S(K))$. 那么就有如下问题：求出 $q(\boldsymbol{x};g_K) \in \mathbb{P}_m(S(K))$，使其在点集 $\mathcal{I}_K$ 上能够在最小二乘的意义下逼近 $g_K(\boldsymbol{x})$. 由 (2.2) 和 (2.3) 可知，$q(\boldsymbol{x};g_K)$ 满足下述方程.

$$q(\boldsymbol{x};g_K) := \arg\min_{p(\boldsymbol{x})\in\mathbb{P}_m} \|p(\boldsymbol{x}) - g_K(\boldsymbol{x})\|_{\mathcal{I}_K}^2. \tag{3.7}$$

此时，也称 $q(\boldsymbol{x};g_K)$ 为函数 $g_K(\boldsymbol{x})$ 在 $S(K)$ 上的m 次重构多项式.

同样地也可以定义重构算子 $\mathcal{R}_m^K$：

$$\begin{aligned}\mathcal{R}_m^K &: C^0(S(K)) \mapsto \mathbb{P}_m(S(K)),\\ &\mathcal{R}_m^K g_K \to q(\boldsymbol{x};g_K).\end{aligned} \tag{3.8}$$

类似地，定义全局的 m 阶重构算子 $\mathcal{R}_m$，其在每个单元上的限制满足

$$(\mathcal{R}_m g)\big|_K := \left(\mathcal{R}_m^K g_K\right)\big|_K. \tag{3.9}$$

于是，$\mathcal{R}_m$ 将 $C^0(\Omega)$ 中的函数 $g(\boldsymbol{x})$ 映射成了分片的 m 阶多项式. 不难知道, $\mathcal{R}_m^K$ 和 $\mathcal{R}_m$ 也都是线性算子.

通过比较 $\widetilde{\mathcal{R}}_m^K$ 和 $\mathcal{R}_m^K$ 可以看出，两个算子只是定义域有所不同. 而且如果定义 $\boldsymbol{v}_K := \left(g_K(\boldsymbol{x}_1), g_K(\boldsymbol{x}_2), \cdots, g_K(\boldsymbol{x}_{n_K})\right)^{\mathrm{T}} \in \mathbb{R}^{n_K}$，那么有

$$\widetilde{\mathcal{R}}_m^K\boldsymbol{v}_K \equiv \mathcal{R}_m^K g_K.$$

同样的，如果定义 $\boldsymbol{v} := \Big(g(\boldsymbol{x}_1), g(\boldsymbol{x}_2), \cdots, g(\boldsymbol{x}_l)\Big)^{\mathrm{T}} \in \mathbb{R}^l$，那么有

$$\widetilde{\mathcal{R}}_m \boldsymbol{v} \equiv \mathcal{R}_m g. \tag{3.10}$$

这就意味着这两组算子在某种意义下是等价的，因此为简便起见，下面将只讨论 $\mathcal{R}_m^K$ 和 $\mathcal{R}_m$.

接下来，对于点集 $\mathcal{I}_K$ 提出如下的假设：

假设 A. 对任意的 $K \in \mathcal{T}_h$ 和 $g \in \mathbb{P}_m(S(K))$，

$$g|_{\mathcal{I}_K} = 0 \quad \Rightarrow \quad g|_{S(K)} \equiv 0.$$

于是就有下面的结论.

定理 3.3. *如果假设A成立，那么最小二乘问题(3.4)和(3.7) 都存在唯一解.*

证明. 由假设A可以自然地得到多项式函数空间 $\mathbb{P}_m(S(K))$ 的基函数系关于点集 $\mathcal{I}_K$ 是线性无关的，于是由定理2.1和定理2.2可以知道最小二乘问题(3.4)和(3.7)都存在唯一解. □

接下来，对于任意的 $K \in \mathcal{T}_h$ 和 $g \in C^0(S(K))$，定义

$$\Lambda(m, \mathcal{I}_K) = \max_{g \in \mathbb{P}_m(S(K))} \frac{\max_{\boldsymbol{x} \in S(K)} |g(\boldsymbol{x})|}{\max_{\boldsymbol{x} \in \mathcal{I}_K} |g(\boldsymbol{x})|}. \tag{3.11}$$

由假设A和空间 $\mathbb{P}_m(S(K))$ 的性质，有

$$\Lambda(m, \mathcal{I}_K) < \infty. \tag{3.12}$$

下面的定理给出了关于重构算子 $\mathcal{R}_m^K$ 的性质.

定理 3.4. *如果假设A成立,那么对于任意的 $K \in \mathcal{T}_h$ 和 $g(\boldsymbol{x}) \in C^0(S(K))$,由(3.8)定义的重构算子 $\mathcal{R}_m^K$，其连续性表述为*

$$\left\| \mathcal{R}_m^K g \right\|_{L^\infty(K)} \leqslant \Lambda(m, \mathcal{I}_K) \sqrt{\#\mathcal{I}_K} \max_{\boldsymbol{x} \in \mathcal{I}_K} |g(\boldsymbol{x})|, \tag{3.13}$$

逼近性表述为

$$\left\| g - \mathcal{R}_m^K g \right\|_{L^\infty(K)} \leqslant \left(1 + \Lambda(m, \mathcal{I}_K) \sqrt{\#\mathcal{I}_K}\right) \inf_{p \in \mathbb{P}_m(S(K))} \|g - p\|_{L^\infty(S(K))}. \tag{3.14}$$

上述结果可以在很多地方找到，如 [96, Theorem 2.1]. 现在将其证明过程叙述如下：

证明. 由假设A以及 $\mathcal{R}_m^K$ 的定义可知，

$$\mathcal{R}_m^K g = g, \quad \forall g \in \mathbb{P}_m(S(K)). \tag{3.15}$$

于是最小二乘问题(3.7)的意义是在离散范数 $\|\cdot\|_{\mathcal{I}_K}$ 意义下使 $g_K(\boldsymbol{x})$ 到 $\mathbb{P}_m(S(K))$ 的距离最小化，于是 $\mathcal{R}_m^K$ 可看作从 $C^0(S(K))$ 到 $\mathbb{P}_m(S(K))$ 的投影算子.

由 (3.12)，有

$$\left\|\mathcal{R}_m^K g\right\|_{L^\infty(K)} \leqslant \left\|\mathcal{R}_m^K g\right\|_{L^\infty(S(K))} \leqslant \Lambda(m, \mathcal{I}_K) \max_{\boldsymbol{x}\in\mathcal{I}_K} |R_m^K g(\boldsymbol{x})|.$$

又由投影算子 $\mathcal{R}_m^K$ 的性质，可以得到

$$\left\|\mathcal{R}_m^K g\right\|_{\mathcal{I}_K} \leqslant \|g\|_{\mathcal{I}_K}.$$

由范数 $\|\cdot\|_{\mathcal{I}_K}$ 的定义，很自然的有

$$\|g\|_{\mathcal{I}_K} \leqslant \sqrt{\#\mathcal{I}_K} \max_{\boldsymbol{x}\in\mathcal{I}_K} |g(\boldsymbol{x})|.$$

结合上述三个不等式，可以得到式 (3.13).

接下来，选择 $p_0 \in \mathbb{P}_m(S(K))$ 使得

$$\|g - p_0\|_{L^\infty(S(K))} = \inf_{p\in\mathbb{P}_m(S(K))} \|g - p\|_{L^\infty(S(K))}.$$

在式 (3.13)中用 $g - p_0$ 替换 g，并且由 (3.15)，可以得到

$$\begin{aligned}
\left\|\mathcal{R}_m^K g - p_0\right\|_{L^\infty(K)} &= \left\|\mathcal{R}_m^K (g - p_0)\right\|_{L^\infty(K)} \\
&\leqslant \Lambda(m, \mathcal{I}_K)\sqrt{\#\mathcal{I}_K} \max_{\boldsymbol{x}\in\mathcal{I}_K} |(g - p_0)(\boldsymbol{x})| \\
&\leqslant \Lambda(m, \mathcal{I}_K)\sqrt{\#\mathcal{I}_K}\, \|g - p_0\|_{L^\infty(S(K))} \\
&= \Lambda(m, \mathcal{I}_K)\sqrt{\#\mathcal{I}_K} \inf_{p\in\mathbb{P}_m(S(K))} \|g - p\|_{L^\infty(S(K))}.
\end{aligned}$$

因此，

$$\begin{aligned}\left\|g-\mathcal{R}_m^K g\right\|_{L^\infty(K)} &\leqslant \|g-p_0\|_{L^\infty(K)}+\left\|\mathcal{R}_m^K g-p_0\right\|_{L^\infty(K)}\\ &\leqslant \left(1+\Lambda(m,\mathcal{I}_K)\sqrt{\#\mathcal{I}_K}\right)\inf_{p\in\mathbb{P}_m(S(K))}\|g-p\|_{L^\infty(S(K))}.\end{aligned}$$

而这给出了式 (3.14). □

由 $\mathcal{R}_m$ 的定义以及定理3.8，不难得到关于 $\mathcal{R}_m$ 的下述结论.

推论 3.5. 如果假设A成立，那么对于任意的 $K\in\mathcal{T}_h$ 和 $g(\boldsymbol{x})\in C^0(\Omega)$，由(3.9)定义的重构算子 $\mathcal{R}_m$，其连续性表述为

$$\|\mathcal{R}_m g\|_{L^\infty(\Omega)}\leqslant \Lambda_m \max_{\boldsymbol{x}\in I_\Omega}|g(\boldsymbol{x})|, \tag{3.16}$$

逼近性表述为

$$\|g-\mathcal{R}_m g\|_{L^\infty(\Omega)}\leqslant (1+\Lambda_m)\max_{K\in\mathcal{T}_h}\inf_{p\in\mathbb{P}_m(S(K))}\|g-p\|_{L^\infty(S(K))}, \tag{3.17}$$

其中

$$\Lambda_m=\max_{K\in\mathcal{T}_h}\Lambda(m,\mathcal{I}_K)\sqrt{\#\mathcal{I}_K}.$$

下文中将 Λ_m 称为全局的重构常数，该常数的大小直接关系着重构算子的稳定性. 由其定义不难发现，为了给出 Λ_m 的大小，则必须先对 $\Lambda(m,\mathcal{I}_K)$ 和 $\#\mathcal{I}$ 作出估计. 下一节将集中讨论这个问题.

3.4 重构算子的稳定性

由 (3.14)可以知道，如果 $\Lambda(m,\mathcal{I}_K)\sqrt{\#\mathcal{I}_K}$ 可以被控制，那么离散最小二乘估计 $\mathcal{R}_m^K g$ 就是函数 g 的最佳一致逼近多项式. 因此，重构算子的稳定性集中体现在对重构常数的估计上.

考虑一个特殊的情形，当

$$\#\mathcal{I}_K=\binom{m+d}{d}=\dim\mathbb{P}_m(\mathbb{R}^d),$$

此时最小二乘问题退化为插值问题，因此可以将 $\Lambda(m,\mathcal{I}_K)\sqrt{\#\mathcal{I}_K}$ 替换为 Lebesgue

常数 $\mathcal{L}(S(K))$，其定义为

$$\mathcal{L}(S(K)) := \max_{\boldsymbol{x}\in S(K)} \sum_{\boldsymbol{x}_\ell\in\mathcal{I}_K} |w_{\boldsymbol{x}_\ell}(\boldsymbol{x})|,$$

这里 $w_{\boldsymbol{x}_\ell}$ 是关于 $\boldsymbol{x}_\ell$ 点的 Lagrange 插值基函数. 由假设A和 $\#\mathcal{I}_K = \binom{m+d}{d}$，有 $\mathcal{R}_m^K g = \Pi g$，这里 Π 是 Lagrange 插值算子. 于是对任意 $K \in \mathcal{T}_h$ 和 $g \in C^0(S(K))$，有

$$\left\|\mathcal{R}_m^K g\right\|_{L^\infty(K)} \leqslant \mathcal{L}(S(K)) \max_{\boldsymbol{x}\in\mathcal{I}_K} |g(\boldsymbol{x})|.$$

不幸的是，对于高维情形下的 Lebesgue 常数人们知之甚少 [103].

下面首先考虑对 $\#\mathcal{I}_K$ 进行估计. 为了强调和层数 t 的关系，记 $\mathcal{I}_t(K) = \mathcal{I}_K$，$S_t(K) = S(K)$. 在假设(3.1)和(3.3)下，有如下定理.

定理 3.6. *如果 $\mathcal{T}_h$ 满足逆假设(3.1)，且采样点集满足(3.3)，对于(3.2)定义的单元模板 $S_t(K)$，对于其上的采样点集 $\mathcal{I}_t(K)$，有*

$$\#\mathcal{I}_t(K) \leqslant 4M(t+1)^2\nu^2, \quad d = 2, \tag{3.18}$$

$$\#\mathcal{I}_t(K) \leqslant 8M(t+1)^3\nu^3, \quad d = 3. \tag{3.19}$$

证明. 由(3.2)可知，

$$\#\mathcal{I}_t(K) \leqslant M\#S_t(K).$$

于是只需给出 $\#S_t(K)$ 的估计.

又由(3.2)的定义，当 $d = 2$ 时，对于任意的 $K \in \mathcal{T}_h$，$S_t(K)$ 可以被一个以 K 的重心为圆心，以 $(t+1)h$ 为半径的圆所覆盖，于是

$$|S_t(K)| \leqslant \pi(t+1)^2h^2.$$

同时，由于 ρ_K 是 K 的内接圆半径，显然有

$$|S_t(K)| \geqslant \#S_t(K) \min_{K\in S_t(K)} \frac{\pi}{4}\rho_K^2,$$

则结合上述两式以及逆假设(3.1)，可知

$$\#S_t(K) \leqslant 4(t+1)^2\nu^2,$$

于是便得到了(3.18).

同理，在 $d=3$ 时可以得到

$$\#S_t(K) \leqslant 8(t+1)^3\nu^3,$$

也就得到了(3.19). □

需要指出的是，对 $\#\mathcal{I}_K$ 更具体的估计在很大程度上依赖于网格和采样点分布的具体性质. 比如在 [78] 中，采用了单纯形网格，并且采样点选为 $S(K)$ 中所有单元顶点的集合，于是有如下结论.

定理 3.7. *如果 $\mathcal{T}_h$ 满足逆假设(3.1)，并且 $S_t(K)$ 为凸，那么*

$$\#\mathcal{I}_t(K) \leqslant \pi(t^2+3t)\nu+3, \quad d=2, \tag{3.20}$$

$$\#\mathcal{I}_t(K) \leqslant \frac{2}{3}(2t^3+9t^2+13t)\nu^3+2t+4, \quad d=3. \tag{3.21}$$

证明. 对于任意的 $K\in\mathcal{T}_h$，$S_t(K)$ 可以被一个以 K 的重心为圆心，以 $(t+1)h$ 为半径的圆所覆盖. 注意到 $S_t(K)$ 为凸，于是有

$$p_t(K) \leqslant 2\pi(t+1)h,$$

这里 $p_t(K)$ 是 $S_t(K)$ 的周长. 同时还有

$$p_t(K) \geqslant \#v_{\mathrm{ex}} \min_{K\in S_t(K)} h_K \geqslant \#v_{\mathrm{ex}} \min_{K\in S_t(K)} \rho_K,$$

这里 $\#v_{\mathrm{ex}}$ 表示位于模板 $S_t(K)$ 的边界上的顶点的个数. 结合上述两个不等式以及逆假设(3.1)，可以得到

$$\#v_{\mathrm{ex}} \leqslant 2\pi(t+1)\nu.$$

考虑到 $\#v_{\mathrm{ex}} = \#I_{t,K} - \#I_{t-1,K}$，就有如下关系式.

$$\#I_{t,K} - \#I_{t-1,K} \leqslant 2\pi(t+1)\nu.$$

由上述递归关系即可得到(3.20).

对于 $d=3$，首先对单元 K 的任意面 F 的面积作出下界估计. 用 m_d 表示 K 中由 F 所对的顶点到 F 的距离，于是可知

$$|K| = \frac{m_d}{3}|F| = \frac{h_K}{3}|F|,$$

并且

$$|K| \geqslant \frac{4\pi}{3}\left(\frac{\rho_K}{2}\right)^3,$$

由此便有

$$|F| \geqslant \frac{\pi}{2\nu}\rho_K^2.$$

用 $\#f_t(K)$ 表示 $S_t(K)$ 边界上的面的个数. 注意到因为 $S_t(K)$ 为凸，则 $S_t(K)$ 的外表面的面积小于 $4\pi(t+1)^2h^2$. 将上面的不等式和这个结论相结合，就能得到

$$\#f_t(K)\min_{K\in S_t(K)}\frac{\pi}{2\nu}\rho_K^2 \leqslant 4\pi(t+1)^2H^2,$$

这就给出了 $\#f_t(K)$ 的上界估计.

$$\#f_t(K) \leqslant 8\nu^3(t+1)^2. \tag{3.22}$$

接下来，由欧拉公式[1]，

$$\#v_{\text{ex}} - \#e + \#f_t(K) = 2,$$

这里 v_{ex} 和 $\#e$ 分别是 $S_t(K)$ 的外表面上的顶点和边的个数. 考虑到每条边同时属于两个面，因此有

$$\#e = \frac{3}{2}\#f_t(K).$$

综合上述两个等式有

$$\#v_{\text{ex}} = \frac{1}{2}\#f_t(K) + 2.$$

上式可以写为如下的关系式

$$\#I_{t,K} - \#I_{t-1,K} = \frac{1}{2}\#F + 2.$$

将 (3.22)带入到上述等式中即可得到关于 $\#I_{t,K}$ 的递推关系，由此便可得到 (3.21). □

接下来考虑参数 $\Lambda(m,\mathcal{I}_K)$. 如果 $d=1$ 并且采样点是等距分布的，那么

[1] 简单多面体的顶点数 V，面数 F 及棱数 E 之间满足欧拉公式：

$$V + F - E = 2.$$

Coppersmitht 和 Rivlin [43] 证明了

$$\Lambda(m,\mathcal{I}_K) \simeq \exp\left(\frac{cm^2}{\#\mathcal{I}_K}\right), \tag{3.23}$$

这里 c 是一个绝对常数. 关于这个结果的讨论详见 [94].

下面的定理在二维和三维情形下给出了拟一致网格上的 $\Lambda(m,\mathcal{I}_K)$ 的上界估计. 在定理中，如果 $\Lambda(m,\mathcal{I}_K)$ 满足一定的条件，那么它将服从一致的上界，且上界为 2.

定理 3.8. *假设 $S(K)$ 是一个凸多边形 (多面体)，并且满足条件*

$$\begin{aligned} m &< \left(\frac{\pi\#S(K)}{16\#v_{ex}}\right)^{1/2}\nu^{-1}, \qquad d=2,\\ m &< \left(\frac{\pi\#S(K)}{12\#f_{ex}}\right)^{1/2}\nu^{-3/2}, \qquad d=3, \end{aligned} \tag{3.24}$$

那么

$$\Lambda(m,\mathcal{I}_K) \leqslant 2, \tag{3.25}$$

这里 $\#v_{ex}$ 是 $S(K)$ 边界上的节点个数，并且 $\#f_{ex}$ 是 $S(K)$ 边界上的面的个数.

由3.3节对网格中采样点分布作出的假设以及 (3.3)，于是有

$$\#S(K) \simeq \#\mathcal{I}_K.$$

因此在 $d=2$ 时有

$$\#v_{\mathrm{ex}} \simeq \sqrt{\#S(K)} \simeq \sqrt{\#\mathcal{I}_K}. \tag{3.26}$$

在 $d=3$ 时，有

$$\#f_{\mathrm{ex}} \simeq \left(\#S(K)\right)^{2/3} \simeq \left(\#\mathcal{I}_K\right)^{2/3}. \tag{3.27}$$

将(3.26)和(3.27)带入(3.24)，可知采样点个数需满足

$$\#\mathcal{I}_K \simeq m^{2d} \tag{3.28}$$

才能保证式 (3.24)成立. 这一点是和一维情形下的估计(3.23)相容的. 注意到式 (3.28)中要求的采样点个数要远大于 $\dim\mathbb{P}_m(\mathbb{R}^d)$,而这恰恰是为了能让 $\Lambda(m,\mathcal{I}_K)$ 服从一致上界所要付出的代价. 而且由(3.26) 和(3.27)可知 $\#S(K)$ 的增长速度总是要快过 $\#v_{\mathrm{ex}}$ 和 $\#f_{\mathrm{ex}}$，因此理论上总是可以通过扩大模板 $S(K)$ 的规模来使条件(3.24)成立.

证明. 首先对二维情形进行证明. 令 $\widetilde{\boldsymbol{x}} \in S(K)$ 使得 $|p(\widetilde{\boldsymbol{x}})| = \max\limits_{\boldsymbol{x}\in S(K)} |p(\boldsymbol{x})|$. 记 $\widetilde{\boldsymbol{x}}_\ell := \arg\min\limits_{\boldsymbol{y}\in\mathcal{I}_K} |\widetilde{\boldsymbol{x}} - \boldsymbol{y}|$. 由 Taylor 展开有

$$p(\widetilde{\boldsymbol{x}}_\ell) = p(\widetilde{\boldsymbol{x}}) + (\widetilde{\boldsymbol{x}}_\ell - \widetilde{\boldsymbol{x}}) \cdot \nabla p(\xi_{\boldsymbol{x}}),$$

其中 $\xi_{\boldsymbol{x}}$ 在 $\widetilde{\boldsymbol{x}}$ 和 $\widetilde{\boldsymbol{x}}_\ell$ 为端点的线段上. 于是有

$$|p(\widetilde{\boldsymbol{x}})| \leqslant |p(\widetilde{\boldsymbol{x}}_\ell)| + h_K \max_{\boldsymbol{x}\in S(K)} |\nabla p(\boldsymbol{x})|,$$

这里

$$|\nabla p(\boldsymbol{x})| = \left(\sum_{i=1}^{d} \left| \frac{\partial p}{\partial x_i}(\boldsymbol{x}) \right|^2 \right)^{1/2}.$$

由 Markov 不等式 [111]，有

$$\max_{\boldsymbol{x}\in S(K)} |\nabla p(\boldsymbol{x})| \leqslant \frac{4m^2}{w(K)} \max_{\boldsymbol{x}\in S(K)} |p(\boldsymbol{x})|, \tag{3.29}$$

这里 $w(K)$ 是凸多边形 $S(K)$ 的宽度.

接下来确定 $w(K)$ 的下界. 为此, 引入如下关于平面凸集的不等式 [74]

$$2\,|S(K)| \leqslant w(K) p_K, \tag{3.30}$$

这里 p_K 表示 $S(K)$ 的直径. 容易看到

$$p_K \leqslant \#v_{\mathrm{ex}} h.$$

令 $\#S(K)$ 为 $S(K)$ 中单元的个数，那么有

$$|S(K)| \geqslant \#S(K) \min_{K\in\mathcal{T}_h} \frac{\pi}{4} \rho_K^2.$$

将上述两个不等式代入 (3.30)，就得到关于 $w(K)$ 的一个下界：

$$w(K) \geqslant \frac{\pi}{2} \min_{K\in S(K)} \frac{\rho_K^2}{h} \frac{\#S(K)}{\#v_{\mathrm{ex}}}. \tag{3.31}$$

将以上不等式代入 (3.29)，且考虑到 (3.24)中的第一个不等式，便有

$$\max_{\boldsymbol{x}\in S(K)}|p(\boldsymbol{x})| \leqslant |p(\widetilde{\boldsymbol{x}}_\ell)| + \frac{1}{2}\max_{\boldsymbol{x}\in S(K)}|p(\boldsymbol{x})|.$$

这样就在 $d=2$ 的情形下得到了 (3.25).

对于 $d=3$ 的情况下的证明在本质上与二维情形相同. 这是因为不等式 (3.29)和 (3.30) 对多面体也成立. □

定理3.8给出了一个一般情况下的 $\Lambda(m,\mathcal{I}_K)$ 具有一致上界的充分条件，但该条件 (3.24)显得过于复杂，容易让人困惑. 下面的定理将在 $d=2$ 且为三角形网格剖分的情况下给出一个更加清晰明确的条件，其中只包含 m、ν 和 t. 在该条件成立时依然能得到估计式 (3.25).

