2020 年 2 月 计 算 数 学 第 42 卷第 1 期
Feb., 2020 MATHEMATICA NUMERICA SINICA Vol.42, No.1

自适应稀疏伪谱逼近新方法*

林济铿 袁恺明 申丹枫
(同济大学电子与信息工程学院, 上海 201804)
罗萍萍
(上海电力大学电气工程学院, 上海 200090)
刘阳升
(同济大学电子与信息工程学院, 上海 201804)

摘 要

自适应稀疏伪谱逼近法是广义混沌多项式类方法的最新进展, 相对于其它方法具有计算精度高、速度快的优点. 但它仍存在如下缺点: 1) 终止判据对逼近误差的估计精度偏低; 2) 只适用于单输出问题. 本文提出了适用于多输出问题且具有更高逼近精度的自适应稀疏伪谱逼近新方法. 本文首先提出了新型终止判据及基于此新型终止判据的自适应稀疏伪谱逼近新方法, 并以命题的形式证明了新型终止判据相比于现有终止判据具有更高的估计精度, 从而使基于此的逼近函数精度更接近于预期精度; 进而, 本文基于指标集的统一策略和新型终止判据, 提出了适用于多输出问题的自适应稀疏伪谱逼近新方法, 该方法因能充分利用各输出变量的抽样结果, 具有比将单输出方法直接推广到多输出问题更高的计算效率. 多个算例验证了本文所提出新方法的有效性和正确性.

关键词: 自适应稀疏伪谱逼近法; 终止判据; 逼近误差; 单输出; 多输出
MR (2010) 主题分类: 65D15, 65Y20

1. 引 言

不确定性传递问题 (uncertainty propagation) 是一类研究如何量化由输入不确定性传递到的输出不确定性的问题. 蒙特卡罗类方法、摄动法和矩方程法是三类传统的不确定性传递问题的分析方法, 广义混沌多项式法 (generalized polynomial chaos, gPC) 则是近些年提出的新型分析方法.

蒙特卡罗类方法以蒙特卡罗法 (Monte Carlo method, MCM) [1] 为代表, 其收敛速度与随机变量维数无关, 故可方便地应用于大规模系统, 但因其误差与抽样数的平方根成反比, 当计算精度要求很高时所需的抽样次数很大、计算时间很长. 摄动法 (perturbation method) [2] 是一类非抽样方法, 它在随机变量的均值附近对输入输出随机变量间的函数关系进行泰勒展开, 无需抽样就可快速求出输出变量的期望、方差等各阶矩, 但由于其只能进行低阶泰勒展开 (最多 2 阶, 超过 2 阶时因展开过程过于复杂而无法进行), 所以无法完全反映输出随机变量与输入随机变量的函数关系, 因而当输入随机变量波动范围较大时以及输入输出函数关系的非线性程度较大时, 其所求取的各阶矩存在较大的误差. 矩方程法 (moment equation method) [3] 从表示输入输出函数关系的系统方程出发, 直接推导含有输出随机变量各阶矩的方程组, 然后求解该方程组而得到输出随机变量的各阶矩. 该方法的优点在于能准确给出不超过既定阶次

* 2018 年 6 月 22 日收到.

的各阶矩的表达式, 但由于在绝大多数情况下其所给出的各阶矩的表达式中会出现阶次比既定阶次更高的矩, 为了能够具体算出不超过既定阶次的各阶矩, 就不得不对高于既定阶次的矩进行一定的假设, 相应不可避免地引入了误差.

广义混沌多项式法 [4–6](generalized polynomial chaos, gPC) 是一种基于多项式函数逼近理论的不确定性传递问题的分析方法, 其思想与前面三类方法完全不同. 该类方法的基本思想是采用以输入随机变量为自变量的正交多项式基函数的线性组合逼近输入输出变量间的函数关系, 从而可快速地求解出输出变量的各阶矩和概率密度. 该方法因无需进行误差较大的低阶展开, 而完全克服摄动法因只能进行较低阶次的展开而导致输入随机变量在较大范围波动时, 其相应解的误差较大的缺点. 该方法无需任何假设条件就能准确地求出输出变量的各阶矩, 因而其计算结果的精度和可靠性要明显高于矩方程法. 该方法达到相同计算精度所需要的抽样次数及相关的计算量远小于蒙特卡罗法. 基于如上特点, 该类方法备受数学理论界及工程界的重视 [7–10].

根据基函数系数的计算方法不同, 广义混沌可进一步划分为两类方法: 1) 随机 Galer- kin 法 (stochastic Galerkin method, SGM); 2) 随机配置点法 (stochastic collocation method, SCM).

1) 随机 Galerkin 法 [9]. 该类方法先用由系数待定的正交多项式基函数组成的逼近函数代替输入输出变量间的函数关系, 并将其代入关于输入和输出变量的随机非线性方程组, 然后将该随机非线性方程组转化为确定性的非线性方程组进行求解, 从而得到各基函数的系数以及逼近函数. 该方法无需抽样即可达到很高的计算精度; 但只适用于系统模型相对简单且随机变量数较小的场景.