定理 3.9. *如果在三角形网格剖分 $\mathcal{T}_h$ 中，$S(K)$ 是一个凸多边形，并且*

$$m < \left(\frac{3\sqrt{3}(t^2+4\nu)}{8\pi(t+1)}\right)^{1/2}\nu^{-3/2}, \tag{3.32}$$

那么

$$\Lambda(m,\mathcal{I}_K) \leqslant 2. \tag{3.33}$$

证明. 令 $\widetilde{\boldsymbol{x}}\in S(K)$ 满足 $|p(\widetilde{\boldsymbol{x}})| = \max\limits_{\boldsymbol{x}\in S_t(K)}|p(\boldsymbol{x})|$. 如果 $\widetilde{\boldsymbol{x}}\in\mathcal{I}_K$，那么取 $\Lambda(m,\mathcal{I}_K)=1$. 否则，记 $\widetilde{\boldsymbol{x}}_\ell = \arg\min\limits_{\boldsymbol{y}\in\mathcal{I}_K}|\boldsymbol{y}-\widetilde{\boldsymbol{x}}|$. 如果 $\widetilde{\boldsymbol{x}}$ 在 $S_t(K)$ 的边界上，那么 $|\widetilde{\boldsymbol{x}}_\ell - \widetilde{\boldsymbol{x}}| \leqslant h_K/2$，由 Taylor 展开，便有

$$p(\widetilde{\boldsymbol{x}}_\ell) = p(\widetilde{\boldsymbol{x}}) + (\widetilde{\boldsymbol{x}}_\ell - \widetilde{\boldsymbol{x}})\cdot\nabla p(\xi_{\boldsymbol{x}}),$$

其中点 $\xi_{\boldsymbol{x}}$ 位于以 $\widetilde{\boldsymbol{x}}$ 和 $\widetilde{\boldsymbol{x}}_\ell$ 为端点的线段上. 于是得到了方程

$$|p(\widetilde{\boldsymbol{x}})| \leqslant |p(\widetilde{\boldsymbol{x}}_\ell)| + \frac{h}{2}\max_{\boldsymbol{x}\in S(K)}|\nabla p(\boldsymbol{x})|. \tag{3.34}$$

如果 $\widetilde{\boldsymbol{x}}$ 在 $S_t(K)$ 的内部，那么 $\nabla p(\widetilde{\boldsymbol{x}})=0$，由 Taylor 展开有

$$p(\widetilde{\boldsymbol{x}}_\ell) = p(\widetilde{\boldsymbol{x}}) + \frac{1}{2}(\widetilde{\boldsymbol{x}}_\ell - \widetilde{\boldsymbol{x}})^2\cdot\nabla^2 p(\xi_{\boldsymbol{x}}),$$

这里点 $\xi_{\boldsymbol{x}}$ 位于以 $\widetilde{\boldsymbol{x}}$ 和 $\widetilde{\boldsymbol{x}}_\ell$ 为端点的线段上. 这就推出了

$$|p(\widetilde{\boldsymbol{x}})| \leqslant |p(\widetilde{\boldsymbol{x}}_\ell)| + \frac{h^2}{2}\max_{\boldsymbol{x}\in S_t(K)}|\nabla^2 p(\boldsymbol{x})|. \tag{3.35}$$

将 Markov 不等式(3.29)分别应用于(3.34)和(3.35)，可以得到

$$\max_{\boldsymbol{x}\in S_t(K)}|p(\boldsymbol{x})| \leqslant \max_{\boldsymbol{x}\in \mathcal{I}_K}|p(\boldsymbol{x})| + \frac{2m^2h}{w(K)}\max_{\boldsymbol{x}\in S_t(K)}|p(\boldsymbol{x})|,$$

和

$$\max_{\boldsymbol{x}\in S_t(K)}|p(\boldsymbol{x})| \leqslant \max_{\boldsymbol{x}\in \mathcal{I}_K}|p(\boldsymbol{x})| + \frac{8m^4h^2}{w^2(K)}\max_{\boldsymbol{x}\in S_t(K)}|p(\boldsymbol{x})|.$$

由于 K 是一个凸多边形，于是有

$$p_t(K) \geqslant 2t \min_{K\in S_t(K)} h_K \geqslant 2t \min_{K\in S_t(K)} \rho_K.$$

注意到

$$p_t \leqslant (\#\mathcal{I}_t(K) - \#\mathcal{I}_{t-1}(K))\, h.$$

由上述两个不等式能得到递推关系

$$\#\mathcal{I}_t(K) - \#\mathcal{I}_{t-1}(K) \geqslant 2t/\nu,$$

同时由 $\#\mathcal{I}_0(K) = 3$，可以得到

$$\#\mathcal{I}_t(K) \geqslant 3 + t(t+1)/\nu.$$

接下来，由欧拉公式，有

$$\#S_t(K) = \#\mathcal{I}_t(K) + \#\mathcal{I}_{t-1}(K) + 2.$$

然后由上面的三个关系式，可知

$$\begin{aligned}\#S_t(K) &= \#\mathcal{I}_t(K) - \#\mathcal{I}_{t-1}(K) + 2\#\mathcal{I}_{t-1}(K) + 2\\ &\geqslant 8 + \frac{2t^2}{\nu}.\end{aligned}$$

注意到模板 $S(K)$ 能被以 K 的重心为圆心，以 $(t+1)h$ 为半径的圆所覆盖因此有

$$p_t(K) \leqslant 2\pi(t+1)h.$$

由 Finsler-Hadwiger 不等式 [54] 有

$$|K| \geqslant \frac{3\sqrt{3}}{4}\rho_K^2.$$

由上述三个不等式以及 (3.30)，有如下的关于宽度 $w(K)$ 的下界估计：

$$w(K) \geqslant \frac{3\sqrt{3}\,(4+t^2\nu^{-1})}{2\pi(t+1)\nu} \min_{K\in S_t(K)} \rho_K.$$

由条件 (3.32)，可知

$$\max_{\boldsymbol{x}\in S_t(K)} |p(\boldsymbol{x})| \leqslant \max_{\boldsymbol{x}\in \mathcal{I}_K} |p(\boldsymbol{x})| + \frac{1}{2} \max_{\boldsymbol{x}\in S_t(K)} |p(\boldsymbol{x})|,$$

由此就得到了 (3.33). □

由定理3.6到定理3.8可以得到这样的结论：在一定条件下 $\#\mathcal{I}_K$ 和 $\Lambda(m,\mathcal{I}_K)$ 是可以被一致上界控制的，于是重构常数 Λ_m 的上界也可以显式地写出来. 需要强调的是，上面几个定理只是给出了 Λ_m 有界的一个充分条件. 而在实际计算中，并不是说这些条件必须满足才能进行计算。下面将通过数值算例来说明这一点.

3.5 数值算例

在定理 3.8和定理 3.9中，均要求 $S(K)$ 是凸的，但在实际中这个假设并不总是能被满足. 本节将通过几个数值算例来对这一问题加以阐述. 下面通过两个例子来计算重构常数 $\Lambda(m,\mathcal{I}_K)$，这两个例子都是在三角形单元网格上进行计算. 模板 $S(K)$ 的构造遵循(3.2)所规定的方式，并且点集 $\mathcal{I}_K$ 选为模板 $S(K)$ 中所有单元顶点的集合.

例 3.1 在单元 K 的模板 $S(K)$ 上考虑函数 $g(\boldsymbol{x})$，

$$\begin{aligned} g(\boldsymbol{x}) = {} & (x_1+x_2-0.5)^2(x_1-x_2+1)^2(5-4(x-0.2)^2-4(x_2-0.6)^2) \\ & + \frac{1}{2.1+\sin\left(\dfrac{3}{2}\pi x_1-3\right)+\cos(2\pi x_2)}, \quad \boldsymbol{x}\in S(K),\ K\in\mathcal{T}_h, \end{aligned}$$

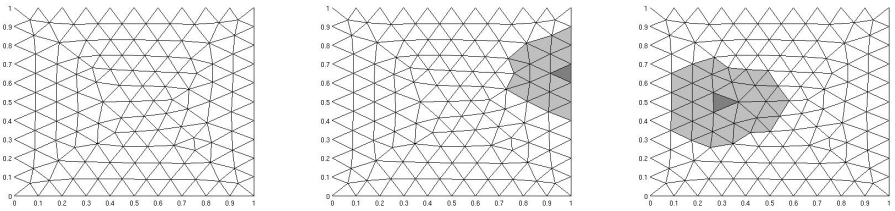
并且令 $p(\boldsymbol{x}) := \mathcal{R}_m^K g(\boldsymbol{x})$. 定义正则化的离散范数

$$\|p\|_{l^\infty(\mathcal{I}_K)} = \max_{\boldsymbol{x}\in\mathcal{I}_K} |p(\boldsymbol{x})|,$$

则重构系数可用下式估计

$$\Lambda(m,\mathcal{I}_K) \simeq \frac{\|p\|_{L^\infty(S(K))}}{\|p\|_{l^\infty(\mathcal{I}_K)}}.$$

下面，在三角形单元网格中选取两种类型的单元来测试. 如图 3.2(a)所示，一个单元在区域边界上，另一个单元在区域内部. 对每个单元，分别向外扩展两层来构造模板 $S_2(K)$，见图 3.2(b)和图 3.2(c). 接下来，在两个模板上分别计算 $\Lambda(m,\mathcal{I}_K)$，并将结果展示在表 3.1 和表 3.2中. 从表中可以明显的看出，$\Lambda(m,\mathcal{I}_K)$ 只是比 1.0 稍大. 在其他单元的模板上进行计算也能得到类似的结果. 由此可知在形状规则的网格上，$\Lambda(m,\mathcal{I}_K)$ 是可以被控制住的. 注意到图 3.2(b)和图 3.2(c)中，$S(K)$ 都是非凸的，这一现象暗示着定理 3.8和定理 3.9中的凸性假设也许可以去掉.



(a) 网格尺寸 $h = 0.1$　　(b) K 在区域边界　　(c) K 在区域内部

图 *3.2: 例3.1*：网格以及两个非凸模板 $S(K)$ 的例子.

m	$\|p\|_{L^\infty(S(K))}$	$\|p\|_{\ell^\infty(\mathcal{I}_K)}$	$\Lambda(m,\mathcal{I}_K)$
2	6.2361	6.1981	1.0061
3	6.2534	6.2210	1.0052
4	6.2575	6.2229	1.0059

表 *3.1: 例3.1*：K 在区域边界时的重构常数 $\Lambda(m,\mathcal{I}_K)$.

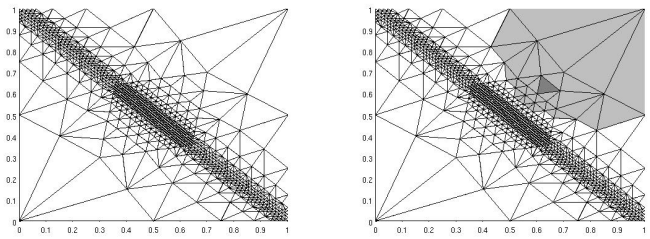
m	$\|p\|_{L^\infty(S(K))}$	$\|p\|_{\ell^\infty(\mathcal{I}_K)}$	$\Lambda(m,\mathcal{I}_K)$
2	4.8370	4.7774	1.0125
3	4.9150	4.8217	1.0193
4	6.9050	6.3542	1.0867

表 *3.2: 例3.1*：K 在区域内部时的重构常数 $\Lambda(m,\mathcal{I}_K)$.

例 3.2 在这个例子中，令函数 $g(\boldsymbol{x})$ 为

$$g(\boldsymbol{x}) = 0.01 + \exp\left(-\frac{(x_1-0.85)^2+(x_2-0.85)^2}{2\sigma^2}\right), \quad \boldsymbol{x} \in S(K),\ K \in \mathcal{T}_h,$$

考虑一个自适应加密的三角形网格. 如图3.3(a)所示，越靠近区域反对角线的网格被加密得越细致. 这样的网格的 ν 要比拟一致网格的大，因此需要更大的单元模板才能满足定理 3.8 中的条件. 选取一个区域内部的单元 K，并构造两层的模板 $S_2(K)$，见图3.3(b). 在该模板上计算重构常数 $\Lambda(m,\mathcal{I}_K)$，并将结果展示在表3.3中. 可以看出 $\Lambda(m,\mathcal{I}_K)$ 会随着多项式次数 m 的增长而增长. 但如果增加单元模板的规模，即增加层数 t，那么如表3.4所示，重构常数 $\Lambda(m,\mathcal{I}_K)$ 会逐渐的变小，并被控制到 2.0 以内. 因此，通过增加 $S(K)$ 的规模来控制重构常数的方法是可行的.



(a) 自适应加密网格 (b) 单元及其模板 $S(K)$

图 *3.3: 例3.2：自适应加密网格和单元模板 $S(K)$ 的例子.*

m	$\|p\|_{L^\infty(S(K))}$	$\|p\|_{\ell^\infty(\mathcal{I}_K)}$	$\Lambda(m,\mathcal{I}_K)$
2	0.2257	0.2234	1.0101
3	0.3903	0.2193	1.7794
4	0.8162	0.2198	3.7131

表 *3.3: 例3.2：自适应网格上的重构常数 $\Lambda(m,\mathcal{I}_K)$.*

	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$
$m=2$	1.0101	1.0000	1.0000	1.0000	1.0000
$m=3$	1.7794	1.4165	1.1808	1.0768	1.0341
$m=4$	3.7131	1.6604	1.9327	1.8771	1.7860

表 *3.4: 例3.2：当层数 t 增长时重构常数 $\Lambda(m,\mathcal{I}_K)$ 的变化.*

3.6　小结

本章给出了网格上的多项式重构算子的定义，并讨论了它的连续性和逼近性. 对于不等式中的重构常数，尝试着给出了其在一般情形下以及特殊情形下的上界估计，并且得出了该常数在一定条件下具有一致的上界. 最后用数值算例给出了实际计算中的重构常数的例子，这些例子说明通常情况下重构常数并不大. 即使在有些情况下重构常数较大，也可以通过扩大单元模板来降低该参数的值.

第四章　基于重构有效系数的异质多尺度方法

在异质多尺度有限元方法中，宏观求解器的缺失数据是数值积分点上的有效系数矩阵. 在传统的 HMM-FEM 方法 [114, 87, 2, 6] 中，往往是通过求解数值积分点上的单胞问题来给出有效系数矩阵的估计. 而大量单胞问题的数值求解恰恰是 HMM-FEM 方法的时间代价的主要来源. 尤其是当采用高阶宏观求解器时，算法的时间消耗会随着单胞问题数量的快速增加而迅速增长. 杜锐和明平兵 [48] 对于线性元和二次元的宏观求解器提出了一种新的数值积分准则. 该准则倾向于尽量选取单元边界中点或单元顶点来作为数值积分点，这一点和所谓的好的数值积分公式 [93] 所要求的数值积分点尽可能在单元内部相违背. 但是，这种积分格式使得一个单胞问题的计算结果能够被多个单元所使用，也就变相地降低了单胞问题的个数，在保持原有精度的前提下提高了 HMM-FEM 方法的计算效率. 汪康 [109] 应用该方法实现了三维情形下的数值算例，进一步验证了该方法的有效性. 不幸的是，由于高阶的数值积分点会更多地出现在单元内部 [108]，因此这种方法很难进一步推广到高阶元的情形.

本章将提出一种新的基于最小二乘重构的异质多尺度有限元方法. 对于 P_k 元的宏观求解器，该方法将网格中的所有节点选作采样点，而采样值则是采样点上的有效系数矩阵的近似值，这些近似值是通过在采样点上求解局部的单胞问题而得到的. 然后，对每个单元在其周围选取一个单元模板，并在单元模板上用 m 阶的最小二乘重构来对所有采样点上的采样值进行拟合. 最后，拟合出的多项式就作为单元上的有效系数矩阵的近似来进行数值计算. 在这个过程中，m 的大小是由 k 来决定的 (详见 §4.4节). 对于标准的网格，顶点的个数比边的数量以及高阶的数值积分点的数量要少的多，因此在顶点上计算单胞问题将对计算效率带来数倍的提升，而且这样一来不必为网格结构制定特殊的数值积分格式，使得整个方法的实现更加灵活，可以支持高阶的宏观求解器。本章还从理论以及数值算例两方面验证了新方法能够达到最优收敛阶，但是时间代价仅仅和线性元宏观求解器相当. 图4.1用示意图的方式更加清楚地展示了本章提出的方法和传统方法之间的

区别.

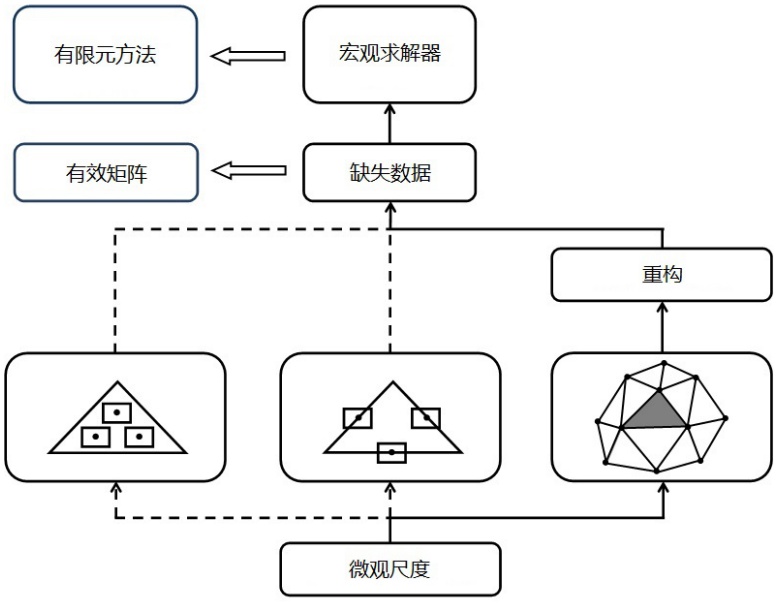


图 *4.1: HMM-FEM* 方法的演变示意图

对于方程 (1.8)，已经有不少学者 [51, 1, 48, 47] 对 HMM-FEM 方法的收敛性进行了研究. 他们的结果主要集中在周期边界条件和 Dirichlet 边界条件的子问题的离散误差上，但是少有关于其它边界条件 [63] 的结果. 而在本章中，将对满足 Dirichlet 边界条件、Neumann 边界条件以及周期边界条件的子问题给出一个统一性的离散误差估计. 其中的证明过程基于一个均匀化理论 [16] 中的插值算子，并要求子问题的解满足合理的正则性假设.

本章的 §4.1节简单回顾了经典的均匀化理论和渐进展开的相关结论; 在 §4.2节中给出了基于重构的高阶异质多尺度方法的方法的描述; §4.3 节对满足三种边界条件的子问题给出了统一性的误差分析，并基于这个结果得到了所提出的方法的收敛性结果; §4.4节给出了相应的数值算例; 最后 §4.5 节对本章的内容做了小结.

4.1　椭圆形均匀化问题

本节将以具有快速振荡系数的二阶椭圆型方程(1.8)为例来对经典的均匀化理论中 H-收敛和周期渐进展开的结论进行回顾.

假设方程中的系数 $a^\varepsilon \in \mathcal{M}(\alpha, \beta, \Omega)$(定义见 §2.3节)，但不要求其对称. 在经典的均匀化理论中，关于方程(1.8)有如下结论. 在 H-收敛 (见 [105]) 的意义下，对

于每个 $a^\varepsilon \in \mathcal{M}(\alpha,\beta,\Omega)$ 和 $f \in H^{-1}(\Omega)$，方程(1.8)的解序列 $\{u^\varepsilon\}$ 满足

$$\begin{aligned} u^\varepsilon &\rightharpoonup U_0, && \text{在}H_0^1(\Omega)\text{中弱收敛,} \\ a^\varepsilon \nabla u^\varepsilon &\rightharpoonup \mathcal{A}\nabla U_0, && \text{在}[L^2(\Omega)]^d\text{中弱收敛,} \end{aligned}$$

其中均匀化解 U_0 满足均匀化方程

$$\begin{cases} -\operatorname{div}\left(\mathcal{A}(\boldsymbol{x})\nabla U_0(\boldsymbol{x})\right) = f(\boldsymbol{x}), & \boldsymbol{x}\in\Omega, \\ U_0(\boldsymbol{x}) = 0, & \boldsymbol{x}\in\partial\Omega, \end{cases} \tag{4.1}$$

在(4.1)中，$\mathcal{A}(\boldsymbol{x})$ 被称为均匀化系数或有效系数，用来描述系统在 $O(1)$ 尺度上的性质. 除了一维情形和拟一维情形以外，有效系数 $\mathcal{A}(\boldsymbol{x})$ 一般没有显式表达式 [105].

系数 $a^\varepsilon = \boldsymbol{a}(\boldsymbol{x},\boldsymbol{x}/\varepsilon) = \boldsymbol{a}(\boldsymbol{x},\boldsymbol{y})$ 包含两个尺度：$O(\varepsilon)$ 尺度和 $O(1)$ 尺度. 如果系数 a^ε 是一个局部周期矩阵，也就是说 $\boldsymbol{a}(\boldsymbol{x},\boldsymbol{y})$ 关于 $\boldsymbol{y}$ 是 Y-周期函数，那么根据周期结构的多尺度渐进展开 [16]，那么均匀化系数 $\mathcal{A}(\boldsymbol{x})$ 可以表示为

$$\mathcal{A}_{ij}(\boldsymbol{x}) = \fint_Y \left(a_{ij} + a_{ik}\frac{\partial\chi^j}{\partial y_k}\right)(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}, \quad \forall i,j = 1,2,\cdots,d, \tag{4.2}$$

其中辅助函数 $\boldsymbol{\chi}(\boldsymbol{x},\boldsymbol{y}) = \{\chi^j(\boldsymbol{x},\boldsymbol{y})\}_{j=1}^d$ 是关于 $\boldsymbol{y}$ 的 Y-周期函数，并且满足

$$\frac{\partial}{\partial y_i}\left(a_{ik}\frac{\partial\chi^j}{\partial y_k}\right)(\boldsymbol{x},\boldsymbol{y}) = -\left(\frac{\partial}{\partial y_i}a_{ij}\right)(\boldsymbol{x},\boldsymbol{y}) \quad \boldsymbol{y}\in Y, \quad \int_Y \chi^j(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = 0. \tag{4.3}$$

根据定理2.7，这个问题的解是存在的 [16]. 在(4.3)两边同时乘以 χ^j，然后在 Y 上积分. 由分部积分得到：对任意的 $j = 1,2,\cdots,d$，有

$$\left\|\nabla_{\boldsymbol{y}}\chi^j(\boldsymbol{x},\boldsymbol{y})\right\|_{L^2(Y)} \leqslant \beta/\alpha, \quad \forall\boldsymbol{x}\in\Omega,\ \forall\boldsymbol{y}\in Y. \tag{4.4}$$

直接的推导可知有效系数 $\mathcal{A}\in\mathcal{M}(\alpha,\beta^2/\alpha,\Omega)$，特别地，如果系数矩阵 a^ε 是对称的，那么 $\mathcal{A}\in\mathcal{M}(\alpha,\beta,\Omega)$[105]. 在一维情形下，局部周期振荡系数对应的有效系数 $\mathcal{A}(\boldsymbol{x})$ 可以由其调和平均给出，即

$$\mathcal{A}(\boldsymbol{x}) = \left(\fint_Y \frac{1}{a(\boldsymbol{x},\boldsymbol{y})}\,\mathrm{d}\boldsymbol{y}\right)^{-1}.$$

在高维情形下，分层材料 (layered materials) 的弹性系数仅与一个方向的自变量

有关且是周期的，相应的有效系数也具有显式的表达式，可参见 [33, 定理 5.10].

4.2　算法描述

记 $\mathcal{T}_H$ 为区域 Ω 上如3.1节所定义的三角形剖分网格，其网格尺寸为 H. 为了区别起见，下文中将把 $\mathcal{T}_H$ 称为宏观网格. 选取传统的 k 次有限元 P_k 作为宏观求解器，相应的有限元空间记为 $\mathcal{V}_H$，即

$$\mathcal{V}_H := \{v \in H_0^1(\Omega) \mid v|_K \in P_k(K),\ K \in \mathcal{T}_H\}.$$

异质多尺度有限元方法的解 $U_H \in \mathcal{V}_H$ 满足

$$a_H(U_H, V) = (f, V), \quad \forall V \in \mathcal{V}_H, \tag{4.5}$$

其中双线性算子 a_H 定义为：对任意 $V, W \in \mathcal{V}_H$，

$$a_H(V, W) = \sum_{K\in\mathcal{T}_H} \int_K \nabla W(\boldsymbol{x}) \cdot \mathcal{A}_H(\boldsymbol{x}) \nabla V(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}. \tag{4.6}$$

在实际计算中，(4.6)式的积分是由数值积分格式来计算的. 但 $\mathcal{A}_H(\boldsymbol{x})$ 在每个数值积分点上的值是缺失数据，因此需要通过某种方式来给出 $\mathcal{A}_H(\boldsymbol{x})$ 在每个数值积分点上的近似.

为此，首先对网格中的每个节点 $\boldsymbol{x}_\ell$ 构造单胞

$$I_\delta := \boldsymbol{x}_\ell + \delta Y, \quad Y := \left(-\frac{1}{2}, -\frac{1}{2}\right)^d,$$

其中 δ 用来表示单胞的尺寸. 对单胞进行三角形剖分，得到微观网格 $\mathcal{T}_h$，网格尺寸为 h. 接下来在 I_δ 上选取 $P_{k'}$ 为微观求解器并求解如下问题：求 $v_h^\varepsilon - V_\ell \in \mathcal{V}_h$，使得

$$(a^\varepsilon \nabla v_h^\varepsilon, \nabla \varphi)_{L^2(I_\delta)} = 0, \quad \forall \varphi \in \mathcal{V}_h, \tag{4.7}$$

这里 $V_\ell \equiv V(\boldsymbol{x}_\ell) + (\boldsymbol{x} - \boldsymbol{x}_\ell) \cdot \nabla V(\boldsymbol{x}_\ell)$ 是 V 在 $\boldsymbol{x}_\ell$ 点的线性近似. 这个方程被称为**单胞问题**或**子问题**，其中的有限元空间 $\mathcal{V}_h$ 因边界条件而异.

1. 如果

$$\mathcal{V}_h = \mathcal{V}_{D,h} := \left\{ v \in H_0^1\Big(I_\delta(\boldsymbol{x}_\ell)\Big) \mid v|_K \in \mathbb{P}_{k'}(K),\ \ K \in \mathcal{T}_h \right\}.$$

则称 (4.7)为 Dirichlet 子问题.

2. 如果

$$\mathcal{V}_h = \mathcal{V}_{N,h} := \left\{ v \in H^1\Big(I_\delta(\boldsymbol{x}_\ell)\Big) \mid v|_K \in \mathbb{P}_{k'}(K), \langle \nabla v \rangle_{I_\delta} = 0 \ \ K \in \mathcal{T}_h \right\}.$$

则称 (4.7)为 Neumann 子问题.

3. 如果

$$\mathcal{V}_h = \mathcal{V}_{P,h} := \left\{ v \in H^1_{\mathrm{per}}\Big(I_\delta(\boldsymbol{x}_\ell)\Big) \mid v|_K \in \mathbb{P}_{k'}(K), \langle v \rangle_{I_\delta} = 0 \ \ K \in \mathcal{T}_h \right\},$$

则称 (4.7)为周期子问题. 这里 $H^1_{\mathrm{per}}\Big(I_\delta(\boldsymbol{x}_\ell)\Big)$ 是 $C^\infty_{\mathrm{per}}\Big(I_\delta(\boldsymbol{x}_\ell)\Big)$ 在 H^1 范数下生成的闭包.

对于这三种子问题的详细讨论可以参考 [114] 和 [47]. 求解这些子问题之后，就可以按照如下方式估计出每个顶点 $\boldsymbol{x}_\ell$ 上 $\widetilde{\mathcal{A}}_H(\boldsymbol{x}_\ell)$ 的值

$$\widetilde{\mathcal{A}}_H(\boldsymbol{x}_\ell) \langle \nabla v_h^\varepsilon \rangle_{I_\delta} := \langle a^\varepsilon \nabla v_h^\varepsilon \rangle_{I_\delta}. \tag{4.8}$$

在这里用 $\langle \cdot \rangle_{I_\delta}$ 表示单胞 I_δ 上的积分平均.

接下来，为每个单元 K 构造一个凸的单元模板 $S(K)$，构造方法可以参考(3.2). 在每个模板上，令采样点集 $\mathcal{I}_K$ 为 $S(K)$ 上所有网格节点的集合. 于是就可以用最小二乘方法来重构单元 K 上的有效系数矩阵，即

$$(\mathcal{A}_H)_{ij} = \arg \min_{p \in \mathbb{P}_m(S(K))} \left\| \Big(\widetilde{\mathcal{A}}_H\Big)_{ij} - p \right\|^2_{\mathcal{I}_K}. \tag{4.9}$$

注. 在此假设 $S(K)$ 是凸的，是为了满足定理 3.7和定理 3.8的条件. 然而正如上一章中所论述的 (详见 §3.5节)，本节所提的方法也能应用到非凸的单元模板 $S(K)$ 上.

考虑到越靠近单元的节点带给单元的信息一般也会越多，因此还可以考虑另外一种定义有效系数矩阵的方式. 具体来说，是采用带等式约束的最小二乘方法：

$$(\mathcal{A}_H)_{ij} = \arg \min_{p \in \mathbb{P}_m(S(K))} \left\| \Big(\widetilde{\mathcal{A}}_H\Big)_{ij} - p \right\|^2_{\mathcal{I}_K} \tag{4.10}$$

并服从约束

$$p(\boldsymbol{x}_\ell) = (\widetilde{\mathcal{A}}_H)_{ij}(\boldsymbol{x}_\ell), \quad \forall \boldsymbol{x}_\ell \in K.$$

§4.4节中的数值算例验证了两种定义方式具有同样的收敛阶，但是带等式约束的方式在数值结果上要更精确一些.

4.3　收敛性分析

本节将对上一节提出的方法的适定性和收敛性进行理论分析.

4.3.1　适定性证明

首先定义

$$e(\mathrm{HMM}) := \max_{\substack{\boldsymbol{x}\in K\\ K\in\mathcal{T}_H}} \|(\mathcal{A}-\mathcal{A}_H)(\boldsymbol{x})\|_F\,,$$

这里 $\|\cdot\|_F$ 是矩阵的 Frobenius 范数. 关于算法的适定性，有如下的引理.

引理 4.1. *如果 $e(HMM)<\alpha$，那么对于所有的 $V,W\in\mathcal{V}_H$，有*

$$\begin{aligned} a_H(V,V) &\geqslant (\alpha - e(HMM))\,\|\nabla V\|^2_{L^2(\Omega)}\,,\\ |a_H(V,W)| &\leqslant \left(\frac{\beta^2}{\alpha}+\alpha\right)\|\nabla V\|_{L^2(\Omega)}\,\|\nabla W\|_{L^2(\Omega)}\,. \end{aligned} \tag{4.11}$$

证明. 由有效系数矩阵 $\mathcal{A}$ 的椭圆性和 $e(\mathrm{HMM})$ 的定义，有

$$\begin{aligned} a_H(V,V) &= \int_\Omega \nabla V\cdot\mathcal{A}(\boldsymbol{x})\nabla V\,\mathrm{d}\boldsymbol{x} + \sum_{K\in\mathcal{T}_H}\int_K \nabla V\cdot(\mathcal{A}_H-\mathcal{A})(\boldsymbol{x})\nabla V\,\mathrm{d}\boldsymbol{x}\\ &\geqslant (\alpha - e(\mathrm{HMM}))\,\|\nabla V\|^2_{L^2(\Omega)}\,. \end{aligned}$$

这给出了 $a_H(\cdot,\cdot)$ 的强制性. 用类似的方法也能得到 $a_H(\cdot,\cdot)$ 的连续性. □

下文中的定理4.3证明了 $e(\mathrm{HMM})$ 的收敛性，因此当网格充分小的时候，$e(\mathrm{HMM})<\alpha$ 总是可以被满足的. 于是结合 Lax-Milgram 引理，上述引理保证了变分问题(4.5)的解的存在和唯一性.

4.3.2　误差分析

下面考虑算法的误差估计. 在本节中，令 $(\cdot,\cdot)$ 表示 $L^2(\Omega)$ 上的内积.

下面的引理基于引理4.1和 Berger, Scott and Strang 的定理 [18, 定理 1.1] 而得到的. 类似的结果也能在 [51, 定理 1.1] 和 [47, 引理 2.1] 中找到. 值得一提的是，本文的证明中显式地刻画了误差与系数 a^ε 的关系.

引理 4.2. 令 U_0 和 U_H 分别为方程 (4.1)和 (4.5) 的解. 如果 $e(HMM) < \alpha/2$，那么，

$$\|\nabla(U_0 - U_H)\|_{L^2(\Omega)} \leqslant \frac{\beta}{\alpha} \inf_{V\in\mathcal{V}_H} \|\nabla(U_0 - V)\|_{L^2(\Omega)} + \frac{2c_p}{\alpha^2}\|f\|_{H^{-1}(\Omega)}e(HMM), \quad (4.12)$$

这里 c_p 是下述 *Poincaré* 不等式 [118, 定理 1.4.2] 中的常数

$$\|V\|_{H^1(\Omega)} \leqslant c_p \|\nabla V\|_{L^2(\Omega)}, \qquad \forall V \in \mathcal{V}_H.$$

同时，如果有 $f \in L^2(\Omega)$ 使得 $U_0 \in H^2(\Omega)$，那么存在常数 C 使得

$$\|U_0 - U_H\|_{L^2(\Omega)} \leqslant C\left(H \inf_{V\in\mathcal{V}_H} \|\nabla(U_0 - V)\|_{L^2(\Omega)} + e(HMM)\right). \quad (4.13)$$

证明. 令 $\hat{U}_0 \in \mathcal{V}_H$ 满足

$$\left(\mathcal{A}\nabla\hat{U}_0, \nabla V\right) = \left(\mathcal{A}\nabla U_0, \nabla V\right), \quad \forall V \in \mathcal{V}_H.$$

根据许进超和 Zikatanov[113] 的结论有

$$\left\|\nabla(U_0 - \hat{U}_0)\right\|_{L^2(\Omega)} \leqslant \frac{\beta}{\alpha} \inf_{V\in\mathcal{V}_H} \|\nabla(U_0 - V)\|_{L^2(\Omega)}. \quad (4.14)$$

定义 $W := U_H - \hat{U}_0$. 由上述等式可以得到

$$\begin{aligned}(\mathcal{A}\nabla W, \nabla W) &= (\mathcal{A}\nabla U_H, \nabla W) - \left(\mathcal{A}\nabla\hat{U}_0, \nabla W\right) = (\mathcal{A}\nabla U_H, \nabla W) - (\mathcal{A}\nabla U_0, \nabla W)\\ &= (\mathcal{A}\nabla U_H, \nabla W) - \sum_{K\in\mathcal{T}_H}\int_K \nabla W(\boldsymbol{x})\cdot\mathcal{A}_H(\boldsymbol{x})\nabla U_H \,\mathrm{d}\boldsymbol{x}\\ &= \sum_{K\in\mathcal{T}_H}\int_K \nabla W\cdot(\mathcal{A}(\boldsymbol{x}) - \mathcal{A}_H(\boldsymbol{x}))\nabla U_H \,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

因此，

$$\left\|\nabla(U_H - \hat{U}_0)\right\|_{L^2(\Omega)} \leqslant \alpha^{-1}e(\mathrm{HMM})\|\nabla U_H\|_{L^2(\Omega)}. \quad (4.15)$$

又由 $\mathcal{A}_H$ 的椭圆性和 Poincaré 不等式，可知

$$\begin{aligned}\|\nabla U_H\|^2_{L^2(\Omega)} &\leqslant (\alpha - e(\mathrm{HMM}))^{-1}a_H(\nabla U_H, \nabla U_H)\\ &\leqslant \frac{2c_p}{\alpha}\|f\|_{H^{-1}(\Omega)}\|\nabla U_H\|_{L^2(\Omega)}.\end{aligned}$$

于是得到

$$\|\nabla U_H\|_{L^2(\Omega)} \leqslant \frac{2c_p}{\alpha} \|f\|_{H^{-1}(\Omega)}. \tag{4.16}$$

结合(4.14)、(4.15)和(4.16)，即可推出(4.12). 而(4.13)的结论是显然的. □

从(4.13)可以看出，误差估计的关键就在于 e(HMM) 的估计. 为此，本节将证明下述主要结论.

定理 4.3. *考虑 $a^\varepsilon = a(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)$，设对任意的 $i,j = 1,\ldots,d$ 都有 $a_{ij}(\boldsymbol{x},\boldsymbol{y})$ 关于 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 光滑，并且 $a(\boldsymbol{x},\boldsymbol{y})$ 关于 $\boldsymbol{y}$ 是 Y-周期的. $\boldsymbol{\chi}(\boldsymbol{x},\boldsymbol{y})$ 如 (4.3)所定义，并且满足*

$$\|\boldsymbol{\chi}\|_{H^{k'+1}(Y)} < \infty. \tag{4.17}$$

采用 m 阶重构算子. 如果假设A (详见 §3.3节) 成立，且条件 (3.24)满足，那么就有

$$e(HMM) \leqslant C\left(H^{m+1} + \delta + \frac{\varepsilon}{\delta} + \frac{h^{2k'}}{\varepsilon^{2k'}}\right). \tag{4.18}$$

首先来对 e(HMM) 作一简单分析. 由(4.9)和(4.10) 以及重构算子的定义(3.8)有

$$e(\mathrm{HMM}) \leqslant \max_{\substack{\boldsymbol{x}\in K\\ K\in\mathcal{T}_H}} \left(\left\|(\mathcal{A} - \mathcal{R}_m^K\mathcal{A})(\boldsymbol{x})\right\|_F + \left\|\mathcal{R}_m^K(\mathcal{A} - \widetilde{\mathcal{A}}_H)(\boldsymbol{x})\right\|_F\right).$$

在上式中，由定理3.4可知，对于任意的 $K \in \mathcal{T}_H$ 和 $\boldsymbol{x} \in K$，

$$\begin{aligned}
\left\|(\mathcal{A} - \mathcal{R}_m^K\mathcal{A})(\boldsymbol{x})\right\|_F &\leqslant C(1+\Lambda_m)H^{m+1},\\
\left\|\mathcal{R}_m^K(\mathcal{A} - \widetilde{\mathcal{A}}_H)(\boldsymbol{x})\right\|_F &\leqslant \Lambda_m \max_{\substack{\boldsymbol{x}_\ell\in\mathcal{I}_K\\ K\in\mathcal{T}_H}} \left\|(\mathcal{A} - \widetilde{\mathcal{A}}_H)(\boldsymbol{x}_\ell)\right\|_F.
\end{aligned}$$

于是，问题转化为在网格的每个节点上，对理论有效系数矩阵 $\mathcal{A}(\boldsymbol{x}_\ell)$ 和数值有效系数矩阵 $\widetilde{\mathcal{A}}_H(\boldsymbol{x}_\ell)$ 之间误差的估计.

为此，引入 $\widehat{\mathcal{A}}_H(\boldsymbol{x}_\ell)$，它表示对于子问题(4.7)进行精确求解并计算出的有效系数矩阵. 具体定义为 $\widehat{\mathcal{A}}_H(\boldsymbol{x}_\ell)$ 满足

$$\widehat{\mathcal{A}}_H(\boldsymbol{x}_\ell)\left\langle\nabla v^\varepsilon\right\rangle_{I_\delta} \equiv \left\langle a^\varepsilon\nabla v^\varepsilon\right\rangle_{I_\delta},$$

其中 $v^\varepsilon - V_\ell \in \mathcal{V}$ 满足

$$(a^\varepsilon\nabla v^\varepsilon, \nabla\varphi)_{L^2(I_\delta)} = 0, \quad \forall\varphi\in\mathcal{V}. \tag{4.19}$$

这里 $\mathcal{V}$ 可能为 $\mathcal{V}_D$，$\mathcal{V}_N$ 或 $\mathcal{V}_P$：

$$\begin{aligned}
\mathcal{V}_D &\equiv H_0^1\Big(I_\delta(\boldsymbol{x}_\ell)\Big),\\
\mathcal{V}_N &\equiv \Big\{\, v \in H^1\Big(I_\delta(\boldsymbol{x}_\ell)\Big) \;\big|\; \langle \nabla v\rangle_{I_\delta} = 0 \,\Big\},\\
\mathcal{V}_P &\equiv \Big\{\, v \in H^1_{\mathrm{per}}\Big(I_\delta(\boldsymbol{x}_\ell)\Big) \;\big|\; \langle v\rangle_{I_\delta} = 0 \,\Big\}.
\end{aligned}$$

首先介绍一个引理，该引理在网格节点上对 $\mathcal{A}$ 和 $\widehat{\mathcal{A}}_H$ 之间的误差进行了估计.

引理 4.4. [47, 定理 3.4] 如果 $a^\varepsilon = a(\boldsymbol{x}, \boldsymbol{x}/\varepsilon)$ 满足 $a(\boldsymbol{x}, \boldsymbol{y}) \in C^{0,1}(\Omega; L^\infty(Y))$ 并且 $a(\boldsymbol{x}, \boldsymbol{y})$ 关于 $\boldsymbol{y}$ 是 Y-周期函数，那么在每个网格节点 $\boldsymbol{x}_\ell$ 上有

$$\Big\|(\mathcal{A} - \widehat{\mathcal{A}}_H)(\boldsymbol{x}_\ell)\Big\|_F \leqslant C\frac{\beta^4}{\alpha^3}\left(\delta + \frac{\varepsilon}{\delta}\right). \tag{4.20}$$

接下来考虑 $\widetilde{\mathcal{A}}_H$ 和 $\widehat{\mathcal{A}}_H$ 之间的误差.

引理 4.5. 在每个网格节点 $\boldsymbol{x}_\ell$ 上有

$$\begin{aligned}
\Big\|(\widetilde{\mathcal{A}}_H - \widehat{\mathcal{A}}_H)(\boldsymbol{x}_\ell)\Big\|_F \leqslant \frac{\beta^3}{\alpha^2} &\left(\sum_{i=1}^d \inf_{v\in\mathcal{V}_h} \|\nabla(v_i^\varepsilon - x_i - v)\|^2_{L^2(I_\delta(\boldsymbol{x}_\ell))}\right)^{1/2}\\
&\times\left(\sum_{i=1}^d \inf_{v\in\mathcal{V}_h} \|\nabla(\widetilde{v}_i^\varepsilon - x_i - v)\|^2_{L^2(I_\delta(\boldsymbol{x}_\ell))}\right)^{1/2},
\end{aligned} \tag{4.21}$$

其中 v_i^ε 定义为当 $V_\ell = x_i$ 时方程 (4.19)的解，$\widetilde{v}_i^\varepsilon$ 定义为 $V_\ell = x_i$ 并且把 a^ε 替换为其转置 $(a^\varepsilon)^{\mathrm{T}}$ 时方程 (4.19)解.

证明. 由 $\widehat{\mathcal{A}}_H$ 的定义和(4.19)，对每个点 $\boldsymbol{x}_\ell$ 有

$$\begin{aligned}
\widehat{\mathcal{A}}_H(\boldsymbol{x}_\ell)_{ij} &= \langle \nabla v_i^\varepsilon\rangle_{I_\delta(\boldsymbol{x}_\ell)} \cdot \widehat{\mathcal{A}}_H(\boldsymbol{x}_\ell)\left\langle \nabla v_j^\varepsilon\right\rangle_{I_\delta(\boldsymbol{x}_\ell)} = \nabla x_i \cdot \left\langle a^\varepsilon \nabla v_j^\varepsilon\right\rangle_{I_\delta(\boldsymbol{x}_\ell)}\\
&= \left\langle \nabla x_i \cdot a^\varepsilon \nabla v_j^\varepsilon\right\rangle_{I_\delta(\boldsymbol{x}_\ell)} = \left\langle \nabla \widetilde{v}_i^\varepsilon \cdot a^\varepsilon \nabla v_j^\varepsilon\right\rangle_{I_\delta(\boldsymbol{x}_\ell)}.
\end{aligned}$$

由同样的方式可以得到

$$\widetilde{\mathcal{A}}_H(\boldsymbol{x}_\ell)_{ij} = \left\langle \nabla \widetilde{v}_{i,h}^\varepsilon \cdot a^\varepsilon \nabla v_{j,h}^\varepsilon\right\rangle_{I_\delta(\boldsymbol{x}_\ell)},$$

其中 $\widetilde{v}_{i,h}^\varepsilon$ 是当 a^ε 替换为其转置时方程 (4.8)的解. 结合上述两个不等式就有

$$\left(\widehat{\mathcal{A}}_H - \widetilde{\mathcal{A}}_H\right)_{ij}(\boldsymbol{x}_\ell) = \left\langle \nabla(\widetilde{v}_i^\varepsilon - \widetilde{v}_{i,h}^\varepsilon)\cdot a^\varepsilon \nabla v_j^\varepsilon\right\rangle_{I_\delta(\boldsymbol{x}_\ell)} + \left\langle \nabla \widetilde{v}_{i,h}^\varepsilon \cdot a^\varepsilon \nabla(v_j^\varepsilon - v_{j,h}^\varepsilon)\right\rangle_{I_\delta(\boldsymbol{x}_\ell)}.$$

由于 $\widetilde{v}_i^\varepsilon - \widetilde{v}_{i,h}^\varepsilon \in \mathcal{V}$，因此上述方程等号右端的第一项为 0. 对于等号右端的第二项则有

$$\begin{aligned}\left(\widehat{\mathcal{A}}_H - \widetilde{\mathcal{A}}_H\right)_{ij}(\boldsymbol{x}_\ell) &= \left\langle \nabla\widetilde{v}_{i,h}^\varepsilon \cdot a^\varepsilon \nabla(v_j^\varepsilon - v_{j,h}^\varepsilon)\right\rangle_{I_\delta(\boldsymbol{x}_\ell)} \\ &= \left\langle \nabla(v_j^\varepsilon - v_{j,h}^\varepsilon)\cdot (a^\varepsilon)^{\mathrm{T}} \nabla\widetilde{v}_{i,h}^\varepsilon\right\rangle_{I_\delta(\boldsymbol{x}_\ell)} \\ &= \left\langle \nabla(v_j^\varepsilon - v_{j,h}^\varepsilon)\cdot (a^\varepsilon)^{\mathrm{T}} \nabla(\widetilde{v}_{i,h}^\varepsilon - \widetilde{v}_i^\varepsilon)\right\rangle_{I_\delta(\boldsymbol{x}_\ell)}. \end{aligned} \tag{4.22}$$

同时，又因为

$$\left\|\nabla(v_i^\varepsilon - v_{i,h}^\varepsilon)\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))} \leqslant \frac{\beta}{\alpha}\inf_{v\in\mathcal{V}_h}\left\|\nabla(v_i^\varepsilon - x_i - v)\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}, \tag{4.23}$$

和

$$\left\|\nabla(\widetilde{v}_j^\varepsilon - \widetilde{v}_{j,h}^\varepsilon)\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))} \leqslant \frac{\beta}{\alpha}\inf_{v\in\mathcal{V}_h}\left\|\nabla(\widetilde{v}_j^\varepsilon - x_j - v)\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}.$$

于是从(4.22)可以得到 (4.21). □

如果 a^ε 是对称矩阵，那么 (4.21)将变成 [1]

$$\left\|(\widetilde{\mathcal{A}}_H - \widehat{\mathcal{A}}_H)(\boldsymbol{x}_\ell)\right\|_F \leqslant \frac{\beta^3}{\alpha^2}\sum_{i=1}^d \inf_{v\in\mathcal{V}_h}\left\|\nabla(v_i^\varepsilon - x_i - v)\right\|^2_{L^2(I_\delta(\boldsymbol{x}_\ell))}. \tag{4.24}$$

到现在为止，除了要求 $a^\varepsilon \in \mathcal{M}(\alpha,\beta,D)$ 之外，尚未对系数 a^ε 作出任何的假设. 对于 $i = 1,\ldots,d$，令 Πv_i^ε 为 v_i^ε 的 Lagrange 插值函数. 在 (4.23)中令 $v = \Pi(v_i^\varepsilon - x_i)$，于是有

$$v_i^\varepsilon - x_i - v = v_i^\varepsilon - x_i - \Pi(v_i^\varepsilon - x_i) = v_i^\varepsilon - \Pi v_i^\varepsilon. \tag{4.25}$$

假如 $\|v_i^\varepsilon\|_{H^{k'+1}(I_\delta(\boldsymbol{x}_\ell))}$ 是有界的，那么就有如下的估计结果

$$\left\|\nabla(v_i^\varepsilon - x_i - v)\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))} = \left\|\nabla(v_i^\varepsilon - \Pi v_i^\varepsilon)\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))} \leqslant Ch^{k'}\left\|v_i^\varepsilon\right\|_{H^{k'+1}(I_\delta(\boldsymbol{x}_\ell))}.$$

然而，这个正则化结果对于 $k' > 1$ 可能不成立 [73]. 此外，即使这个结果成立，也必须搞清楚 $\|v_i^\varepsilon\|_{H^{k'+1}(I_\delta(\boldsymbol{x}_\ell))}$ 对参数 ε 和 δ 的依赖关系，而这一点是比较困难的.

[1] 如果 a^ε 是对称的，那么通过能量范数 $\|v\|_a \equiv \left(\int_D \nabla v\cdot a^\varepsilon\nabla v\,\mathrm{d}\boldsymbol{x}\right)^{1/2}$，可以将 (4.21)中的 β^3/α^2 替换为 β，见 [32, 注 2.4.1].

如果 $k'=1$，并且假设

$$|\nabla a^\varepsilon(x)| \leqslant \frac{C}{\varepsilon} \quad a.e.,\ x\in\Omega,$$

杜锐和明平兵 [48] 证明了

$$\|\nabla v_i^\varepsilon\|_{H^1(I_\delta(\boldsymbol{x}_\ell))} \leqslant \frac{C}{\varepsilon}.$$

由此，可以直接得出

$$\inf_{v\in\mathcal{V}_h} \|\nabla(v_i^\varepsilon - x_i - v)\|_{L^2(I_\delta(\boldsymbol{x}_\ell))} \leqslant C\frac{h}{\varepsilon}, \tag{4.26}$$

这里 C 是独立于 ε、δ 和 h 的常数.

如果 a^ε 是局部周期的，即 $a^\varepsilon(\boldsymbol{x}) = a(\boldsymbol{x},\boldsymbol{x}/\varepsilon)$ 并且 $a(\boldsymbol{x},\boldsymbol{y})$ 关于 $\boldsymbol{y}$ 是 Y-周期函数，那么情况就会稍有不同. 如果 $k'=1$，子问题服从周期边条件且 $\delta=\varepsilon$，那么 Abdulle [1] 在 $\|\boldsymbol{\chi}\|_{W^{2,\infty}(Y)}$ 有界的假设下证明了 (4.26). 对于满足 $\delta/\varepsilon\in\mathbb{N}$ 的周期子问题，可以采用 [81, 第三章] 中的方法来得到

$$\|v_i^\varepsilon\|_{H^{k'+1}(I_\delta(\boldsymbol{x}_\ell))} \leqslant C\varepsilon^{-k'}.$$

于是对于满足 $\delta/\varepsilon\in\mathbb{N}$ 和 $k'>1$ 的周期子问题有

$$\inf_{v\in\mathcal{V}_h} \|\nabla(v_i^\varepsilon - x_i - v)\|_{L^2(I_\delta(\boldsymbol{x}_\ell))} \leqslant C\left(h/\varepsilon\right)^{k'}.$$

在 [47, 推论 3.10] 用不同的论证方法得到了类似的估计. 然而，尚不清楚上述正则化结果对 $\delta/\varepsilon\notin\mathbb{N}$ 时是否成立.

下文中将给出服从 Dirichlet、Neumann 和周期边条件的子问题的离散误差估计. 在接下来的证明中不再采用类似式 (4.25)中的 Lagrange 插值函数，而是基于均匀化理论和下面的引理来构造一个特殊的插值函数.

引理 4.6. [47, 引理 3.2] 令 v^ε 为方程 (4.19)的解，并且定义

$$\widehat{V}^\varepsilon \equiv V_\ell + \varepsilon(\boldsymbol{\chi}\cdot\nabla)V_\ell,$$

这里 $\boldsymbol{\chi}(\boldsymbol{x},\boldsymbol{y}) = \{\chi^j(\boldsymbol{x},\boldsymbol{y})\}_{j=1}^d$ 的定义见(4.3). 那么存在一个常数 C 使得

$$\left\|\nabla(v^\varepsilon - \widehat{V}^\varepsilon)\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))} \leqslant C\frac{\beta^2}{\alpha^2}\left(\frac{\varepsilon}{\delta}\right)^{1/2}\|\nabla V_\ell\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}. \tag{4.27}$$

上述结果的一个直接推论就是

推论 4.7. *定义*

$$\widetilde{V}^{\varepsilon} = V_\ell + \varepsilon(\varrho^{\varepsilon}\boldsymbol{\chi}\cdot\nabla)V_\ell,$$

这里 $\varrho^{\varepsilon}\in C_0^\infty(I_\delta)$ *是一个截断函数，并且满足* $|\nabla\varrho^{\varepsilon}|\leqslant C/\varepsilon$，*且*

$$\varrho^{\varepsilon}(\boldsymbol{x}) = \begin{cases} 1, & \text{如果 } \operatorname{dist}(\boldsymbol{x},\partial I_\delta(\boldsymbol{x}_\ell))\geqslant 2\varepsilon,\\ 0, & \text{如果 } \operatorname{dist}(\boldsymbol{x},\partial I_\delta(\boldsymbol{x}_\ell))\leqslant \varepsilon, \end{cases}$$

那么就有

$$\left\|\nabla(v^{\varepsilon}-\widetilde{V}^{\varepsilon})\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}\leqslant C\frac{\beta^2}{\alpha^2}\left(\frac{\varepsilon}{\delta}\right)^{1/2}\|\nabla V_\ell\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}, \tag{4.28}$$

其中 $\operatorname{dist}(\boldsymbol{x},\partial I_\delta(\boldsymbol{x}_\ell))$ *表示点* $\boldsymbol{x}$ *到* $\partial I_\delta(\boldsymbol{x}_\ell)$ *的距离.*

证明. 注意到 $\widehat{V}^{\varepsilon}-\widetilde{V}^{\varepsilon}=(\widehat{V}^{\varepsilon}-V_\ell)(1-\varrho^{\varepsilon})$，由 [47, 引理 3.1]，得到

$$\left\|\nabla[(\widehat{V}^{\varepsilon}-V_\ell)(1-\varrho^{\varepsilon})]\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}\leqslant C\frac{\beta}{\alpha}\left(\frac{\varepsilon}{\delta}\right)^{1/2}\|\nabla V_\ell\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}.$$

又因为 (4.27)，于是有

$$\begin{aligned}\left\|\nabla(v^{\varepsilon}-\widetilde{V}^{\varepsilon})\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))} &\leqslant \left\|\nabla(v^{\varepsilon}-\widehat{V}^{\varepsilon})\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}+\left\|\nabla(\widehat{V}^{\varepsilon}-\widetilde{V}^{\varepsilon})\right\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}\\ &\leqslant C\frac{\beta^2}{\alpha^2}\left(\frac{\varepsilon}{\delta}\right)^{1/2}\|\nabla V_\ell\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}.\end{aligned}$$

□

基于上述结果，现在可以估计 v^{ε} 和 v_h^{ε} 之间的误差.

引理 4.8. *令* v^{ε} *和* v_h^{ε} *分别为问题* (4.19)*和* (4.7)*的解. 如果*(4.17)*成立，那么*

$$\|\nabla(v^{\varepsilon}-v_h^{\varepsilon})\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}\leqslant C\left(\left(\frac{\varepsilon}{\delta}\right)^{1/2}+\frac{h^{k'}}{\varepsilon^{k'}}\right)\|\nabla V_\ell\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}. \tag{4.29}$$

证明. 类似 (4.23)，可知

$$\|\nabla(v^{\varepsilon}-v_h^{\varepsilon})\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}\leqslant\frac{\beta}{\alpha}\inf_{v\in\mathcal{V}_h}\|\nabla(v^{\varepsilon}-V_\ell-v)\|_{L^2(I_\delta(\boldsymbol{x}_\ell))}. \tag{4.30}$$

注意到 $\Pi\widetilde{V}^{\varepsilon}=V_\ell+\Pi(\widetilde{V}^{\varepsilon}-V_\ell)$，并且由 $\widetilde{V}^{\varepsilon}-V_\ell\in H_0^1(I_\delta(\boldsymbol{x}_\ell))$ 可知 $\Pi(\widetilde{V}^{\varepsilon}-V_\ell)\in H_0^1(I_\delta(\boldsymbol{x}_\ell))$，于是得到 $\Pi\widetilde{V}^{\varepsilon}-V_\ell\in\mathcal{V}_{D,h}$.

对于 Neumann 子问题，由分部积分可知

$$\left\langle \nabla(\Pi\widetilde{V}^{\varepsilon}-V_{\ell})\right\rangle_{I_{\delta}(\boldsymbol{x}_{\ell})}=\left\langle \nabla[\Pi(\widetilde{V}^{\varepsilon}-V_{\ell})]\right\rangle_{I_{\delta}(\boldsymbol{x}_{\ell})}=\mathbf{0},$$

于是就推出了 $\Pi\widetilde{V}^{\varepsilon}-V_{\ell}\in\mathcal{V}_{N,h}$.

对于周期子问题，用 $\Pi\widetilde{V}^{\varepsilon}-V_{\ell}+c$ 代替 $\Pi\widetilde{V}^{\varepsilon}-V_{\ell}$，这里 c 是一个合适的常数，满足 $\left\langle \Pi\widetilde{V}^{\varepsilon}-V_{\ell}+c\right\rangle_{I_{\delta}(\boldsymbol{x}_{\ell})}=0$. 因此有 $\Pi\widetilde{V}^{\varepsilon}-V_{\ell}+c\in\mathcal{V}_{P,h}$.

在 (4.30)中令 $v=\Pi\widetilde{V}^{\varepsilon}-V_{\ell}$ 或 $\Pi\widetilde{V}-V_{\ell}+c$，并注意到

$$\nabla(v^{\varepsilon}-V_{\ell}-v)=\nabla(v^{\varepsilon}-V_{\ell}-\Pi\widetilde{V}^{\varepsilon}+V_{\ell})=\nabla(v^{\varepsilon}-\Pi\widetilde{V}^{\varepsilon}),$$

且

$$\nabla(v^{\varepsilon}-V_{\ell}-v)=\nabla(v^{\varepsilon}-V_{\ell}-\Pi\widetilde{V}^{\varepsilon}+V_{\ell}-c)=\nabla(v^{\varepsilon}-\Pi\widetilde{V}^{\varepsilon}),$$

这就得到了

$$\begin{aligned}\|\nabla(v^{\varepsilon}-v_{h}^{\varepsilon})\|_{L^{2}(I_{\delta}(\boldsymbol{x}_{\ell}))}&\leqslant\frac{\beta}{\alpha}\left\|\nabla(v^{\varepsilon}-\Pi\widetilde{V}^{\varepsilon})\right\|_{L^{2}(I_{\delta}(\boldsymbol{x}_{\ell}))}\\&\leqslant\frac{\beta}{\alpha}\left(\left\|\nabla(v^{\varepsilon}-\widetilde{V}^{\varepsilon})\right\|_{L^{2}(I_{\delta}(\boldsymbol{x}_{\ell}))}+\left\|\nabla(\widetilde{V}^{\varepsilon}-\Pi\widetilde{V}^{\varepsilon})\right\|_{L^{2}(I_{\delta}(\boldsymbol{x}_{\ell}))}\right).\end{aligned}$$

由经典的插值估计可知

$$\left\|\nabla(\widetilde{V}^{\varepsilon}-\Pi\widetilde{V}^{\varepsilon})\right\|_{L^{2}(I_{\delta}(\boldsymbol{x}_{\ell}))}\leqslant Ch^{k'}\left\|\nabla^{k'+1}\widetilde{V}^{\varepsilon}\right\|_{L^{2}(I_{\delta}(\boldsymbol{x}_{\ell}))}.$$

直接计算得出

$$\begin{aligned}\left\|\nabla^{k'+1}\widetilde{V}^{\varepsilon}\right\|_{L^{2}(I_{\delta}(\boldsymbol{x}_{\ell}))}&\leqslant C\varepsilon^{-k'}\delta^{d/2}\left\|\nabla_{\boldsymbol{y}}^{k'+1}\boldsymbol{\chi}\right\|_{L^{2}(Y)}|\nabla V_{\ell}|\\&=C\varepsilon^{-k'}\left\|\nabla_{\boldsymbol{y}}^{k'+1}\boldsymbol{\chi}\right\|_{L^{2}(Y)}\|\nabla V_{\ell}\|_{L^{2}(I_{\delta}(\boldsymbol{x}_{\ell}))}.\end{aligned}$$

综合上述两个不等式，可以得到

$$\left\|\nabla(\widetilde{V}^{\varepsilon}-\Pi\widetilde{V}^{\varepsilon})\right\|_{L^{2}(I_{\delta}(\boldsymbol{x}_{\ell}))}\leqslant C(h/\varepsilon)^{k'}\left\|\nabla_{\boldsymbol{y}}^{k'+1}\boldsymbol{\chi}\right\|_{L^{2}(Y)}\|\nabla V_{\ell}\|_{L^{2}(I_{\delta}(\boldsymbol{x}_{\ell}))}.$$

同时因为有 (4.28)和 (4.17)，于是就推出了 (4.29). □

由 (4.29)和 (4.21)，可以得到 $\widetilde{\mathcal{A}}_{H}$ 和 $\widehat{\mathcal{A}}_{H}$ 之间的估计.

引理 4.9. *假设引理 4.8的条件成立，那么就有*

$$\left\|(\widetilde{\mathcal{A}}_H - \widehat{\mathcal{A}}_H)(\boldsymbol{x}_\ell)\right\|_F \leqslant C\left(\frac{\varepsilon}{\delta} + \frac{h^{2k'}}{\varepsilon^{2k'}}\right). \tag{4.31}$$

综合引理4.4和引理4.9，可以对定理4.3给出如下证明.

定理4.3的证明. 对于任意的 $K \in \mathcal{T}_H$ 和 $\boldsymbol{x} \in K$，由 (3.14)可得

$$\left\|(\mathcal{A} - \mathcal{R}_m^K\mathcal{A})(\boldsymbol{x})\right\|_F \leqslant C\,(1+\Lambda_m)\,H^{m+1},$$

这里 C 依赖于 $\|\mathcal{A}\|_{W^{m+1,\infty}(S(K))}$. 然后由 (3.13)，可知

$$\begin{aligned} e(\mathrm{HMM}) &\leqslant \max_{\substack{\boldsymbol{x}\in K\\ K\in\mathcal{T}_H}} \left(\left\|(\mathcal{A} - \mathcal{R}_m^K\mathcal{A})(\boldsymbol{x})\right\|_F + \left\|\mathcal{R}_m^K(\mathcal{A} - \widetilde{\mathcal{A}}_H)(\boldsymbol{x})\right\|_F\right) \\ &\leqslant C\,(1+\Lambda_m)\,H^{m+1} + \Lambda_m \max_{\substack{\boldsymbol{x}_\ell\in\mathcal{I}_K\\ K\in\mathcal{T}_H}} \left\|(\mathcal{A} - \widetilde{\mathcal{A}}_H)(\boldsymbol{x}_\ell)\right\|_F. \end{aligned}$$

将上述估计和 (4.20) 以及 (4.31)结合起来，就能得到 (4.18). □

综合引理 4.2和定理4.3，就能对本章提出的基于有效系数重构的高阶异质多尺度方法给出误差估计的表达式.

推论 4.10. *在定理 4.3的条件下有*

$$\begin{aligned} \|\nabla(U_0 - U_H)\|_{L^2(\Omega)} &\leqslant C\left(H^k + H^{m+1} + \delta + \frac{\varepsilon}{\delta} + \frac{h^{2k'}}{\varepsilon^{2k'}}\right), \\ \|U_0 - U_H\|_{L^2(\Omega)} &\leqslant C\left(H^{k+1} + H^{m+1} + \delta + \frac{\varepsilon}{\delta} + \frac{h^{2k'}}{\varepsilon^{2k'}}\right). \end{aligned} \tag{4.32}$$

4.4　数值算例

本节将通过一系列的数值算例来对算法的收敛性和效率进行测试. 本节中所有给出计算时间的例子都是在一台 CPU 主频为 2.50GHz 的 IBM 笔记本上进行的.

例 4.1　第一个例子中将讨论以下内容：重构带来的误差、两种重构方式的比较、与 P_2 边方法的比较以及宏观 P_3 元求解器. 在这个例子中将用以下数据对

问题 (1.8)进行数值实验

$$\begin{cases} a(\boldsymbol{x}, \boldsymbol{x}/\epsilon) = \dfrac{(R_1 + R_2\sin(2\pi x_1))(R_1 + R_2\cos(2\pi x_2))}{(R_1 + R_2\sin(2\pi x_1/\varepsilon))(R_1 + R_2\sin(2\pi x_2/\varepsilon))} I, \\ \qquad f(\boldsymbol{x}) = 1, \\ \qquad u(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \partial\Omega, \end{cases} \tag{4.33}$$

这里 $\varepsilon = 10^{-6}$，$\Omega = (0,1)\times(0,1)$ 且 I 是 2×2 的单位矩阵. 这个问题在 [48] 已经被讨论过.

通过计算可以给出上述函数的有效系数的精确表达式为

$$\mathcal{A}(x_1, x_2) = \frac{(R_1 + R_2\sin(2\pi x_1))(R_1 + R_2\cos(2\pi x_2))}{R_1\sqrt{R_1^2 - R_2^2}} I. \tag{4.34}$$

在数值实验中，取 $R_1 = 2.5$ 和 $R_2 = 1.5$，并采用标准的 Gauss 积分公式来计算 (4.5)中的刚度矩阵. 在每个数值积分点上，用 (4.9)或 (4.10)来计算出 $\mathcal{A}_H$. 为了得到网格节点上的 $\widetilde{\mathcal{A}}_H$ 的每个元素，在 (4.8)中将边界数据 V_ℓ 取为 $\boldsymbol{e}_i \cdot \boldsymbol{x}$，其中 $\{\boldsymbol{e}_i\}_{i=1}^d$ 为标准基. 对区域 Ω 用 *EasyMesh*[2] 进行三角形剖分，网格尺寸为 $H = 1/N$. 单胞 I_δ 也由 *EasyMesh* 进行三角形剖分，网格尺寸为 $h = \delta/M$. 网格如图4.2所示. 为了方便后面的表述，现将误差估计式 (4.32)改写为关于 N 和 M 的表达式.

$$\begin{aligned} \|\nabla(U_0 - U_H)\|_{L^2(\Omega)} &\leqslant C\left(N^{-k} + N^{-m-1} + \delta + \frac{\varepsilon}{\delta} + \frac{\delta^{2k'}}{(M\varepsilon)^{2k'}}\right), \\ \|U_0 - U_H\|_{L^2(\Omega)} &\leqslant C\left(N^{-k-1} + N^{-m-1} + \delta + \frac{\varepsilon}{\delta} + \frac{\delta^{2k'}}{(M\varepsilon)^{2k'}}\right). \end{aligned} \tag{4.35}$$

为了在单元 K 上重构出 $\mathcal{A}_H$，首先要按照(3.2)所描述的方式构造单元模板 $S(K)$. 采样点集 $\mathcal{I}_K$ 取为 $S(K)$ 上的所有网格节点. 假设A的一个直接推论就是

$$\#\mathcal{I}_K \geqslant \binom{m+d}{d}. \tag{4.36}$$

当 $t = 1$，即单元模板只向外扩展一层时，如果 m 较大或者单元靠近区域边界，采样点的个数 $\#\mathcal{I}_K$ 有可能会小于 $\binom{m+d}{d}$. 一个很自然的解决方法就是让 $S_t(K)$ 扩展更多层. 举例来说，三阶多项式重构需要至少十个采样点. 对于图 3.1(b)中的阴影单元，$\#\mathcal{I}_K = 7$，显然是不够的. 为了能得到足够多的采样点，如图 3.1(c)所

[2] 见 *http://www-dinma.univ.trieste.it/nirftc/research/easymesh/*

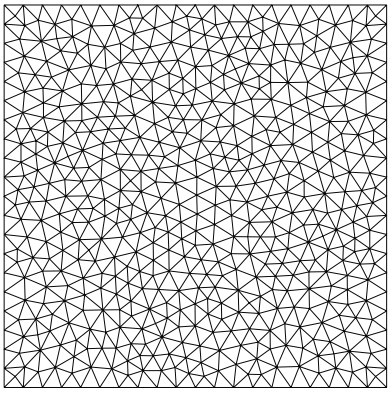


图 *4.2: 例4.1：拟一致三角形网格.*

示，需要采用了一个更大的单元模板，该模板中采样点的个数为 $\#\mathcal{I}_K = 13$. 事实上，条件 (3.24)中所要求的模板单元要更大，因为对 m 阶多项式来说，需要有 $\#\mathcal{I}_K \simeq O(m^{2d})$.

为了检验重构算子的收敛阶，先在网格节点上直接用 (4.34) 式给出有效系数矩阵，并以此作为最小二乘问题的输入数据. 这样一来就相当于对子问题进行精确求解，于是误差估计 (4.32)变为

$$\begin{aligned} \|\nabla(U_0 - U_H)\|_{L^2(\Omega)} &\leqslant C\left(N^{-k} + N^{-m-1}\right), \\ \|U_0 - U_H\|_{L^2(\Omega)} &\leqslant C\left(N^{-k-1} + N^{-m-1}\right). \end{aligned} \tag{4.37}$$

基于 (4.37)可以知道，对于 H^1 或 L^2 误差，宏观求解器中需要采用至少 $k-1$ 或 k 阶多项式重构. 表 4.1 清楚地展示了当 $m = k = 1$ 时，误差估计 (4.37)是最优的. 表 4.2和表4.3中的结果说明了二阶重构算子对于宏观求解器获得最优的 L^2 误差估计是必要的. 表 4.1-4.3 中的结果是基于带约束的最小二乘重构而得到的.

N	L^2 误差	阶	H^1 误差	阶
4	1.687E-01		3.588E-01	
8	4.922E-02	1.78	1.759E-01	1.03
16	1.298E-02	1.92	8.787E-02	1.00
32	3.272E-03	1.99	4.367E-02	1.01

表 *4.1: 例4.1：P_1 元，1 阶多项式重构，收敛阶满阶.*

对于带约束和不带约束的重构，表 4.4-4.5中展示了相应的数值结果. 正如所预期的那样，两种方法都能达到满阶收敛，但带约束的最小二乘重构在数值精度上要优于不带约束的结果.

接下来将通过求解单胞问题来计算出网格顶点上的有效系数. 由(4.35)可知，

N	L^2 误差	阶	H^1 误差	阶
4	7.096E-02		1.251E-01	
8	2.099E-02	1.76	3.973E-02	1.65
16	5.425E-03	1.95	1.042E-02	1.93
32	1.367E-03	1.99	2.637E-03	1.98

表 *4.2:* 例*4.1*：P_2 元，1 阶多项式重构，收敛阶不满阶.

N	L^2 误差	阶	H^1 误差	阶
4	2.010E-02		7.662E-02	
8	2.641E-03	2.93	2.269E-02	1.76
16	2.529E-04	3.38	5.797E-03	1.97
32	2.816E-05	3.17	1.469E-03	1.98

表 *4.3:* 例*4.1*：P_2 元，2 阶多项式重构，收敛阶满阶.

N	L^2 误差	阶	H^1 误差	阶
4	6.297E-02		1.610E-01	
8	1.030E-02	2.61	4.808E-02	1.74
16	8.495E-04	3.55	8.222E-03	2.55
32	6.405E-05	3.73	1.648E-03	2.32

表 *4.4:* 例*4.1*：P_2 元，不带约束的 2 阶多项式重构.

N	L^2 误差	阶	H^1 误差	阶
4	2.010E-02		7.662E-02	
8	2.641E-03	2.93	2.269E-02	1.76
16	2.529E-04	3.38	5.797E-03	1.97
32	2.816E-05	3.17	1.469E-03	1.98

表 *4.5:* 例*4.1*：P_2 元，带约束的 2 阶多项式重构.

除了宏观和微观的离散误差以及重构误差，还有一个误差项 $\delta + \varepsilon/\delta$，它表示所谓的共振误差. 现在已经有大量的数值例子来解释 HMM-FEM 中共振误差的影响，详见 [87, 114, 48] 等. 在这里将不再重复他们的工作. 所以下面的例子将求解周期边条件的子问题并要求 $\delta = \varepsilon$. 于是由(4.35)可知，如果宏观求解器是 P_2 元，微观求解器是 P_1 元且采用二阶带约束最小二乘重构，那么误差估计将变成

$$\|\nabla(U_0 - U_H)\|_{L^2(\Omega)} \leqslant C(N^{-2} + \varepsilon + M^{-2}), \tag{4.38}$$

$$\|U_0 - U_H\|_{L^2(\Omega)} \leqslant C(N^{-3} + \varepsilon + M^{-2}). \tag{4.39}$$

将 ε 取为一个小量，比如 $\varepsilon = 10^{-6}$，那么在上面两式的右端就可以将其忽略. 于是为了能够与宏观求解器的误差收敛阶相匹配，如 [48] 所述，需要按照如下策略来确定微观求解器的尺寸.

$$M = \begin{cases} N, & H^1\text{误差}, \\ N^{3/2}, & L^2\text{误差}. \end{cases}$$

在测试中统一取 $M = N^{3/2}$.

在 [48] 中对于二次元的宏观求解器提出了一种特殊的数值积分格式，其数值积分点全部位于单元的边界中点，本章中将这一方法简称为 P_2 边. 表 4.6 和 4.7分别展示了用本章提出的方法以和 P_2 边方法计算得到的数值结果. 可以看出，采用新方法的 CPU 时间基本上是采用 P_2 边方法的三分之一，这主要是由于 P_2 边方法的运行时间大体上和网格中边的个数成正比，而本章中的方法的运行时间则基本和网格节点的个数成正比. 在网格尺寸趋于 0 时，二维单纯形网格中节点的个数将趋于边的个数的三分之一. 此外，注意到表 4.6中前两行中的误差要大于表4.7中对应的误差，这主要是由于网格较小时新方法会带来较大的重构误差. 当网格加密后，该项误差随之减小，从而在两种方法中得到相近的误差精度.

N	M	CPU 时间 (秒)	L^2 误差	H^1 误差
4	8	0.31	2.049E-02	7.719E-02
8	32	11.85	2.758E-03	2.271E-02
16	64	165.88	2.298E-04	5.800E-03

表 *4.6: 例4.1*：基于重构的高阶算法的数值结果.

上一个例子所采用的宏观-微观匹配策略为 $M = N^{3/2}$，这使得当宏观网格加密时，微观网格的尺寸将迅速变小，从而总的时间代价会大大增加. 如果采用更高阶的宏观求解器，情况将变得更糟. 比如，如果采用三次元宏观求解器，那么误差

N	M	CPU 时间 (秒)	L^2 误差	H^1 误差
4	8	0.46	1.798E-02	7.551E-02
8	32	27.75	1.944E-03	2.310E-02
16	64	445.65	2.585E-04	5.871E-03

表 *4.7:* 例*4.1*：P_2 边方法的数值结果.

估计将会变为

$$\|\nabla(U_0 - U_H)\|_{L^2(\Omega)} \leqslant C(N^{-3} + \varepsilon + M^{-2k'}),$$
$$\|U_0 - U_H\|_{L^2(\Omega)} \leqslant C(N^{-4} + \varepsilon + M^{-2k'}).$$

如果 $k' = 1$，那么 $M = N^2$. 这将导致求解子问题的时间代价随宏观网格的加密而大幅增长. 一个解决该问题的简单方法就是用高阶的微观求解器. 比如，如果 $k' = 2$，那么就有 $M = N$. 由表 4.8和表 4.9 的结果可以看出，至少对于具有光滑微结构的子问题 [5] 来说，高阶微观求解器的优势非常明显.

N	M	CPU 时间 (秒)	L^2 误差	阶	H^1 误差	阶
4	16	1.08	2.158E-02		4.791E-02	
8	64	49.72	2.866E-03	2.91	6.546E-03	2.87
16	256	> 3000	2.175E-04	3.72	5.863E-04	3.48

表 *4.8:* 例*4.1*：宏观 P_3 元，微观 P_1 元，3 阶多项式重构.

N	M	CPU 时间 (秒)	L^2 误差	阶	H^1 误差	阶
4	4	0.28	2.014E-02		4.725E-02	
8	8	2.99	2.777E-03	2.86	6.505E-03	2.86
16	16	46.03	2.121E-04	3.71	5.840E-04	3.48
32	32	803.89	1.334E-05	3.99	7.890E-05	2.89

表 *4.9:* 例*4.1*：宏观 P_3 元，微观 P_2 元，3 阶多项式重构.

例 4.2 下面给出一个三维情形下的例子. 在区域 $[0,1]^3$ 上考虑方程(1.8)，采用的系数矩阵为

$$a(\boldsymbol{x}, \boldsymbol{x}/\varepsilon) = \begin{pmatrix} \dfrac{500}{5 + 3.5\sin(2\pi x_1/\varepsilon)} & 35 & 0 \\ 35 & \dfrac{500}{5 + 3.5\cos(2\pi x_3/\varepsilon)} & 0 \\ 0 & 0 & 200 \end{pmatrix},$$

其中 $\varepsilon = 10^{-6}$. 通过计算，可以知道上述矩阵的有效系数的精确表达式为

$$\mathcal{A}(\boldsymbol{x}) = \begin{pmatrix} 100 & 35 & 0 \\ 35 & 140 & 0 \\ 0 & 0 & 200 \end{pmatrix}.$$

计算所用的网格剖分为拟一致的四面体网格 (图4.3)，宏观求解器取为 P_2 元，微观求解器取为 P_2 元. 此时误差估计式为

$$\begin{aligned} \|\nabla(U_0 - U_H)\|_{L^2(\Omega)} &\leqslant C(N^{-2} + N^{-m-1} + \varepsilon + M^{-4}), \\ \|U_0 - U_H\|_{L^2(\Omega)} &\leqslant C(N^{-3} + N^{-m-1} + \varepsilon + M^{-4}). \end{aligned}$$

于是如前文所述，为了使全局的误差阶达到最优，令多项式重构的阶数 $m = 2$，并且将宏观网格和微观网格的尺寸匹配规则取为 $M = N^{3/4}$. 表4.10中列出了计算结果. 可以看出，在三维情形下，随着网格尺度的减小，L^2 误差和 H^1 误差的收敛阶已经在逐渐趋近于最优结果. 更复杂的三维问题以及更细致的三维网格的测试将是今后工作的方向.

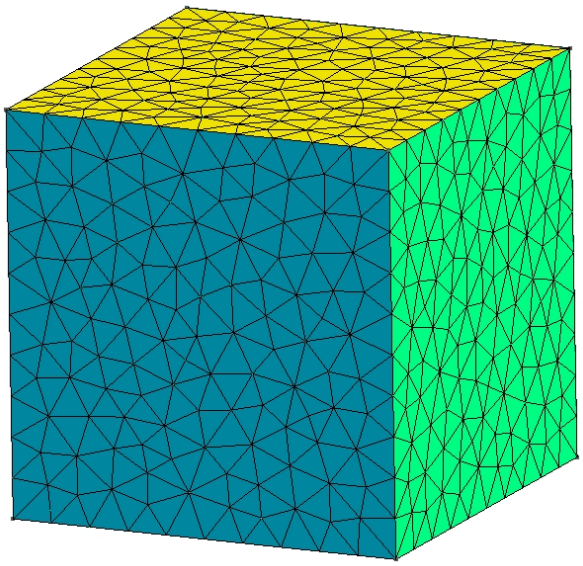


图 *4.3:* 例*4.2*：三维拟一致四面体网格.

N	M	L^2 误差	阶	H^1 误差	阶
2	2	1.449E-01		3.333E-01	
4	4	3.017E-02	2.26	1.370E-01	1.28
8	8	4.583E-03	2.72	4.335E-02	1.66
16	8	6.677E-04	2.78	1.310E-02	1.73
32	16	8.857E-05	2.91	3.702E-03	1.82

表 *4.10:* 例*4.2*：三维四面体网格的例子.

4.5 小结

本章提出了一种新的高阶异质多尺度有限元方法. 该方法通过局部最小二乘重构来恢复单元上的有效系数. 理论和数值结果均显示出，本章中的方法要优于 [51] 和 [48] 中的高阶 HMM-FEM 方法. 对于局部周期问题，本章对于服从 Dirichlet、Neumann 以及周期边条件的子问题给出了统一的离散误差分析，从而完善了异质多尺度有限元方法的理论框架.

第五章　基于重构基函数的间断 Galerkin 方法

本章考虑经典的椭圆问题

$$\begin{cases} -\operatorname{div}(A(\boldsymbol{x})\nabla u(\boldsymbol{x})) = f(\boldsymbol{x}), & \boldsymbol{x} \in \Omega \subset \mathbb{R}^d, \\ \Big(A(\boldsymbol{x})\nabla u(\boldsymbol{x})\Big) \cdot \boldsymbol{n}(\boldsymbol{x}) = b(\boldsymbol{x}), & \boldsymbol{x} \in \partial\Omega, \end{cases} \tag{5.1}$$

其中 $f(\boldsymbol{x}) \in L^2(\Omega)$，$\boldsymbol{n}(\boldsymbol{x})$ 是在 $\partial\Omega$ 上的单位外法向量，且 $A(\boldsymbol{x}) \in \mathcal{M}(\alpha, \beta, \Omega)$(定义见 §2.3节).

在对区域进行网格剖分之后，经典的有限元方法和间断 Galerkin 方法都需要给出形函数空间并构造相应的形函数. 这些形函数一般是在三角形单元或四边形单元 (三维时为四面体单元或六面体单元) 内通过插值给出 [32]. 但是在实际的工程应用中，越来越多地需要用到复杂的多边形 (多面体) 网格，尤其是多种几何体混合的多边形网格. 比如在地质建模中，混合多面体网格能够灵活地描述复杂的地质构造 [79], 而在这种复杂形状的单元上以及多种形状混合的网格上则较难给出合适的形函数，尤其是高阶的形函数. 在计算流体力学中需要在单元边界上计算数值通量，而与四面体单元相比，多面体单元的网格具有更少的边界数量和单元数量，结构更加简洁明了, 有利于提高计算效率 [49]. 在一些实际问题中还可能出现单元退化的情况, 比如地质建模中的“尖灭”现象会使原来网格单元中的一些相邻顶点粘在一起从而变成退化的单元 [83]. 此外，自适应网格和错配网格中因悬点等而产生的多边形网格也可看作退化单元.

在传统的有限元方法中，为了得到高阶的收敛性，往往需要在单元内增加自由度来构造高阶的形函数 [118]. 而 DG 方法使用了间断的形函数，从而带来了更多的自由度. 不管是有限元方法还是 DG 方法，最后形成的线性系统的规模都和整个区域上的自由度的数量直接相关，越多的自由度就意味着线性系统的规模越大，也就给线性系统的求解带来了困难. 传统上，人们往往用预处理器的技术来解决这个问题 ([92, 17, 84, 88, 100] 等).

本章将考虑一种求解二阶椭圆型方程(5.1)的数值方法. 首先在网格上通过最

小二乘重构给出基函数和逼近空间，这个过程并不依赖单元的几何形状，并且构造的逼近空间的维度总是等于网格中单元的个数；然后在间断 Galerkin 方法的框架下给出(5.1)的变分形式，进而求出方程的近似解. 这个方法只需要较少的自由度就可实现高阶收敛，而且适用于任意的多边形 (多面体) 网格.

本章的行文结构如下：§5.1 节介绍了在二阶椭圆型方程中间断 Galerkin 方法的变分形式的推导过程；§5.2 节对重构基函数的间断 Galerkin 方法的进行了详细的描述；§5.3节以一维情形为例，讨论了重构的基函数以及组装出的刚度矩阵的性质；§5.4节从理论上证明了本章所提出的算法的稳定性和收敛性；§5.5节通过一系列的数值算例展示了算法的数值特性；§5.6节对本章进行了总结.

5.1　间断 Galerkin 方法的变分形式

本节将在间断 Galerkin 方法 (简称 DG) 的框架下给出方程(5.1)的变分形式，其中具体推导过程参考 [11].

对于区域 Ω，令 $\mathcal{T}_h$ 为其上的尺寸为 h 的剖分 (见定义3.1)，其中 $d=2$(多边形网格) 或 $d=3$(多面体网格). 定义 Γ 为 $\mathcal{T}_h$ 中所有单元边界的集合，并且记 Γ^0 为区域内部的单元边界的集合 $\Gamma^0:=\Gamma\backslash\partial\Omega$，并且记

$$\int_{\Gamma}\cdot\,\mathrm{d}\boldsymbol{s}:=\sum_{e\in\Gamma}\int_{e}\cdot\,\mathrm{d}\boldsymbol{s},\qquad\int_{\Gamma^0}\cdot\,\mathrm{d}\boldsymbol{s}:=\sum_{e\in\Gamma^0}\int_{e}\cdot\,\mathrm{d}\boldsymbol{s}.$$

令 $\sigma(\boldsymbol{x})=A(\boldsymbol{x})\nabla u(\boldsymbol{x})$，于是(5.1)被改写为一个一阶方程组

$$\begin{cases}\sigma(\boldsymbol{x})=A(\boldsymbol{x})\nabla u(\boldsymbol{x}),\\ -\nabla\cdot\sigma(\boldsymbol{x})=f(\boldsymbol{x}),\end{cases}\quad\boldsymbol{x}\in\Omega,$$

并且满足边界条件

$$\sigma(\boldsymbol{x})\cdot\boldsymbol{n}(\boldsymbol{x})=b(\boldsymbol{x}),\quad\boldsymbol{x}\in\partial\Omega.$$

在两个方程的两边分别乘上试探函数 τ 和 v，并由分部积分有

$$\begin{aligned}\int_{\Omega}\sigma\cdot\tau\,\mathrm{d}\boldsymbol{x}&=-\int_{\Omega}u\nabla\cdot\left(A^{\mathrm{T}}\tau\right)\,\mathrm{d}\boldsymbol{x}+\sum_{K\in\mathcal{T}_h}\int_{\partial K}u\boldsymbol{n}_K\cdot\left(A^{\mathrm{T}}\tau\right)\,\mathrm{d}\boldsymbol{s},\\ \int_{\Omega}\sigma\cdot\nabla v\,\mathrm{d}\boldsymbol{x}&=\int_{\Omega}fv\,\mathrm{d}\boldsymbol{x}+\sum_{K\in\mathcal{T}_h}\int_{\partial K}\sigma\cdot\boldsymbol{n}_Kv\,\mathrm{d}\boldsymbol{s},\end{aligned}$$

这里 $\boldsymbol{n}_K$ 是单元 K 在 ∂K 上的外法向量.

定义分片多项式的函数空间 V_h 和 Σ_h 为

$$V_h := \left\{ v \in L^2(\Omega) \mid v|_K \in P(K), \quad \forall K \in \mathcal{T}_h \right\}, \tag{5.2}$$

$$\Sigma_h := \left\{ \tau \in [L^2(\Omega)]^d \mid \tau|_K \in \Sigma(K), \quad \forall K \in \mathcal{T}_h \right\}, \tag{5.3}$$

其中 $P(K) := \mathbb{P}_m(K)$ 且 $\Sigma(K) := [\mathbb{P}_m(K)]^d$. 于是在 DG 方法中，考虑如下的变分形式 (Cockburn 和舒其望 [40]): 求 $u_h \in V_h$ 和 $\sigma_h \in \Sigma_h$，使得

$$\begin{aligned} \int_\Omega \sigma_h \cdot \tau \,\mathrm{d}\boldsymbol{x} = &- \int_\Omega u_h \nabla_h \cdot \left(A^{\mathrm{T}} \tau \right) \mathrm{d}\boldsymbol{x} \\ &+ \sum_{K \in \mathcal{T}_h} \int_{\partial K} \hat{u}_K \boldsymbol{n}_K \cdot \left(A^{\mathrm{T}} \tau \right) \mathrm{d}\boldsymbol{s}, \qquad \forall \tau \in \Sigma_h, \end{aligned} \tag{5.4}$$

$$\int_\Omega \sigma_h \cdot \nabla_h v \,\mathrm{d}\boldsymbol{x} = \int_\Omega f v \,\mathrm{d}\boldsymbol{x} + \sum_{K \in \mathcal{T}_h} \int_{\partial K} \hat{\sigma}_K \cdot \boldsymbol{n}_K v \,\mathrm{d}\boldsymbol{s}, \qquad \forall v \in V_h, \tag{5.5}$$

其中函数 $\nabla_h v$ 和 $\nabla_h \cdot \tau$ 为分片连续函数，它们在每个单元 K 内的限制分别为 ∇v 和 $\nabla \cdot \tau$. 此外，(5.4)和(5.5)中的 $\hat{\sigma}_K$ 和 $\hat{u}_K$ 为数值通量，分别用来对 K 在边界上的 σ 和 u 进行估计. 通常 $\hat{\sigma}$ 和 $\hat{u}$ 由 σ_h 和 u_h 以及边界条件来表示. 在 DG 方法中，数值通量的选择很敏感，往往会影响数值方法的稳定性、精确性、以及刚度矩阵的稀疏性和对称性 (详见 [40]，[29] 和 [10]).

下面先给出一些符号和空间的定义. 定义分片 Sobolev 空间 $H^l(\mathcal{T}_h)$ 为

$$\forall g \in H^l(\mathcal{T}_h) \;\Rightarrow\; g|_K \in H^l(K), \quad \forall K \in \mathcal{T}_h.$$

因此，显然有 $V_h \subset H^l(\mathcal{T}_h)$ 和 $\Sigma_h \subset [H^l(\mathcal{T}_h)]^d$. 定义集合 $T(\Omega) := \prod_{K \in \mathcal{T}_h} L^2(\partial K)$. $T(\Omega)$ 中的函数在区域内部边界 Γ^0 上为双值，并且在区域边界 $\partial\Omega$ 上都是单值的. 令 K_1 和 K_2 是 $\mathcal{T}_h$ 中两个相邻的单元且具有公共边界 e. 将 K_1 和 K_2 在边界 e 上的外法向量分别记作 $\boldsymbol{n}_1$ 和 $\boldsymbol{n}_2$. 对于任意的 $q(\boldsymbol{x}) \in T(\Gamma)$，记 $q_i := q|_{\partial K_i}$，则令

$$\{q\} := \frac{1}{2}(q_1 + q_2), \quad [\![q]\!] := q_1 \boldsymbol{n}_1 + q_2 \boldsymbol{n}_2, \quad \text{在} e \in \Gamma^0 \text{上}.$$

类似地，对于 $\varphi \in [T(\Gamma)]^d$ 可定义

$$\{\varphi\} := \frac{1}{2}(\varphi_1 + \varphi_2), \quad [\![\varphi]\!] := \varphi_1 \cdot \boldsymbol{n}_1 + \varphi_2 \cdot \boldsymbol{n}_2, \quad \text{在} e \in \Gamma^0 \text{上}.$$

当 $e \in \partial\Omega$ 时，$q \in T(\Gamma)$ 和 $\varphi \in [T(\Gamma)]^d$ 在 e 上的值都是唯一确定的，因此令

$$[\![q]\!] := q\boldsymbol{n}, \quad \{\varphi\} := \varphi \quad \text{在} e \in \partial\Omega \text{上}.$$

在(5.4)和(5.5)中，都涉及到同一类求和公式 $\sum\limits_{K\in\mathcal{T}_h} \int_{\partial K} q_K\varphi_K \cdot \boldsymbol{n}_K \,\mathrm{d}\boldsymbol{s}$. 采用上述定义的符号，通过计算可知，对于所有的 $q \in T(\Gamma)$ 和 $\varphi \in [T(\Gamma)]^d$，都有

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} q_K\varphi_K \cdot \boldsymbol{n}_K \,\mathrm{d}\boldsymbol{s} = \int_\Gamma [\![q]\!] \cdot \{\varphi\} \,\mathrm{d}\boldsymbol{s} + \int_{\Gamma^0} \{q\}[\![\varphi]\!] \,\mathrm{d}\boldsymbol{s}.$$

将上式应用到(5.4)和(5.5)中可得

$$\begin{aligned}
\int_\Omega \sigma_h \cdot \tau \,\mathrm{d}\boldsymbol{x} = &- \int_\Omega u_h \nabla_h \cdot \left(A^{\mathrm{T}}\tau\right) \,\mathrm{d}\boldsymbol{x} \\
&+ \int_\Gamma [\![\hat{u}]\!] \cdot \{A^{\mathrm{T}}\tau\} \,\mathrm{d}\boldsymbol{s} + \int_{\Gamma^0} \{\hat{u}\}[\![A^{\mathrm{T}}\tau]\!] \,\mathrm{d}\boldsymbol{s}, \qquad \forall \tau \in \Sigma_h, \qquad (5.6) \\
\int_\Omega \sigma_h \cdot \nabla_h v \,\mathrm{d}\boldsymbol{x} - &\int_\Gamma \{\hat{\sigma}\} \cdot [\![v]\!] \,\mathrm{d}\boldsymbol{s} - \int_{\Gamma^0} [\![\hat{\sigma}]\!]\{v\} = \int_\Omega fv \,\mathrm{d}\boldsymbol{x}, \qquad \forall v \in V_h, \qquad (5.7)
\end{aligned}$$

其中 $\hat{u} = (\hat{u}_K)_{K\in\mathcal{T}_h}$ 且 $\hat{\sigma} = (\hat{\sigma})_{K\in\mathcal{T}_h}$.

同时，注意到(5.6)式等号右边的第一项可化为

$$\int_\Omega u_h \nabla_h \cdot \left(A^{\mathrm{T}}\tau\right) \,\mathrm{d}\boldsymbol{x} = -\int_\Omega \tau \cdot (A\nabla_h u_h) \,\mathrm{d}\boldsymbol{x} + \int_\Gamma \{A^{\mathrm{T}}\tau\} \cdot [\![u_h]\!] \,\mathrm{d}\boldsymbol{s} + \int_{\Gamma^0} [\![A^{\mathrm{T}}\tau]\!]\{u_h\} \,\mathrm{d}\boldsymbol{s}.$$

将其代入到(5.6)中就有

$$\int_\Omega \sigma_h \cdot \tau \,\mathrm{d}\boldsymbol{x} = \int_\Omega (A\nabla_h u_h) \cdot \tau \,\mathrm{d}\boldsymbol{x} + \int_\Gamma [\![\hat{u} - u_h]\!] \cdot \{A^{\mathrm{T}}\tau\} \,\mathrm{d}\boldsymbol{s} + \int_{\Gamma^0} \{\hat{u} - u_h\}[\![A^{\mathrm{T}}\tau]\!] \,\mathrm{d}\boldsymbol{s}.$$

在上式中取 $\tau = \nabla_h v$，并代入到(5.7)中，即可得

$$\mathcal{B}_h(u_h, v) = \int_\Omega fv \,\mathrm{d}\boldsymbol{x}, \quad \forall v \in V_h, \tag{5.8}$$

其中

$$\begin{aligned}
\mathcal{B}_h(u_h, v) := &\int_\Omega (A\nabla_h u_h) \cdot \nabla_h v_h \,\mathrm{d}\boldsymbol{x} \\
&+ \int_\Gamma \Big((A[\![\hat{u} - u_h]\!]) \cdot \{\nabla_h v\} - \{\hat{\sigma}\} \cdot [\![v]\!] \Big) \,\mathrm{d}\boldsymbol{s} \qquad (5.9) \\
&+ \int_{\Gamma^0} \Big(\{\hat{u} - u_h\}[\![A^{\mathrm{T}}\nabla_h v]\!] - [\![\hat{\sigma}]\!]\{v\} \Big) \,\mathrm{d}\boldsymbol{s}.
\end{aligned}$$

于是(5.8)式就是方程(5.1)在 DG 方法下的变分形式. 其中双线性算子 $\mathcal{B}_h(\cdot,\cdot)$ 中，数值通量 $\hat{u}$ 和 $\hat{\sigma}$ 的选取有多种形式 (见 [13, 46, 40, 98, 28] 等)，在 [11] 中有对各

种数值通量的相容性与稳定性的讨论.

5.2 算法描述

5.2.1 基函数和逼近空间

本节将给出在 $\mathcal{T}_h$ 上利用最小二乘重构来构造基函数的具体方法. 对于每个单元 $K \in \mathcal{T}_h$，设其内部存在一个采样点 $\boldsymbol{x}_K$. 采用 §3.3节的定义，有

$$I_K = \{\boldsymbol{x}_K\}, \quad \forall K \in \mathcal{T}_h,$$

并且

$$I_\Omega = \{\, \boldsymbol{x}_K \mid \forall\ K \in \mathcal{T}_h \,\},$$

则显然有 $\#I_\Omega = n_e$，其中 n_e 为 $\mathcal{T}_h$ 中的单元个数. 对于每个单元 K，按照(3.2)的方式构造单元模板 $S(K)$，于是

$$\mathcal{I}_K = \{\, \boldsymbol{x}_{K'} \mid \forall\ K' \in S(K) \,\}.$$

并且 $n_K = \#\mathcal{I}_K = \#S(K)$.

定义向量 $\boldsymbol{e}_K \in \mathbb{R}^{n_e}$，其分量为

$$\boldsymbol{e}_{K,K'} = \delta_{KK'}, \quad \forall K' \in \mathcal{T}_h,$$

其中

$$\delta_{KK'} = \begin{cases} 1, & K = K', \\ 0, & K \neq K'. \end{cases}$$

于是结合 §3.3节中介绍的重构算子 $\widetilde{\mathcal{R}}_m$，可在 Ω 上定义函数 $\lambda^K(\boldsymbol{x})$ 为

$$\lambda^K(\boldsymbol{x}) := \widetilde{\mathcal{R}}_m \boldsymbol{e}_K. \tag{5.10}$$

由重构算子的定义(3.6)，可以知道 $\lambda^K(\boldsymbol{x})$ 为分片多项式函数，并且

$$\lambda^K(\boldsymbol{x})\Big|_{K'} \equiv \widetilde{\mathcal{R}}_m^{K'} \left(\mathcal{S}_{K'} \boldsymbol{e}_K\right)(\boldsymbol{x}), \quad \boldsymbol{x} \in K'. \tag{5.11}$$

对所有的单元 $K \in \mathcal{T}_h$，都可以按照上述过程构造 $\lambda^K(\boldsymbol{x})$. 由于算子 $\widetilde{\mathcal{R}}_m$ 是线性算子，并且 $\{\boldsymbol{e}_K\}$ 线性无关，因此函数组 $\{\lambda^K(\boldsymbol{x})\}$ 也是线性无关的. 于是定义逼近

空间 U_h，该空间以函数组 $\{\lambda^K\}$ 为基函数.

$$U_h := \operatorname{span}\left\{ \lambda^K(\boldsymbol{x}) \mid \forall K \in \mathcal{T}_h \right\}.$$

不难知道 $\operatorname{rank}(U_h) = n_e$. 显然，$U_h$ 中的函数为分片多项式函数，并且有 $U_h \subset V_h$.

由上述过程得到的 U_h 和 $\{\lambda^K(\boldsymbol{x})\}$ 具有如下三条简单的性质.

性质 5.1. $\lambda^K(\boldsymbol{x})$ 在 Ω 上具有紧支集，且

$$\operatorname{supp}\ \lambda^K(\boldsymbol{x}) = \{\, K' \in \mathcal{T}_h \mid K \in S(K') \,\}. \tag{5.12}$$

其中 supp 表示函数的紧支集.

证明. 由(5.11)式有

$$\begin{aligned} K \notin S(K') &\Leftrightarrow \mathcal{S}_{K'}\boldsymbol{e}_K \equiv \boldsymbol{0} \\ &\Leftrightarrow \widetilde{\mathcal{R}}_m^{K'}(\mathcal{S}_{K'}\boldsymbol{e}_K) \equiv 0 \\ &\Leftrightarrow \lambda^K(\boldsymbol{x})\big|_{K'} \equiv 0, \end{aligned}$$

因此(5.12)成立. □

注. 由这个性质可知，只有 K' 的模板中包含单元 K，那么基函数 $\lambda^K(\boldsymbol{x})$ 才在单元 K' 上不为 0. 在此要强调的是，要区分开 $\lambda^K(\boldsymbol{x})$ 的紧支集和单元模板 $S(K)$，这是两个完全不同的概念.

性质 5.2. 对任意连续函数 $v(\boldsymbol{x}) \in C^0(\Omega)$，都有 $\mathcal{R}_m v \in U_h$，并且

$$\mathcal{R}_m v = \sum_{K\in\mathcal{T}_h} v(\boldsymbol{x}_K)\lambda^K(\boldsymbol{x}). \tag{5.13}$$

证明. 定义 $\boldsymbol{v} := (\cdots, v(\boldsymbol{x}_K) \cdots)^{\mathrm{T}}$, $\forall K \in \mathcal{T}_h$，于是由(3.10)式有

$$\mathcal{R}_m v = \widetilde{\mathcal{R}}_m \boldsymbol{v}.$$

由于 $\boldsymbol{v} = \sum\limits_{K\in\mathcal{T}_h} v(\boldsymbol{x}_K)\boldsymbol{e}_K$，考虑到(5.10)和 $\widetilde{\mathcal{R}}_m$ 是线性算子，于是有

$$\begin{aligned} \widetilde{\mathcal{R}}_m \boldsymbol{v} &= \sum_{K\in\mathcal{T}_h} v(\boldsymbol{x}_K)\widetilde{\mathcal{R}}_m \boldsymbol{e}_K \\ &= \sum_{K\in\mathcal{T}_h} v(\boldsymbol{x}_K)\lambda^K(\boldsymbol{x}). \end{aligned}$$

由此就得到了(5.13)，因此有 $\mathcal{R}_m v \in U_h$.

□

性质 5.3. 设两个相邻单元为 K_1 和 K_2，并且它们的公共边为 e. 如果两个单元的模板 $S(K_1)$ 和 $S(K_2)$ 完全相同，那么对于任意的基函数 $\lambda^K(\boldsymbol{x}), \forall K \in \mathcal{T}_h$，可知

$$[\![\lambda^K(\boldsymbol{x})]\!] = 0, \quad 在e上.$$

即 $\lambda^K(\boldsymbol{x})$ 在 $\overline{K_1 \bigcup K_2}$ 上是连续的.

证明. 由于 $S(K_1) = S(K_2)$，因此它们的采样点集 $\mathcal{I}_{K_1} = \mathcal{I}_{K_2}$，并且 $\mathcal{S}_{K_1}\boldsymbol{e}_K = \mathcal{S}_{K_2}\boldsymbol{e}_K$, $\forall K \in \mathcal{T}_h$. 于是在 $S(K_1)$ 和 $S(K_2)$ 上得到的是完全一样的最小二乘重构问题，记这一问题的解为 $q(\boldsymbol{x})$，那么由基函数的定义(5.11)有

$$\lambda^K(\boldsymbol{x})\Big|_{K_1} = q(\boldsymbol{x})\Big|_{K_1}, \qquad \lambda^K(\boldsymbol{x})\Big|_{K_2} = q(\boldsymbol{x})\Big|_{K_2}.$$

由于 $q(\boldsymbol{x}) \in \mathbb{P}_m(S(K_i)), i = 1, 2$，因此可知 $\lambda^K(\boldsymbol{x})$ 在 $\overline{K_1 \bigcup K_2}$ 上是连续的. □

为了保证基函数 $\lambda^K(\boldsymbol{x})$ 的存在性，假设在每个模板 $S(K')$ 上，假设A(详见§3.3节) 都成立. 于是总能找到一组系数 $C^K_{\boldsymbol{\alpha},K'}$ 满足

$$\lambda^K(\boldsymbol{x})\Big|_{K'} = \sum_{|\boldsymbol{\alpha}|\leqslant m} C^K_{\boldsymbol{\alpha},K'}(\boldsymbol{x} - \boldsymbol{x}_{K'})^{\boldsymbol{\alpha}}, \quad \boldsymbol{x} \in K'.$$

并且，假设定理3.6和定理3.8的条件成立，那么存在和 K 无关的常数，使得

$$|C^K_{\boldsymbol{\alpha},K'}| \leqslant C, \quad \forall \boldsymbol{\alpha},\ K,\ K'.$$

在实际计算中，只需要记录下所有的 $C^K_{\boldsymbol{\alpha},K'}$，即可计算出任意基函数 $\lambda^K(\boldsymbol{x})$ 的值.

5.2.2 二阶椭圆方程的数值求解

由于 $U_h \in V_h$，因此可以在(5.8) 的基础上给出方程(5.1)的变分形式.

本章采用经典的内罚函数法 (简称 IP)[46, 11] 来给出具体的变分形式. 将数值通量取为

$$\hat{u} = \{u_h\},\ 在\Gamma上; \quad \hat{\sigma} = \{A\nabla_h u_h\} - \alpha_j([\![u_h]\!]),\ 在\Gamma^0上; \quad \hat{\sigma} = \{A\nabla_h u_h\},\ 在\partial\Omega上.$$

并且 $\alpha_j(\varphi) = \mu\varphi$. 其中的 $\mu = \eta_e h_e^{-1}$ 为惩罚权重函数，h_e 为边界 e 的长度，η_e 为一个给定的正数. 显然, $\hat{u}$ 和 $\hat{\sigma}$ 在 Γ 上是单值的，因此在 Γ^0 上有 $\{\hat{u} - u_h\} = 0$ 和

$[\![\hat{\sigma}]\!] = 0$，于是在(5.9)中有

$$\int_{\Gamma^0} \{\hat{u} - u_h\} [\![A^{\mathrm{T}} \nabla_h v]\!] \,\mathrm{d}\boldsymbol{s} = 0, \qquad \int_{\Gamma^0} [\![\hat{\sigma}]\!] \{v\} \,\mathrm{d}\boldsymbol{s} = 0.$$

又由于在 Γ^0 上 $[\![\hat{u}]\!] = 0$ 并且在 $\partial\Omega$ 上 $\hat{u} = u_h$，因此

$$\int_{\Gamma} \big(A[\![\hat{u} - u_h]\!]\big) \cdot \{\nabla_h v\} \,\mathrm{d}\boldsymbol{s} = -\int_{\Gamma^0} \big(A[\![u_h]\!]\big) \cdot \{\nabla_h v\} \,\mathrm{d}\boldsymbol{s}.$$

最后，由边界条件 $(A\nabla_h u_h) \cdot \boldsymbol{n} = b$ 有

$$\int_{\Gamma} \{\hat{\sigma}\} \cdot [\![v]\!] \,\mathrm{d}\boldsymbol{s} = \int_{\Gamma^0} \{A\nabla_h u_h\} \cdot [\![v]\!] \,\mathrm{d}\boldsymbol{s} - \int_{\Gamma^0} \alpha_j([\![u_h]\!]) \cdot [\![v]\!] \,\mathrm{d}\boldsymbol{s} + \int_{\partial\Omega} bv \,\mathrm{d}\boldsymbol{s}.$$

将上面的几个式子代入(5.8)并整理，即可得到 IP 方法的变分问题：求 $u_h \in U_h$，使其满足

$$\mathcal{B}_h^{IP}(u_h, v_h) = F(v_h), \quad \forall v_h \in U_h. \tag{5.14}$$

其中的双线性算子为定义为

$$\begin{aligned} \mathcal{B}_h^{IP}(u_h, v_h) = &\int_{\Omega} \big(A(\boldsymbol{x})\nabla_h u_h(\boldsymbol{x})\big) \cdot \nabla_h v_h(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \\ &- \int_{\Gamma^0} \big(A(\boldsymbol{x})[\![u_h(\boldsymbol{x})]\!]\big) \cdot \{\nabla_h v_h(\boldsymbol{x})\} \,\mathrm{d}\boldsymbol{s} \\ &- \int_{\Gamma^0} \big(A(\boldsymbol{x})\{\nabla_h u_h(\boldsymbol{x})\}\big) \cdot [\![v_h(\boldsymbol{x})]\!] \,\mathrm{d}\boldsymbol{s} + \alpha^j(u_h, v_h), \end{aligned} \tag{5.15}$$

其中

$$\alpha^j(u_h, v_h) = \int_{\Gamma^0} \alpha_j([\![u_h(\boldsymbol{x})]\!]) \cdot [\![v_h(\boldsymbol{x})]\!] \,\mathrm{d}\boldsymbol{x}.$$

而右端项 $F(\cdot)$ 为

$$\begin{aligned} F(v_h) &= \int_{\Omega} f(\boldsymbol{x}) v_h(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} b(\boldsymbol{x}) v_h(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{s} \\ &= (f, v_h)_{L^2(\Omega)} + (b, v_h)_{L^2(\partial\Omega)}. \end{aligned}$$

满足这个变分形式的 u_h 即为方程(5.1)的弱解.

进一步，由于 $u_h \in U_h$，于是总能找到 $\boldsymbol{u} = (u_1, u_2, \cdots, u_{n_e})^{\mathrm{T}} \in \mathbb{R}^{n_e}$ 使得

$$u_h(\boldsymbol{x}) = \sum_{K \in \mathcal{T}_h} u_K \lambda^K(\boldsymbol{x}).$$

将 $v_h(\boldsymbol{x})$ 取为 $\lambda^{K'}(\boldsymbol{x})$，则变分形式(5.4)可以被改写为

$$\sum_{K\in\mathcal{T}_h}\mathcal{B}_h^{IP}(\lambda^K,\lambda^{K'})u_K = F(\lambda^{K'}),\ \forall K'\in\mathcal{T}_h. \tag{5.16}$$

这样就得到了下述线性系统

$$\boldsymbol{B}\boldsymbol{u}=\boldsymbol{F}, \tag{5.17}$$

其中的刚度矩阵为

$$\boldsymbol{B}=\begin{pmatrix} \mathcal{B}_h^{IP}(\lambda^{K_1},\lambda^{K_1}) & \mathcal{B}_h^{IP}(\lambda^{K_2},\lambda^{K_1}) & \cdots & \mathcal{B}_h^{IP}(\lambda^{K_{ne}},\lambda^{K_1}) \\ \mathcal{B}_h^{IP}(\lambda^{K_1},\lambda^{K_2}) & \mathcal{B}_h^{IP}(\lambda^{K_2},\lambda^{K_2}) & \cdots & \mathcal{B}_h^{IP}(\lambda^{K_{ne}},\lambda^{K_2}) \\ & \cdots & \cdots & \\ \mathcal{B}_h^{IP}(\lambda^{K_1},\lambda^{K_{ne}}) & \mathcal{B}_h^{IP}(\lambda^{K_2},\lambda^{K_{ne}}) & \cdots & \mathcal{B}_h^{IP}(\lambda^{K_{ne}},\lambda^{K_{ne}}) \end{pmatrix}, \tag{5.18}$$

并且 $\boldsymbol{F}$ 为

$$\boldsymbol{F}=\left(F(\lambda^{K_1}),F(\lambda^{K_2}),\cdots,F(\lambda^{K_{ne}})\right)^{\mathrm{T}}.$$

求解上述线性系统就能得到 $\boldsymbol{u}$，从而给出 $u_h(\boldsymbol{x})$.

注. 注意到 U_h 的自由度总是等于网格 $\mathcal{T}_h$ 中单元的个数，因此如果网格固定，线性系统(5.17)的规模总是不变的. 如果想得到高阶格式，那么只需要提高重构多项式的阶数，而不需要增加自由度.

5.3 一维情形的例子

本节将用一个一维的例子来具体阐述上一节所提出的算法，以便读者能有个更直观的理解. 考虑区间 $[0,1]$ 上的 Neumann 边值问题

$$\begin{cases} -\dfrac{\mathrm{d}^2u}{\mathrm{d}x^2}=f, \quad x\in[0,1], \\ u_x(0)=g_1, \quad u_x(1)=g_2. \end{cases} \tag{5.19}$$

这里的 u_x 表示 $u(x)$ 对 x 的导数. 如图 5.1所示，网格被剖分为五个单元 $K_1,K_2,\cdots,K_5$. 取每个单元的中点为采样点 x_i.

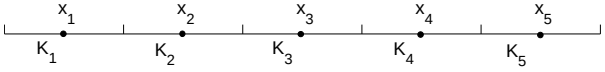


图 *5.1:* [0,1] 区间上的一维网格

5.3.1　基函数

接下来用最小二乘方法来构造基函数. 首先构造出五个单元的单元模板，具体如下:

$$S(K_1) = \{K_1, K_2\}, \quad S(K_5) = \{K_4, K_5\},$$
$$S(K_i) = \{K_{i-1}, K_i, K_{i+1}\},\ i = 2, 3, 4.$$

为简便起见，将基函数记为

$$\lambda_i(x) := \lambda^{K_i}(x), \quad i = 1, 2, 3, 4, 5.$$

接下来在每个单元模板上求解最小二乘问题来给出基函数. 设重构阶数 $m = 1$. 以 $\lambda_1(x)$ 为例，由于 K_1 出现在 $S(K_1)$ 和 $S(K_2)$ 中，因此 $\lambda_1(x)$ 的支集为 $K_1 \bigcup K_2$. 在 $S(K_1)$ 中，由(5.11)可知

$$\lambda_1(x)|_{K_1} = \arg\min_{p(x)\in\mathbb{P}_1} \left(|p(x_1) - 1|^2 + |p(x_2) - 0|^2\right),$$

求解该最小二乘问题可得

$$\lambda_1(x)|_{K_1} = -\frac{1}{h}(x - x_1) + 1.$$

类似地，在 $S(K_2)$ 中可以得到

$$\lambda_1(x)|_{K_2} = -\frac{1}{2h}(x - x_2) + \frac{1}{3}.$$

可以很明显地看出，$\lambda_1(x)$ 是间断的分片线性函数. 它在 K_1 和 K_2 的交点，即 $x = 0.2$ 处有一个跳跃. 下面继续计算出其他的基函数，并在表5.1中列出全部的基函数. 图 5.2是所有的基函数的示意图. 在图中可以清楚的看到每个基函数的支集以及它们各自在单元边界上的跳跃.

在这个例子中, $S(K_i)$ 与 λ_i 的支集是相等的，但这并不总是成立的. 事实上，如果将 $S(K_1)$ 取为 $\{K_1, K_2, K_3\}$，那么有

$$\text{supp } \lambda_3(x) = \{K_1, K_2, K_3, K_4\},$$

但是

$$S(K_3) = \{K_2, K_3, K_4\}.$$

$\lambda_1(x)\|_{K_1} = -\frac{1}{h}(x-x_1)+1,$		$\lambda_1(x)\|_{K_2} = -\frac{1}{2h}(x-x_2)+\frac{1}{3}.$
$\lambda_2(x)\|_{K_1} = \frac{1}{h}(x-x_1),$	$\lambda_2(x)\|_{K_2} = \frac{1}{3},$	$\lambda_2(x)\|_{K_3} = -\frac{1}{2h}(x-x_3)+\frac{1}{3},$
$\lambda_3(x)\|_{K_2} = \frac{1}{2h}(x-x_2)+\frac{1}{3},$	$\lambda_3(x)\|_{K_3} = \frac{1}{3},$	$\lambda_3(x)\|_{K_4} = -\frac{1}{2h}(x-x_4)+\frac{1}{3},$
$\lambda_4(x)\|_{K_3} = \frac{1}{2h}(x-x_3)+\frac{1}{3},$	$\lambda_4(x)\|_{K_4} = \frac{1}{3},$	$\lambda_4(x)\|_{K_5} = -\frac{1}{h}(x-x_5),$
$\lambda_5(x)\|_{K_4} = \frac{1}{2h}(x-x_4)+\frac{1}{3},$		$\lambda_5(x)\|_{K_5} = \frac{1}{h}(x-x_5)+1.$

表 *5.1:* $m=1$ 时重构出的基函数组.

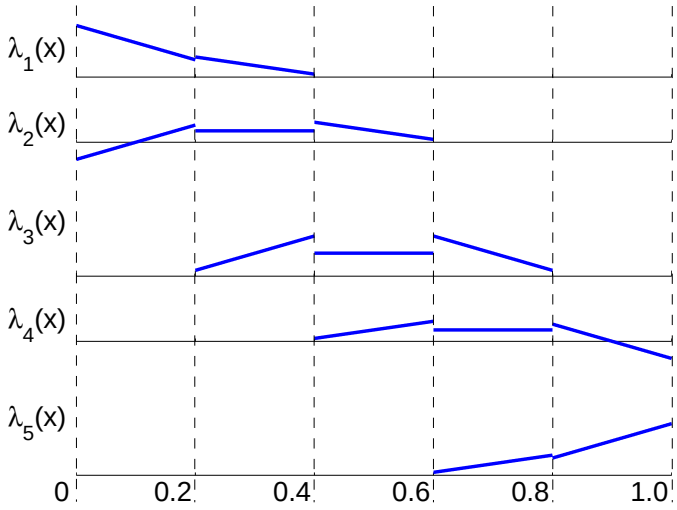


图 *5.2:* 一阶重构的基函数示意图

下面再给出一个 $m=2$ 的例子. 由于重构二次多项式需要至少三个采样点，于是构造如下的单元模板.

$$S(K_1)=\{K_1,K_2,K_3\},\quad S(K_5)=\{K_3,K_4,K_5\},$$
$$S(K_i)=\{K_{i-1},K_i,K_{i+1}\},\ i=2,3,4.$$

在这套模板下，基函数 $\lambda_i(x), i=1,2,4,5$ 的紧支集和 $m=1$ 时是相同的，但是 $\lambda_3(x)$ 的紧支集变为 $\{K_1,K_2,K_3,K_4,K_5\}$. 此外，还需注意到 $S(K_1)$ 和 $S(K_2)$，以及 $S(K_4)$ 和 $S(K_5)$ 分别是相同的.

由最小二乘重构得到基函数组 $\{\lambda_i(x)\}_{i=1}^5$，并列在表5.2中. 同样的，将这些基函数一起展示在图5.3中. 从图上可以明显地看出所有的基函数在 $\overline{K_1\bigcup K_2}$ 和 $\overline{K_1\bigcup K_2}$ 上都是连续的，这也就验证了性质5.3.

$\lambda_1(x)\|_{K_1}=\frac{1}{2h^2}(x-x_1)^2-\frac{3}{2h}(x-x_1)+1,$	$\lambda_1(x)\|_{K_2}=\frac{1}{2h^2}(x-x_2)^2-\frac{1}{2h}(x-x_2).$
$\lambda_2(x)\|_{K_1}=-\frac{1}{h^2}(x-x_1)^2+\frac{2}{h}(x-x_1),$	$\lambda_2(x)\|_{K_2}=-\frac{1}{h^2}(x-x_2)^2+1,$
$\lambda_2(x)\|_{K_1}=\frac{1}{2h^2}(x-x_3)^2-\frac{1}{2h}(x-x_3).$	
$\lambda_3(x)\|_{K_1}=\frac{1}{2h^2}(x-x_1)^2-\frac{1}{2h}(x-x_1),$	$\lambda_3(x)\|_{K_2}=\frac{1}{2h^2}(x-x_2)^2+\frac{1}{2h}(x-x_2),$
$\lambda_3(x)\|_{K_3}=-\frac{1}{h^2}(x-x_3)^2+1,$	$\lambda_3(x)\|_{K_4}=\frac{1}{2h^2}(x-x_4)^2-\frac{1}{2h}(x-x_4),$
$\lambda_3(x)\|_{K_5}=\frac{1}{2h^2}(x-x_5)^2+\frac{1}{2h}(x-x_5).$	
$\lambda_4(x)\|_{K_3}=\frac{1}{2h^2}(x-x_3)^2+\frac{1}{2h}(x-x_3),$	$\lambda_4(x)\|_{K_4}=-\frac{1}{h^2}(x-x_4)^2+1,$
$\lambda_4(x)\|_{K_5}=-\frac{1}{h^2}(x-x_5)^2-\frac{2}{h}(x-x_5).$	
$\lambda_5(x)\|_{K_4}=\frac{1}{2h^2}(x-x_4)^2+\frac{1}{2h}(x-x_4),$	$\lambda_5(x)\|_{K_5}=\frac{1}{2h^2}(x-x_5)^2+\frac{3}{2h}(x-x_5)+1.$

表 5.2: $m=2$ 时重构出的基函数组.

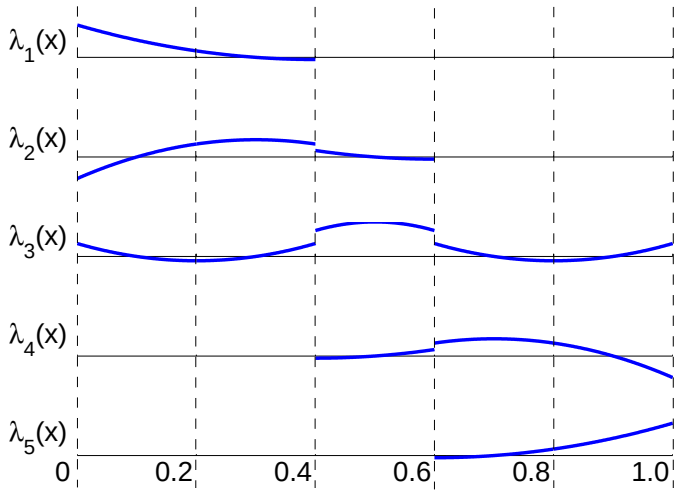


图 *5.3:* 二阶重构的基函数示意图

5.3.2 刚度矩阵

在方程(5.19)中，$d=1$ 且 $A(x)\equiv 1$. 于是由(5.15)，方程的双线性算子简化为

$$\begin{aligned}\mathcal{B}_h^{IP}(u_h,v_h)=&\sum_{K\in\mathcal{T}_h}\int_K\frac{\mathrm{d}u_h}{\mathrm{d}x}\frac{\mathrm{d}v_h}{\mathrm{d}x}\,\mathrm{d}x\\&-\sum_{e\in\Gamma^0}\left(\llbracket u_h\rrbracket\left\{\frac{\mathrm{d}v_h}{\mathrm{d}x}\right\}+\left\{\frac{\mathrm{d}u_h}{\mathrm{d}x}\right\}\llbracket v_h\rrbracket\right)+\mu\sum_{e\in\Gamma^0}\llbracket u_h\rrbracket\llbracket v_h\rrbracket.\end{aligned}\tag{5.20}$$

由此就可以依照(5.18)构造刚度矩阵 $\boldsymbol{B}$.

举例来说，令 $m=1$ 并采用表5.1中构造的基函数. 通过计算可以知道刚度矩阵 $\boldsymbol{B}$ 的具体形式为

$$\frac{1}{144h}\begin{pmatrix}168+2\mu h & -126-5\mu h & -36+4\mu h & -6-\mu h & 0\\ -126-5\mu h & 144+14\mu h & 12-14\mu h & -24+6\mu h & -6-\mu h\\ -36+4\mu h & 12-14\mu h & 48+20\mu h & 12-14\mu h & -36+4\mu h\\ -6-\mu h & -24+6\mu h & 12-14\mu h & 144+14\mu h & -126-5\mu h\\ 0 & -6-\mu h & -36+4\mu h & -126-5\mu h & 168+2\mu h\end{pmatrix}.$$

可以看出矩阵 $\boldsymbol{B}$ 的第一行有 4 个非零元素，对这一点可以做如下解释：从图5.1中可知 $\lambda_2(x)$ 和 $\lambda_3(x)$ 的紧支集都与 $\lambda_1(x)$ 的紧支集相交非空，因此显然有刚度矩

阵第一行的前三个元素非零. 而又由于在 $e := K_2 \bigcap K_3$ 上，

$$[\![\lambda_1]\!], \quad \left\{\frac{\mathrm{d}\lambda_1}{\mathrm{d}x}\right\}, \quad [\![\lambda_4]\!], \quad \left\{\frac{\mathrm{d}\lambda_4}{\mathrm{d}x}\right\}$$

都不为零，因此有 $\mathcal{B}_h^{IP}(\lambda_4, \lambda_1) \neq 0$.

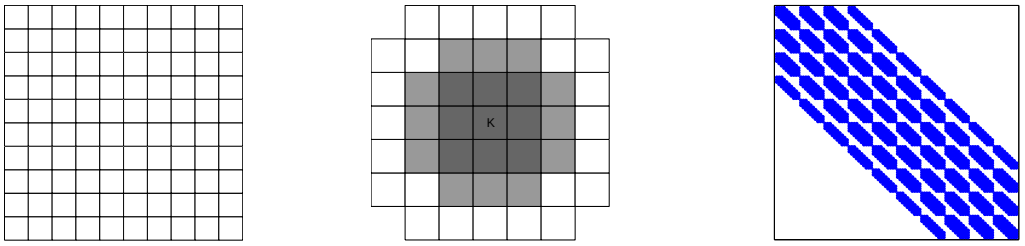
高维时的情况更加复杂，刚度矩阵中非零元素的个数取决于相对应的网格单元的几何形状和位置. 在此只讨论一个简单的例子. 考虑二维四边形单元网格，如图5.4(a)所示. 令重构阶数 $m = 1$ 且模板层数 $t = 1$. 考虑一个位于区域中心位置的单元 K，那么刚度矩阵 $\boldsymbol{B}$ 中对应于 K 的那一行将有 45 个元素. 如果把所有使得

$$\mathcal{B}_h^{IP}(\lambda^{K'}, \lambda^K) \neq 0 \tag{5.21}$$

的单元 K' 标出来，那么就得到了图5.4(b)，图中用三种颜色标出了三类单元.

- 颜色最深的 9 个单元表示 $\lambda^K(\boldsymbol{x})$ 的紧支集，对于这一区域的任意单元 K'，(5.21)显然成立.
- 颜色较深的 12 个单元，由于它们对应的基函数在 K 的紧支集上非零，因此也使得(5.21)成立.
- 白色的单元的基函数的紧支集和 K 的紧支集相邻，由于(5.20)中 $[\![\cdot]\!]$ 和 $\{\cdot\}$ 的存在，而使得 (5.21)成立.

整个稀疏矩阵 $\boldsymbol{B}$ 的示意图见图5.4(c).



(a) 四边形单元网格　(b) K 和周围的单元　(c) 刚度矩阵示意

图 5.4: 四边形单元网格与稀疏矩阵.

5.4　收敛性分析

本节将讨论双线性算子(5.15)的有界性，强制性和收敛性. [11] 中已经证明了在 V_h 空间上(5.9)的有界性和强制性. 由于 $U_h \subset V_h$ 且 LP 双线性算子是(5.9)的特

定形式，因此上述结论在本章中也适用. 在本节中将给出这些证明的完整表述.

对 $v \in H^2(\mathcal{T}_h)$ 定义如下的半范数

$$|v|_{1,h} := \sum_{K\in\mathcal{T}_h} |v|_{H^1(K)}, \qquad |v|_* := \sum_{e\in\Gamma^0} h_e^{-1} \|[\![v]\!]\|^2_{L^2(e)}.$$

接下来，定义内积

$$(w,v)_U := \sum_{K\in\mathcal{T}_h} \int_K (\nabla w\cdot\nabla v)\,\mathrm{d}\boldsymbol{x} + \sum_{e\in\Gamma} h_e^{-1}\int_e [\![w]\!][\![v]\!]\,\mathrm{d}\boldsymbol{s}, \quad w,v\in H^2(\mathcal{T}_h),$$

从而诱导出范数为

$$\|v\|_U := (v,v)_U^{1/2} = \left(|v|^2_{1,h} + |v|^2_*\right)^{1/2}. \tag{5.22}$$

由 [9, 引理 2.1]，可知存在常数 C 使得

$$\|v\|_{L^2(\Omega)} \leqslant C\left(|v|^2_{1,h} + |v|^2_*\right)^{1/2}, \quad \forall v\in H^2(\mathcal{T}_h),$$

因此(5.22)所定义的范数成立. 这使得 U_h 在 $(\cdot,\cdot)_U$ 和 $\|\cdot\|_U$ 下成为一个 Hilbert 空间.

5.4.1 有界性

首先证明算子 $\mathcal{B}_h^{IP}$ 的有界性. 由于 $A\in\mathcal{M}(\alpha,\beta,\Omega)$，显然有

$$\left|\sum_{K\in\mathcal{T}_h}\int_K (A\nabla w\cdot\nabla v)\,\mathrm{d}\boldsymbol{x}\right| \leqslant C|w|_{1,h}|v|_{1,h}, \quad \forall w,v\in H^2(\mathcal{T}_h).$$

接下来考虑 $\int_{\Gamma^0}(A[\![w]\!])\cdot\{\nabla_h v\}\,\mathrm{d}\boldsymbol{s}$ 和 $\int_{\Gamma^0}(A\{\nabla_h w\})\cdot[\![v]\!]\,\mathrm{d}\boldsymbol{s}$. 首先令 e 是单元 K 的边界，于是对任意的 $w\in H^2(K)$，由 [9, (2.5) 式] 可知

$$\left\|\frac{\partial w}{\partial n}\right\|^2_{L^2(e)} \leqslant C(h_e^{-1}|w|^2_{H^1(K)} + h_e|w|^2_{H^2(K)}) \leqslant C|w|_{H^1(K)}.$$

由此可知，对任意的 $w, v \in U_h$ 有

$$\begin{aligned}\left|\int_{\Gamma^0}(A\{\nabla_h w\})\cdot[\![v]\!]\,\mathrm{d}\boldsymbol{s}\right| &= \left|\sum_{e\in\Gamma^0}\int_e (A\{\nabla_h w\})\cdot[\![v]\!]\,\mathrm{d}\boldsymbol{s}\right| \\ &\leqslant C\left[\sum_{K\in\mathcal{T}_h}\left(|w|^2_{H^1(K)}+h_K^2|w|^2_{H^2(K)}\right)\right]^{1/2} \\ &\quad\times\left[\sum_{e\in\Gamma^0}h_e^{-1}\int_e|[\![v]\!]|^2\,\mathrm{d}\boldsymbol{s}\right]^{1/2} \\ &\leqslant C|w|_{1,h}|v|_*.\end{aligned} \tag{5.23}$$

上式最后一个不等号成立是由于逆不等式 [118, 定理 3.4.1]

$$|w|_{H^2(K)}\leqslant Ch_K^{-1}|w|_{H^1(K)},\quad \forall w\in\mathbb{P}_m(K). \tag{5.24}$$

类似的还有

$$\int_{\Gamma^0}(A[\![w]\!])\cdot\{\nabla_h v\}\,\mathrm{d}\boldsymbol{s}\leqslant C|w|_*|v|_{1,h}. \tag{5.25}$$

最后由上述几个不等式就可以得到有界性

$$\mathcal{B}_h^{IP}(w,v)\leqslant C\left\|w\right\|_U\left\|v\right\|_U,\quad \forall w,v\in U_h.$$

5.4.2　强制性

由于 $A\in\mathcal{M}(\alpha,\beta,\Omega)$，因此

$$\int_\Omega(A\nabla_h v)\cdot\nabla_h v\,\mathrm{d}\boldsymbol{x}\geqslant\alpha|v|^2_{1,h},\quad\forall v\in U_h.$$

又因为(5.23)和(5.25)，可得

$$\mathcal{B}_h^{IP}(v,v)\geqslant\alpha|v|^2_{1,h}+\eta_0|v|^2_*-C|v|_{1,h}|v|_*,\quad\forall v\in U_h,$$

其中 $\eta_0:=\min\limits_{e\in\Gamma^0}\eta_e$. 于是由代数平均不等式 $(a^2+b^2\geqslant 2ab)$，当 η_0 充分大时强制性成立，即

$$\mathcal{B}^{IP}(v,v)\geqslant C_s\left\|v\right\|_U^2,\quad\forall v\in U_h,$$

其中 $C_s>0$ 为常数.

5.4.3 收敛性

接下来考虑由(5.14)得到的弱解的收敛性. 由重构算子的定义以及定理3.4，可得到下面一系列的误差估计.

引理 5.4. 如果 $\forall g \in H^{m+1}(\Omega)$，那么存在常数 C 使得

$$\|g - \mathcal{R}_m g\|_{L^2(\Omega)} \leqslant C(1 + \Lambda_m) h^{m+1} |g|_{H^{m+1}(\Omega)}. \tag{5.26}$$

证明. 由标准的插值理论，可知

$$\begin{aligned}
\left\|g - \mathcal{R}_m^K g\right\|_{L^2(K)} &\leqslant |K|^{1/2} \left\|g - \mathcal{R}_m^K g\right\|_{L^\infty(K)} \\
&\leqslant |K|^{1/2}(1 + \Lambda_m) \inf_{p \in \mathbb{P}_m(S(K))} \|g - p\|_{L^\infty(S(K))} \\
&\leqslant C_I(1 + \Lambda_m)|K|^{1/2} h_{S(K)}^{m+1-d/2} |g|_{H^{m+1}(S(K))} \\
&\leqslant \widetilde{C}_I(1 + \Lambda_m) h_K^{m+1} |g|_{H^{m+1}(S(K))}.
\end{aligned}$$

于是就得到

$$\begin{aligned}
\|g - \mathcal{R}_m g\|_{L^2(\Omega)}^2 &= \sum_{K \in \mathcal{T}_h} \left\|g - \mathcal{R}_m^K g\right\|_{L^2(K)}^2 \\
&\leqslant \sum_{K \in \mathcal{T}_h} \widetilde{C}_I^2 (1 + \Lambda_m)^2 h_K^{2(m+1)} |g|_{H^{m+1}(S(K))}^2 \\
&\leqslant C(1 + \Lambda_m)^2 h^{2(m+1)} |g|_{H^{m+1}(\Omega)}^2.
\end{aligned}$$

□

由逆不等式(5.24)可得到如下引理

引理 5.5. 如果 $\forall g \in H^{m+1}(\Omega)$，那么有

$$|g - \mathcal{R}_m g|_{1,h} \leqslant C(1 + \Lambda_m) h^m |g|_{H^{m+1}(\Omega)}. \tag{5.27}$$

证明. 用 Π_m 表示标准的 Lagrange 插值算子. 结合逆不等式(5.24)，就有

$$\begin{aligned}\left\|\nabla(g-\mathcal{R}_m^K g)\right\|_{L^2(K)} &\leqslant \|\nabla(g-\Pi_m g)\|_{L^2(K)} + \left\|\nabla(\Pi_m g-\mathcal{R}_m^K g)\right\|_{L^2(K)}\\ &\leqslant \|\nabla(g-\Pi_m g)\|_{L^2(K)} + C_I h_K^{-1}\left\|\Pi_m g-\mathcal{R}_m^K g\right\|_{L^2(K)}\\ &\leqslant \|\nabla(g-\Pi_m g)\|_{L^2(K)} + C_I h_K^{-1}\|g-\Pi_m g\|_{L^2(K)}\\ &\quad + C_I h_K^{-1}\left\|g-\mathcal{R}_m^K g\right\|_{L^2(K)}\\ &\leqslant C_0 h_K^m |g|_{H^{m+1}(K)} + \widetilde{C}_I \Lambda_m h_K^m |g|_{H^{m+1}(S(K))}.\end{aligned}$$

这就推出了 (5.27). □

接下来，由插值结果 (5.26)和 (5.27)，有如下估计

引理 5.6. 对于 $\forall g \in H^{m+1}(\Omega)$，都有

$$|g-\mathcal{R}_m g|_* \leqslant C(1+\Lambda_m)h^m |g|_{H^{m+1}(\Omega)}. \tag{5.28}$$

证明. 事实上，注意到对每个 $e \in \Gamma^0$，设 $e = \overline{K_1}\bigcap\overline{K_2}$，并应用 [9, (2.4) 式] 可得

$$\begin{aligned}\|[\![g-\mathcal{R}_m g]\!]\|_{L^2(e)}^2 &= \|\left(g-\mathcal{R}_m^{K_1}g\right)-\left(g-\mathcal{R}_m^{K_2}g\right)\|_{L^2(e)}^2\\ &\leqslant 2\left(\|g-\mathcal{R}_m^{K_1}g\|_{L^2(e)}^2 + \|g-\mathcal{R}_m^{K_2}g\|_{L^2(e)}^2\right)\\ &\leqslant C\left(h_{K_1}^{-1}\|g-\mathcal{R}_m^{K_1}g\|_{L^2(K_1)}^2 + h_{K_1}\|\nabla(g-\mathcal{R}_m^{K_1}g)\|_{L^2(K_1)}^2\right.\\ &\quad \left.+h_{K_2}^{-1}\|g-\mathcal{R}_m^{K_2}g\|_{L^2(K_2)}^2 + h_{K_2}\|\nabla(g-\mathcal{R}_m^{K_2}g)\|_{L^2(K_2)}^2\right).\end{aligned}$$

而如果 $e \subset \partial\Omega\bigcap\overline{K}$，则有

$$\|[\![g-\mathcal{R}_m g]\!]\|_{L^2(e)}^2 \leqslant C\left(h_K^{-1}\|g-\mathcal{R}_m^K g\|_{L^2(K)}^2 + h_K\|\nabla(g-\mathcal{R}_m^K g)\|_{L^2(K)}^2\right).$$

于是由引理 5.4和引理5.5，就可以得到 (5.28). □

注. 上面的插值结果并没有用到逆假设(3.1).

综合引理5.4，引理5.5 和引理5.6，在 $\|\cdot\|_U$ 范数的意义下有如下估计:

引理 5.7. 如果假设A成立，并且定理*3.6* 和定理*3.8*的条件成立，那么对于 $\forall g \in H^{m+1}(\Omega)$，存在 $g_h \in U_h$ 使得

$$\|g-g_h\|_U \leqslant Ch^m |g|_{H^{m+1}(\Omega)}. \tag{5.29}$$

证明. 由(5.26)，(5.27)和(5.28)有

$$\|g - \mathcal{R}_m g\|_U \leqslant Ch^m |g|_{H^{m+1}(\Omega)}.$$

于是引理得证. □

定理 5.8. 令 $u \in H^{m+1}(\Omega)$ 为(5.1)的解，$u_h \in U_h$ 为离散变分问题(5.4)的解. 如果假设A成立，且定理*3.6* 和定理*3.8*的条件成立，那么有

$$\|u - u_h\|_U \leqslant Ch^m |u|_{H^{m+1}(\Omega)}, \tag{5.30}$$

$$\|u - u_h\|_{L^2(\Omega)} \leqslant Ch^{m+1} |u|_{H^{m+1}(\Omega)}. \tag{5.31}$$

证明. 由 Céa's 引理和引理5.7，立刻有

$$\|u - u_h\|_U \leqslant C \inf_{v_h \in U_h} \|u - v_h\|_U \leqslant Ch^m |u|_{H^{m+1}(\Omega)}.$$

而(5.31)中的结果可以由 Aubin-Nistche 技术得到. 令 w 为如下对偶问题的解.

$$\begin{aligned} -\nabla \cdot (A^{\mathrm{T}}(\boldsymbol{x}) \nabla w(\boldsymbol{x})) &= u - u_h, \quad \text{in } \Omega, \\ n(\boldsymbol{x})^{\mathrm{T}} A^{\mathrm{T}}(\boldsymbol{x}) \nabla w(\boldsymbol{x}) &= 0, \quad \text{on } \partial\Omega. \end{aligned}$$

由 $u - u_h \in L^2(\Omega)$ 可知 $w \in H^2(\Omega)$ 以及 $|w|_{H^2(\Omega)} \leqslant C\|u - u_h\|_{L^2(\Omega)}$. 于是对上述方程取 $u - u_h$ 作试探函数并在 Ω 上做积分，可得

$$\|u - u_h\|^2_{L^2(\Omega)} = \mathcal{B}_h^{IP}(u - u_h, w).$$

注意到对于任意的 $w_h \in U_h$ 有

$$\mathcal{B}_h^{IP}(u - u_h, w_h) = 0,$$

因此，令 $w_I := \mathcal{R}_m w$，有

$$\|u - u_h\|^2_{L^2(\Omega)} = \mathcal{B}_h^{IP}(u - u_h, w - w_I).$$

由双线性算子 $\mathcal{B}_h^{IP}$ 的有界性，并在(5.29)中令 $m = 1$，就有

$$\begin{aligned} \|u - u_h\|^2_{L^2(\Omega)} &\leqslant C\|u - u_h\|_U \|w - w_I\|_U \\ &\leqslant Ch\|u - u_h\|_U |w|_{H^2(\Omega)} \\ &\leqslant Ch\|u - u_h\|_U \|u - u_h\|_{L^2(\Omega)}, \end{aligned}$$

于是给出了

$$\|u-u_h\|_{L^2(\Omega)} \leqslant Ch\|u-u_h\|_U \leqslant Ch^{m+1}|u|_{H^{m+1}(\Omega)}.$$

□

5.5　数值算例

本节将用一系列的数值算例来说明上述方法的数值特性. 这些算例将验证定理5.8中的收敛性结果，并展示本章所提出的方法在不同的多边形网格上的适用性. 最后还将考查采样点选取的灵活性问题.

例 5.1　在单位正方形 $[0,1]^2$ 上考虑 Neumann 边值问题(5.1)，并且将精确解取为

$$u(\boldsymbol{x}) = \sin(2\pi x_1)\sin(4\pi x_2).$$

系数矩阵设为单位矩阵 $A(\boldsymbol{x}) = I$.

考虑两种拟一致多边形网格，如图5.5所示. 第一个网格是有三角形单元和四边形单元混合组成 (图5.5(a)). 这些网格是由 *gmsh* [1]软件生成. 第二组则是由六边形单元构成的网格 (图 5.5(b)). Brezzi, Lipnikov 和 Simoncini 在 [26] 中也使用了一套类似的网格. 与他们稍有不同的是，本文是先生成规则的结构化六边形网格，然后对区域内部的所有网格节点 $\boldsymbol{x}_\ell = (x_{\ell,1}, x_{\ell,2})$ 做如下的调整

$$\begin{aligned} x_{\ell,1} &= x_{\ell,1} + \mu\sin(2\pi x_{\ell,1})\sin(2\pi x_{\ell,2}),\\ x_{\ell,2} &= x_{\ell,2} + \mu\sin(2\pi x_{\ell,2})\sin(2\pi x_{\ell,1}), \end{aligned}$$

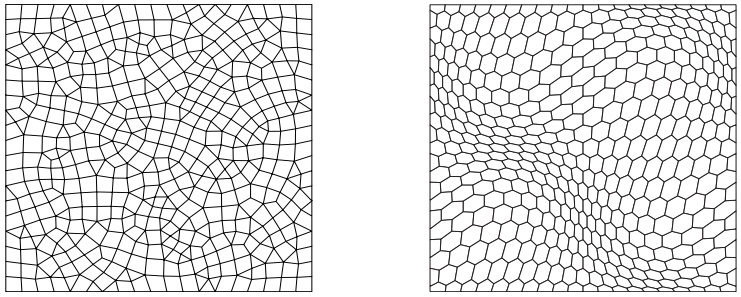
其中 $\mu = 0.075$ 是一个可调的参数. 在所有单元中将每个单元的重心取为采样点.

在两组例子中，离散步长为 $h = 1/n$. 表 5.3和表5.4中，U 和L^2 分别表示衡量误差的范数 $\|\cdot\|_U$ 和 $\|\cdot\|_{L^2}$. 本节所关注的是在用不同阶数 $(m = 1,2,3,4,5,6)$ 的多项式重构时新方法的收敛性表现如何. 表中的结果显示出，数值解在 U-范数下 m 阶收敛，在 L^2 范数下 $m+1$ 阶收敛，而这和(5.30)、(5.31) 中的结果非常吻合.

例 5.2　考虑区域 $[0,1]^2$ 上的 Neumann 边值问题(5.1)，精确解为

$$u(\boldsymbol{x}) = x_1^2 x_2 + \sin(2\pi(x_1+x_2))\sin(2\pi x_2),$$

[1] http://geuz.org/gmsh/



(a) 混合单元网格 (b) 六边形单元网格

图 *5.5: 例5.1：拟一致多边形网格.*

m	$\|\cdot\|$	$n=20$	$n=40$		$n=80$		$n=160$	
		误差	误差	阶	误差	阶	误差	阶
1	U	3.748E-00	1.606E-00	1.22	6.420E-01	1.32	2.754E-01	1.22
	L^2	2.270E-01	7.980E-02	1.51	2.130E-02	1.91	5.241E-03	2.02
2	U	4.718E-00	1.075E-00	2.13	1.966E-01	2.45	4.156E-02	2.24
	L^2	3.115E-01	5.606E-02	2.47	5.290E-03	3.41	4.484E-04	3.56
3	U	8.197E-01	6.891E-02	3.57	7.847E-03	3.13	9.183E-04	3.10
	L^2	3.746E-02	1.292E-03	4.86	6.220E-05	4.38	3.496E-06	4.15
4	U	3.471E-01	1.185E-02	4.87	7.590E-04	3.96	4.244E-05	4.16
	L^2	1.302E-02	1.721E-04	6.24	5.416E-06	4.99	1.436E-07	5.23
5	U	1.563E-01	3.264E-03	5.58	1.038E-04	4.97	3.145E-06	5.04
	L^2	4.921E-03	4.697E-05	6.71	7.798E-07	5.91	1.189E-08	6.04
6	U	6.445E-02	6.372E-04	6.66	1.094E-05	5.86	1.524E-07	6.17
	L^2	1.871E-03	7.723E-06	7.92	6.883E-08	6.81	4.787E-10	7.17

表 *5.3: 例5.1：三角形单元和四边形单元混合的网格上的数值结果.*

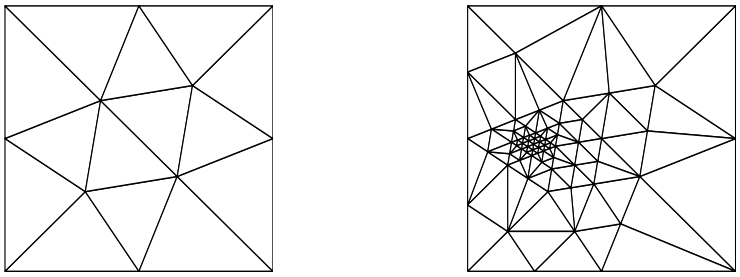
m	$\|\cdot\|$	$n=20$	$n=40$		$n=80$		$n=160$	
		误差	误差	阶	误差	阶	误差	阶
1	U	4.512E-00	1.983E-00	1.19	7.288E-01	1.44	2.694E-01	1.44
	L^2	3.114E-01	1.209E-01	1.34	3.482E-02	1.80	9.239E-03	1.91
2	U	5.935E-00	3.038E-00	0.97	7.764E-01	1.97	1.363E-01	2.51
	L^2	4.157E-01	1.992E-01	1.06	3.880E-02	2.36	4.249E-03	3.19
3	U	2.717E-00	3.284E-01	3.05	2.024E-02	4.02	2.014E-03	3.33
	L^2	1.576E-01	1.287E-02	3.61	2.764E-04	5.54	6.340E-06	5.45
4	U	1.935E-00	1.833E-01	3.40	8.316E-03	4.46	4.240E-04	4.29
	L^2	1.088E-01	7.361E-03	3.89	1.624E-04	5.50	2.987E-06	5.76
5	U	1.096E-00	3.985E-02	4.78	6.498E-04	5.94	1.736E-05	5.23
	L^2	5.449E-02	1.162E-03	5.55	6.398E-06	7.50	3.195E-08	7.65
6	U	5.037E-01	1.496E-02	5.07	1.827E-04	6.36	2.724E-06	6.07
	L^2	2.288E-02	4.391E-04	5.70	2.088E-06	7.72	1.272E-08	7.36

表 *5.4: 例5.1：六边形单元网格上数值结果.*

为了更一般化，将系数矩阵设定为

$$A(\boldsymbol{x}) = \begin{pmatrix} (x_1+1)^2 & -x_1x_2 \\ -x_1x_2 & (x_2+1)^2 \end{pmatrix}.$$

在定理5.8的证明过程中，为了保证重构常数的一致有界性而采用了逆假设(3.1). 但实际上，如果抛开重构常数的大小这一点，整个证明过程其实并没有要求网格一定是拟一致的. 为了验证这一点，考虑一组自适应三角形网格，见图5.6(b). 这些网格是从拟一致的三角形网格 (图5.6(a)) 出发，然后以单元中心到点 $[0.25, 0.5]$ 的距离作为指示子对网格进行加密而得到的. 在每个自适应网格中，最小单元的尺寸大致是最大单元的尺寸的 $1/256$. 采样点仍选在每个单元的重心上. 从表5.5中的结果来看，本章所提出的方法在自适应网格上也能达到最优的收敛阶.



(a) 拟一致网格　　(b) 自适应网格

图 *5.6: 例5.2: 拟一致三角形网格和自适应加密网格.*

例 5.3　在区域 $[0,1]^2$ 上考虑 Neumann 边值问题(5.1)，取精确解为

$$u(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_1-0.5)^2+(x_2-0.5)^2}{2}\right),$$

并且令系数矩阵为

$$A(\boldsymbol{x}) = \begin{pmatrix} 3+\cos(2\pi x_1) & x_1-x_2 \\ x_1-x_2 & 3-\sin(2\pi x_2) \end{pmatrix}.$$

在之前的例子中，采样点都取在了单元的重心上. 而这个算例将对算法中采样点位置的敏感性进行测试. 采用如图5.7(b)所示的三角形剖分的网格. 因此这个例子中不再将采样点选为单元的重心，而是在每个单元内随机选取一个采样点.

m	$\|\cdot\|$	$n=20$	$n=40$		$n=80$		$n=160$	
		误差	误差	阶	误差	阶	误差	阶
1	U	3.507E-00	1.788E-00	0.97	7.296E-01	1.29	3.243E-01	1.17
	L^2	2.442E-01	1.006E-01	1.28	2.436E-02	2.05	4.983E-03	2.29
2	U	3.683E-00	2.099E-00	0.81	8.038E-01	1.38	1.498E-01	2.42
	L^2	2.548E-01	1.288E-01	0.98	3.871E-02	1.73	4.349E-03	3.15
3	U	2.696E-00	6.828E-01	1.98	1.190E-01	2.52	1.240E-02	3.26
	L^2	1.697E-01	3.393E-02	2.32	3.694E-03	3.20	1.363E-04	4.76
4	U	2.127E-00	3.004E-01	2.82	3.188E-02	3.24	2.087E-03	3.93
	L^2	1.234E-01	1.156E-02	3.42	6.846E-04	4.08	1.779E-05	5.27
5	U	1.786E-00	2.599E-01	2.78	1.241E-02	4.39	3.034E-04	5.35
	L^2	1.022E-01	1.200E-02	3.09	2.501E-04	5.58	1.834E-06	7.09
6	U	1.529E-00	8.988E-02	4.09	2.730E-03	5.04	4.488E-05	5.93
	L^2	8.278E-02	2.791E-03	4.89	5.021E-05	5.80	3.410E-07	7.20

表 *5.5: 例5.2：自适应加密网格上的数值结果.*

具体地，设每个单元 K 的三个顶点为 $\boldsymbol{x}_K^{(i)}, i=1,2,3$，选取两个服从 [0,1] 上的均匀分布的随机数，分别记为 ζ_K 和 ξ_K，那么采样点 $\boldsymbol{x}_K$ 取为

$$\boldsymbol{x}_K = \zeta_K\xi_K\boldsymbol{x}_K^{(1)} + \xi_K(1-\zeta_K)\boldsymbol{x}_K^{(2)} + (1-\xi_K)\boldsymbol{x}_K^{(3)}.$$

这样可以保证采样点位于三角形单元内部，如图5.7所示. 相应的数值结果展示在表 5.6中，从表中可以看出无论是 $\|\cdot\|_U$ 还是 $\|\cdot\|_{L^2}$ 都保持了最优的收敛阶，这就意味着算法对于采样点位置的选择提供了较高的自由度. 在实际问题中，完全可以根据实际需要而自由决定采样点的位置.

(a) 采样点位于单元重心　　(b) 随机选取采样点

图 *5.7: 例5.3：采样点的不同选取方法的对比.*

m	$\|\cdot\|$	$n=20$	$n=40$		$n=80$		$n=160$	
		误差	误差	阶	误差	阶	误差	阶
1	U	1.181e-00	5.675e-01	1.06	2.825e-01	1.01	1.410e-01	1.00
	L^2	3.730e-02	8.485e-03	2.14	2.067e-03	2.04	5.104e-04	2.02
2	U	1.037e-00	1.955e-01	2.41	4.053e-02	2.27	9.458e-03	2.10
	L^2	3.805e-02	3.620e-03	3.39	3.029e-04	3.58	2.683e-05	3.50
3	U	2.884e-01	3.450e-02	3.06	4.386e-03	2.98	5.516e-04	2.99
	L^2	7.600e-03	2.790e-04	4.77	1.184e-05	4.56	6.366e-07	4.22
4	U	1.407e-01	9.388e-03	3.91	6.038e-04	3.96	3.860e-05	3.97
	L^2	2.741e-03	6.438e-05	5.41	2.132e-06	4.92	6.680e-08	5.00
5	U	1.893e-02	7.187e-04	4.72	2.342e-05	4.94	7.360e-07	4.99
	L^2	2.016e-04	3.572e-06	5.82	5.428e-08	6.04	8.169e-10	6.05
6	U	9.107e-03	1.247e-04	6.19	2.053e-06	5.92	3.332e-08	5.95
	L^2	1.409e-04	8.121e-07	7.44	6.660e-09	6.93	5.380e-11	6.95

表 *5.6:* 例*5.3*：随机选取采样点的数值结果.

5.6　小结

本章提出了一个基于最小二乘重构的间断 Galerkin 方法. 该方法选取的逼近空间是分片多项式空间的子空间，并且每一个基函数在各个单元上的限制都是通过局部的最小二乘重构来得到的. 这样选取的逼近空间的维度总是和网格单元的数量相等，而且在保持自由度不变的情况下，只靠改变重构的阶数就能改变算法的收敛阶. 另一方面，该方法没有采用传统的形函数空间，因此适用于任意的多边形网格，具有很大的灵活性. 本章的理论分析部分证明了该方法的稳定性并给出了收敛阶. 在数值实验部分则在多种不同的网格上测试了一系列的算例，并对考察了采样点的随机分布对结果的影响. 结果表明本章所提出的方法是一种求解椭圆型方程的有效、稳定的高阶数值方法。

总结与展望

本文中的内容是我博士期间的主要研究工作，主要探讨了如何通过最小二乘重构来实现椭圆型方程的高阶方法. 文中定义的最小二乘重构算子可以看作是离散最小二乘方法的泛函化表述. 为了能够对重构常数给出一个确定的上界，本文用到了两个关键假设：一个是假设单元模板为凸多边形，另一个是逆假设成立. 然而在实际计算中，这两个假设看起来都不是必须的. 本文通过数值算例展示了这一点.

基于最小二乘重构，本文针对两个不同类型的椭圆型方程分别提出了不同的高阶方法. 第一个方法是在异质多尺度有限元方法的框架下，通过在网格节点上求解单胞问题并用最小二乘重构来恢复单元上的有效矩阵信息，从而得到高阶的宏观求解器，但算法的计算复杂度仍和线性求解器相当. 第二个方法是利用每个单元周围的局部信息来重构出分片多项式的基函数和逼近空间，并应用间断 Galerkin 方法在经典的二阶椭圆型方程上的变分形式来得到数值解. 这个方法可以在不改变逼近空间自由度的前提下达到高阶收敛，同时还可适用于任意的多边形单元网格.

进一步的工作主要包括以下几个方面：

- 改进重构算子稳定性的证明. 如果能够去掉上面提到的两个假设，那么最小二乘重构的稳定性就真正有了保障.

- 对于单元模板，考虑不同的构造方式对重构结果带来的影响. 进一步可考虑对特殊问题单独定制合适的单元模板.

- 对于本文所提出的重构有效系数的异质多尺度方法，更全面、细致地测试三维情形下的例子，同时将该方法应用到实际问题中，比如复合材料的优化设计等. 此外，该方法的思想还可以应用到其他的多尺度问题中.

- 对于重构基函数的间断 Galerkin 方法，考虑通过实际中的科学与工程计算

问题来对其进行进一步的检验与测试. 此外，还可以考虑试探函数是分片常数的情形，该情形下将得到类似有限体积的格式，在计算流体力学、对流扩散方程等问题中具有重要意义.

参考文献

[1] A. Abdulle. On a priori error analysis of fully discrete heterogeneous multiscale FEM. *Multiscale Model. Simul.*, 4:447–459, 2005.

[2] A. Abdulle. The finite element heterogeneous multiscale method: a computational strategy for multiscale PDEs. *GAKUTO Internat. Ser. Math. Sci. and Appl.*, 31:133–181, 2009.

[3] A. Abdulle and Y. Bai. Reduced basis finite element heterogeneous multiscale method for high-order discretizations of elliptic homogenization problems. *J. Comput. Phys.*, 231(21):7014–7036, 2012.

[4] A. Abdulle, W. E, B. Engquist, and E. Vanden-Eijnden. The heterogeneous multiscale method. *Acta Numer.*, 21:1–87, 2012.

[5] A. Abdulle and B. Engquist. Finite element heterogeneous multiscale methods with near optimal computational complexity. *Multiscale Model. Simul.*, 6(4):1059–1084, 2007.

[6] A. Abdulle and A. Nonnenmacher. A short and versatile finite element multiscale code for homogenization problems. *Comput. Methods Appl. Mech. Engrg.*, 198(37):2839–2859, 2009.

[7] R.A. Adams and J.J.F. Fournier. *Sobolev Spaces.* 2nd eds., Academic Press, New York, 2003.

[8] R.E. Alcouffe, A. Brandt, J.E. Dendy, and J.W. Painter. The multi-grid method for the diffusion equation with strongly discontinuous coefficients. *SIAM J. Sci. Statist. Comput.*, 2(4):430–454, 1981.

[9] D.N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19(4):742–760, 1982.

[10] D.N. Arnold, F. Brezzi, B. Cockburn, and D. Marini. Discontinuous Galerkin methods for elliptic problems. In *Discontinuous Galerkin Methods*, pages 89–101. Springer, 2000.

[11] D.N. Arnold, F. Brezzi, B. Cockburn, and L.D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39:1749–1779, 2002.

[12] I. Babuška and M. Zlámal. Nonconforming elements in the finite element method with penalty. *SIAM J. Numer. Anal.*, 10(5):863–875, 1973.

[13] F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *J. Comput. Phys.*, 131(2):267–279, 1997.

[14] C.E. Baumann and J.T. Oden. A discontinuous hp finite element method for convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 175(3):311–341, 1999.

[15] C.E. Baumann and J.T. Oden. A discontinuous hp finite element method for the Euler and Navier-Stokes equations. *Internat. J. Numer. Methods Fluids*, 31(1):79–95, 1999.

[16] A. Bensoussan, J.L. Lions, and G.C. Papanicolaou. *Asymptotic Analysis for Periodic Structures*. North-Holland, Amsterdam, 1978.

[17] M. Benzi. Preconditioning techniques for large linear systems: a survey. *J. Comput. Phys.*, 182(2):418–477, 2002.

[18] A. Berger, R. Scott, and G. Strang. Approximate boundary conditions in the finite element method. *Symposia Mathematica*, X:295–313, 1972.

[19] J.T. Betts. An improved penalty function method for solving constrained parameter optimization problems. *J. Optim. Theory Appl.*, 16(1-2):1–24, 1975.

[20] J.T. Betts. Solving the nonlinear least square problem: Application of a general method. *J. Optim. Theory Appl.*, 18(4):469–483, 1976.

[21] Å. Björck. Methods for sparse linear least squares problems. In *Sparse Matrix Computations*, pages 177–199. Academic Press New York, 1976.

[22] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, 1996.

[23] Å. Björck and C.C. Paige. Solution of augmented linear systems using orthogonal factorizations. *BIT Numer. Math.*, 34(1):1–24, 1994.

[24] P.B. Bochev and M.D. Gunzburger. Finite element methods of least-squares type. *SIAM Review*, 40(4):789–837, 1998.

[25] F. Brezzi, L.P. Franca, T.J.R. Hughes, and A. Russo. $b = \int g$. *Comput. Methods Appl. Mech. Engrg.*, 145:329–339, 1996.

[26] F. Brezzi, K. Lipnikov, and V. Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Math. Models Methods Appl. Sci.*, 15(10):1533–1551, 2005.

[27] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous finite elements for diffusion problems. *Atti Convegno in onore di F. Brioschi (Milano 1997), Istituto Lombardo, Accademia di Scienze e Lettere*, pages 197–217, 1999.

[28] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous Galerkin approximations for elliptic problems. *Numer. Methods Partial Differential Equations*, 16(4):365–378, 2000.

[29] P. Castillo. An optimal estimate for the local discontinuous Galerkin method. In *Discontinuous Galerkin Methods*, pages 285–290. Springer, 2000.

[30] P. Castillo. Performance of discontinuous Galerkin methods for elliptic PDEs. *SIAM J. Sci. Comput.*, 24(2):524–547, 2002.

[31] P. Castillo, B. Cockburn, D. Schötzau, and C. Schwab. Optimal a priori error estimates for the hp-version of the local discontinuous Galerkin method for convection-diffusion problems. *Math. Comp.*, 71(238):455–478, 2002.

[32] P.G. Ciarlet. *The Finite Element Method for the Elliptic Problems*. North-Holland, Amsterdam, 1978.

[33] D. Cioranescu and P. Donato. *An Introduction to Homogenization*. Oxford University Press, Oxford, 1999.

[34] B. Cockburn and C. Dawson. Some extensions of the local discontinuous Galerkin method for convection-diffusion equations in multidimensions. In *The Proceedings of the Conference on the Mathematics of Finite Elements and Applications: MAFELAP X. Elsevier*, pages 225–238, 2000.

[35] B. Cockburn, S. Hou, and C.W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. *Math. Comp.*, 54(190):545–581, 1990.

[36] B. Cockburn, G.E. Karniadakis, and C.W. Shu. *The Development of Discontinuous Galerkin Methods*. Springer, 2000.

[37] B. Cockburn, S.Y. Lin, and C.W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one-dimensional systems. *J. Comput. Phys.*, 84(1):90–113, 1989.

[38] B. Cockburn and C.W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: General framework. *Math. Comp.*, 52(186):411–435, 1989.

[39] B. Cockburn and C.W. Shu. The Runge-Kutta local projection P^1-discontinuous Galerkin method for scalar conservation laws. *RAIRO Modél. Math. Anal. Numér.*, 25:337–361, 1991.

[40] B. Cockburn and C.W. Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM J. Numer. Anal.*, 35(6):2440–2463, 1998.

[41] B. Cockburn and C.W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems. *J. Comput. Phys.*, 141(2):199–224, 1998.

[42] B. Cockburn and C.W. Shu. Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. *J. Sci. Comput.*, 16(3):173–261, 2001.

[43] D. Coppersmith and T.J. Rivlin. The growth of polynomials bounded at equally spaced points. *SIAM J. Math. Anal.*, 23:970–983, 1992.

[44] J. Dennis, E. John, and R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, 1996.

[45] M. Dorobantu and B. Engquist. Wavelet-based numerical homogenization. *SIAM J. Numer. Anal.*, 35(2):540–559, 1998.

[46] J. Douglas and T. Dupont. Interior penalty procedures for elliptic and parabolic Galerkin methods. In *Computing Methods in Applied Sciences*, pages 207–216. Springer, 1976.

[47] R. Du and P.B. Ming. Convergence of the heterogeneous multiscale finite element method for elliptic problem with nonsmooth microstructures. *Multiscale Model. Simul.*, 8:1770–1783, 2010.

[48] R. Du and P.B. Ming. Heterogeneous multiscale finite element method with novel numerical integration schemes. *Commun. Math. Sci.*, 8:863–885, 2010.

[49] P. Dvorak. New element lops time off CFD simulations. *Machine Design*, 78(5):154–155, 2006.

[50] W. E and B. Engquist. The heterogeneous multi-scale methods. *Commun. Math. Sci.*, 1:87–132, 2003.

[51] W. E, P.B. Ming, and P.W. Zhang. Analysis of the heterogeneous multiscale method for elliptic homogenization problems. *J. Amer. Math. Soc.*, 18:121–156, 2005.

[52] Y.R. Efendiev, T.Y. Hou, and X.H. Wu. Convergence of a nonconforming multiscale finite element method. *SIAM J. Numer. Anal.*, 37(3):888–910, 2000.

[53] C. Farhat, I. Harari, and L.P. Franca. The discontinuous enrichment method. *Comput. Methods Appl. Mech. Engrg.*, 190(48):6455–6479, 2001.

[54] P. Finsler and H. Hadwiger. Einige relationen im dreieck. *Comment. Math.*, 10 (1):316–326, 1937.

[55] J. Fish and Z. Yuan. Multiscale enrichment based on partition of unity. *Internat. J. Numer. Methods Engrg.*, 62(10):1341–1359, 2005.

[56] J.E. Flaherty, R.M. Loy, M.S. Shephard, and J.D. Teresco. Software for the parallel adaptive solution of conservation laws by discontinuous Galerkin methods. In *Discontinuous Galerkin Methods*, pages 113–123. Springer, 2000.

[57] R. Fletcher. Modified Marquardt subroutine for nonlinear least squares. Technical report, Atomic Energy Research Establishment, Harwell, England, 1971.

[58] W.M. Gentleman. Error analysis of QR decompositions by Givens transformations. *Linear Algebra Appl.*, 10(3):189–197, 1975.

[59] G.H. Golub. Numerical methods for solving linear least squares problems. *Numer. Math.*, 7(3):206–216, 1965.

[60] G.H. Golub and J.H. Wilkinson. Note on the iterative refinement of least squares solution. *Numer. Math.*, 9(2):139–148, 1966.

[61] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73(2):325–348, 1987.

[62] H.O. Hartley. The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. *Technometrics*, 3(2):269–280, 1961.

[63] S. Hazanov and C. Huet. Order relationships for boundary conditions effect in heterogeneous bodies smaller than representative volume. *J. Mech. Phys. Solids*, 42:1995–2011, 1994.

[64] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Bur. Standards*, 49:409–436, 1952.

[65] N.J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 2002.

[66] T. Hou, X.H. Wu, and Z.Q. Cai. Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients. *Math. Comp.*, 68(227):913–943, 1999.

[67] T.Y. Hou and X.H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169–189, 1997.

[68] C.Q. Hu and C.W. Shu. Weighted essentially non-oscillatory schemes on triangular meshes. *J. Comput. Phys.*, 150(1):97–127, 1999.

[69] T.J. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Comput. Methods Appl. Mech. Engrg.*, 127(1):387–401, 1995.

[70] C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic problems. *Comput. Methods Appl. Mech. Engrg.*, 45(1):285–312, 1984.

[71] A. Jones. Spiral – A new algorithm for non-linear parameter estimation using least squares. *Computer Journal*, 13(3):301–308, 1970.

[72] G. Karniadakis and S. Sherwin. *Spectral/hp Element Methods for Computational Fluid Dynamics*. Oxford University Press, 2013.

[73] V.A. Kondrat'ev. Boundary value problems for elliptic equations in domains with conical or angular points. *Trans. Moscow Math. Soc.*, 16:227–313, 1967.

[74] T. Kubota. Einige Ungleischheitsbezichungen über Eilinien und Eiflähen. *Sci. Rep. of the Tŏhoku Univ. Ser. (1)*, 12:45–65, 1923.

[75] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems*. SIAM, 1974.

[76] P. Lesaint and P.A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*, number 33, pages 89–123. C.A. deBoor(Ed), Academic Press, 1974.

[77] K. Levenberg. A method for the solution of certain problems in least squares. *Quart. Appl. Math.*, 2:164–168, 1944.

[78] R. Li, P.B. Ming, and F.Y. Tang. An efficient high order heterogeneous multiscale method for elliptic problems. *Multiscale Model. Simul.*, 10:259–283, 2012.

[79] K.A. Lie, S. Krogstad, I.S. Ligaarden, J.R. Natvig, H.M. Nilsen, and B. Skaflestad. Open-source MATLAB implementation of consistent discretisations on complex grids. *Comput. Geosci.*, 16(2):297–322, 2012.

[80] J.L. Lions. Problèmes aux limites non homogènes à donées irrégulières: Une méthode d' approximation. *Numerical Analysis of Partial Differential Equations (C.I.M.E. 2 Ciclo, Ispra, 1967)*, pages 283–292, 1968.

[81] B. Lipman, F. John, and M. Schechter. *Partial Differential Equations*. Intersciences Publishers, 1964.

[82] Y. Liu and Y.T. Zhang. A robust reconstruction for unstructured WENO schemes. *J. Sci. Comput.*, 54(2-3):603–621, 2013.

[83] S.L. Lyons, R.R. Parashkevov, and X.H. Wu. A family of H^1-conforming finite element spaces for calculations on 3D grids with pinch-outs. *Numer. Linear Algebra Appl.*, 13(9):789–799, 2006.

[84] T.A. Manteuffel. An incomplete factorization technique for positive definite linear systems. *Math. Comp.*, 34(150):473–497, 1980.

[85] D.W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *J. SIAM*, 11(2):431–441, 1963.

[86] R.R. Meyer and P.M. Roth. Modified damped least squares: an algorithm for non-linear estimation. *IMA J. Appl. Math.*, 9(2):218–233, 1972.

[87] P.B. Ming and X.Y. Yue. Numerical methods for multiscale elliptic problems. *J. Comput. Phys.*, 214(1):421–445, 2006.

[88] N. Munksgaard. Solving sparse symmetric sets of linear equations by preconditioned conjugate gradients. *ACM Trans. Math. Software*, 6(2):206–219, 1980.

[89] W. Niethammer, J.D. Pillis, and R.S. Varga. Convergence of block iterative methods applied to sparse least-squares problems. *Linear Algebra Appl.*, 58:327–341, 1984.

[90] J. Nitsche. Über ein Variationsprinzip zur lösung von Dirichlet-Problemen bei verwendung von teilräumen, die keinen Randbedingungen unterworfen sind. In *Abh. Math. Sem. Univ. Hamburg*, volume 36, pages 9–15. Springer, 1971.

[91] E.E. Osborne. On least squares solutions of linear equations. *J. ACM*, 8(4):628–636, 1961.

[92] P.O. Persson and J. Peraire. Newton-GMRES preconditioning for discontinuous Galerkin discretizations of the Navier-Stokes equations. *SIAM J. Sci. Comput.*, 30(6):2709–2733, 2008.

[93] P. Rabinowitz and N. Richter. Perfectly symmetric two-dimensional integration formulas with minimal numbers of points. *Math. Comput.*, 23:765–779, 1969.

[94] E.A. Rakhmanov. Bounds for polynomials with a unit discrete norm. *Ann. of Math. (2)*, 165:55–88, 2007.

[95] W.H. Reed and T.R. Hill. Triangularmesh methods for the neutron transport equation. *Los Alamos Report LA-UR-73-479*, 1973.

[96] L. Reichel. On polynomial approximation in the uniform norm by the discrete least squares method. *BIT*, 26:349–368, 1986.

[97] G.R. Richter. The discontinuous Galerkin method with diffusion. *Math. Comp.*, 58(198):631–643, 1992.

[98] B. Rivière, M.F. Wheeler, and V. Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. *Comput. Geosci.*, 3(3-4):337–360, 1999.

[99] G. Sangalli. Capturing small scales in elliptic problems using a residual-free bubbles finite element method. *Multiscale Model. Simul.*, 1(3):485–503, 2003.

[100] M.A. Saunders. *Sparse Least Squares by Conjugate Gradients: A Comparison of Preconditioning Methods*. Systems Optimization Laboratory, Department of Operations Research, Stanford University, 1979.

[101] E. Schmidt. Über die auflösung linearer Gleichungen mit unendlich vielen Unbekannten. *Rend. Circ. Mat. Palermo. Ser.*, 25(1):53–77, 1908.

[102] H. Schwarz. Die Methode der konjugierten Gradienten in der Ausgleichsrechnung. *ZfV*, 95:130–140, 1970.

[103] S.J. Smith. Lebesgue constants in polynomial interpolation. In *Annales Mathematicae et Informaticae*, volume 33, pages 1787–5021. Eszterházy Károly College, Institute of Mathematics and Computer Science, 2006.

[104] Y.Z. Song and O. Goro. Methodes iteratives de type SOR pow resoudre les problemes des moindres cares. *Intern. J. Computer Math.*, 68:99–118, 1998.

[105] L. Tartar. H-convergence. In *Course Peccot, Collége de France. Partially written by F. Murat. Séminaire d'Analyse Fonctionnelle et Numérique de l'Université d'Alger, 1977–1978*. March 1977.

[106] O. Taussky. Note on the condition of matrices. *Math. Comp.*, 4(30):111–112, 1950.

[107] A. Toselli and O. Widlund. *Domain Decomposition Methods: Algorithms and Theory*. Springer, 2005.

[108] P. Šolín, K. Segeth, and I. Doležel. *Higher-Order Finite Element Methods*. Chapman & Hall/CRC, Boca Raton, F.L., 2004.

[109] K. Wang. *Numerical Simulation for 3D Heterogeneous Multiscale Elliptic Equation*. Master thesis, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 2010.

[110] M.F. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.*, 15(1):152–161, 1978.

[111] D.R. Wilhelmsen. A Markov inequality in several dimensions. *J. Approx. Theory*, 11:216–220, 1974.

[112] J.H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, UK, 1965.

[113] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numer. Mathe.*, 94(1):195–202, 2003.

[114] X.Y. Yue and W. E. The local microscale problem in the multiscale modeling of strongly heterogeneous media: effects of boundary conditions and cell size. *J. Comput. Phys.*, 222:556–572, 2007.

[115] Z.M. Zhang and A. Naga. A new finite element gradient recovery method: superconvergence property. *SIAM J. Sci. Comput.*, 26:1192–1213, 2005.

[116] O.C. Zienkiewicz and J.Z. Zhu. The superconvergence patch recovery and a posteriori error estimates. Part 1: The recovery technique. *Internat. J. Numer. Methods Engrg.*, 33:1331–1364, 1992.

[117] 徐树方，张平文. 数值线性代数. 北京大学出版社, 2000.

[118] 石钟慈，王鸣. 有限元方法. 科学出版社, 2010.

在学期间研究成果

- R. Li, P.B. Ming, and F.Y. Tang. *An efficient high order heterogeneous multiscale method for elliptic problems*, published in SAIM Multiscale Modelling and Simulation, 10:259–283, 2012.

- R. Li, P.B. Ming, and F.Y. Tang. *Reconstructed basis discontinuous Galerkin method for elliptic problems*, in preparation, 2015.

致谢

本文是在我的导师李若教授和明平兵研究员的指导下完成的. 多年来，两位老师悉心指导我的学习与科研，不仅为我把握研究方向、开拓研究思路，而且毫无保留地为我创造各种科研条件，细心地答疑解惑. 两位老师深厚的知识基础、广阔的学术视野、敏锐的数学直觉、高超的编程技术，一直是我学习的楷模. 他们严谨创新的治学态度、积极进取的人生追求、真诚宽容的处事方式，都将是我今后工作和生活的准则. 在此向两位老师致以最诚挚的感谢.

其次，感谢北京大学数学科学学院提供的良好的学习和科研环境. 感谢曾经为我传道授业解惑的所有老师，包括张平文教授、汤华中教授、徐树方教授、李治平教授、李铁军教授、王鸣教授、周铁教授、高立教授等，是他们的精彩讲解帮助我在学业的道路上步步前行. 感谢教务的袁燕老师、蔡贤川老师、张婧老师在日常生活中的关心和帮助.

同时，感谢师门的兄弟姐妹们，包括王端师姐、赵伟波师兄、袁钢师兄、邓剑师兄、蔡振宁师兄、罗毅师兄、李蔚明师姐以及吴朔男、王艳莉、周宏升、张涛、刘姝、樊玉伟、李君、陈里等，大家的相互陪伴让生活变得多姿多彩. 感谢吕唐杰和唐浩哲，我们一同度过从本科到博士的时光，相互鼓励，共同努力.

特别感谢我的父母，是他们含辛茹苦地将我养育成人，在我成长的道路上处处伴随着他们的关怀和奉献. 对于他们的付出，我只有在今后用加倍的努力来回报. 最后还要感谢我的妻子常红燕女士，和她在一起的时光是我博士期间最美好的回忆. 在本文的完成期间，她不仅在生活上对我无微不至地照顾，而且每晚认真细致地帮我修改文章，这都给了我莫大的帮助和鼓励. 此生有你是我最大的满足.

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校在 □ 一年 / □ 两年 / □ 三年以后在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：　　日期：　　年　　月　　日