2) 随机配置点法 [10]. 该类方法首先从原系统获得与输入随机变量的各个抽样点 (即配置点) 对应的输出变量的值, 然后利用这些配置点和相应的函数值计算出各基函数的系数, 从而得到逼近函数. 因为只需要配置点信息而不必考虑系统的具体模型, 所以该类方法适用范围广、计算速度快. 该类方法又可分为插值型随机配置点法和伪谱型随机配置点法 (简称伪谱方法), 其中伪谱方法因为能使逼近函数在输入随机变量定义域内各点的误差总体最小, 所以相对于插值型随机配置点法能获得具有更高逼近精度的逼近函数. 伪谱方法可分为经典伪谱方法 [11] 和自适应稀疏伪谱逼近法 [12, 13](adaptive sparse pseudospectral approximation method). 其中, 经典伪谱方法因为缺乏数值积分规则的选取准则, 所以存在当所用数值积分阶次过低时 (即欠积分) 基函数系数计算不准确, 或所用数值积分阶次过高时 (即过积分) 计算量大的缺点. 而 A-SPAM 法, 又可称为新型伪谱方法, 综合利用了一般化 Smolyak 稀疏网格 [14]、自适应算法 [15] 以及张量积逼近网格与张量积积分规则的配对技术, 从而可获得比经典伪谱方法更高逼近精度的逼近函数. 该方法相对于经典伪谱方法具有如下优点: 1) 该方法由于基函数集形式灵活且可自适应扩展, 所以能够根据给定的计算精度要求自适应地确定逼近函数的展开形式, 从而避免了经典伪谱方法的展开阶次难以选择的问题; 2) 该方法基于张量积逼近网格与张量积积分规则配对的技术, 相应克服了经典伪谱方法因欠积分和过积分而导致的问题, 使得其计算效率更高.

尽管 A-SPAM 具有上述优点, 但该方法仍存在如下缺点: 1) 其终止判据对于当前逼近函数的逼近精度的估计不准确 (大多数情况下偏大), 导致所得逼近函数的逼近精度偏离预期精度; 2) 该方法只适用于单输出随机变量的量化问题, 在把它直接推广到多输出量化问题时, 由于抽样信息没有得到充分利用而使得计算效率偏低.

基于如上综述, 本文提出了适用于多输出量化问题的自适应稀疏伪谱逼近新方法 (new adaptive sparse pseudospectral approximation method for multi-output problems, NA-SPAM-

MOP). 本文首先提出了 A-SPAM 的新型终止判据, 并以命题的形式证明所提新型判据的有效性; 然后提出了基于此新型终止判据的适用于单输出量化问题的自适应稀疏伪谱逼近新方法 (new adaptive sparse pseudospectral approximation method, NA-SPAM). 新型终止判据能够更准确地估计出逼近函数的当前精度, 从而使得 NA-SPAM 的逼近函数的精度更接近于预期精度. 进而, 针对多输出量化问题, 本文基于指标集的统一策略, 提出了适用于多输出量化问题的自适应稀疏伪谱逼近新方法 (NA-SPAM-MOP). 指标集统一策略可使抽样信息得到充分利用, 从而使得结合指标集统一策略的 NA-SPAM-MOP 在求解多输出量化问题时, 具有比直接把 NA-SPAM 推广到多输出问题的方法更高的计算效率.

为方便阅读及叙述, 本文将文中涉及到的部分定义及算法置于附录中.

2. 自适应稀疏伪谱逼近法 (A-SPAM) 简介

记输出变量 U 与输入随机变量 $\mathbf{Z}$ 的函数关系式为:

$$U = f(\mathbf{Z}), \tag{2.1}$$

式中 $\mathbf{Z} = (Z_1, \ldots, Z_d)$ 是定义在 Ω 上的独立随机变量向量, 其联合概率密度函数为 $\rho(\mathbf{z}) = \rho_1(z_1) \times \cdots \times \rho_d(z_d)$, 其中 $\rho_i(z_i), i = 1, \ldots, d$ 为随机变量 Z_i 的概率密度函数. 如何快速准确地求出输出变量 U 的各阶矩和概率密度, 为一类常见的不确定性传递问题. 若式 (2.1) 中的函数 $f(\mathbf{z})$ 满足平方可积条件:

$$\int_{\Omega} |f(\mathbf{z})|^2 \cdot \rho(\mathbf{z})\, \mathrm{d}\mathbf{z} < \infty,$$

则该不确定性传递问题可采用 A-SPAM 等 gPC 法有效求解. 并且 $f(\mathbf{z})$ 越光滑, 采用 A-SPAM 等 gPC 法求解上述问题的计算效率越高, 特别是当 $f(\mathbf{z})$ 无穷可微时, A-SPAM 具有指数收敛特性 [15].

A-SPAM 对上述不确定性传递问题的基本求解过程为: 先根据稀疏伪谱逼近法 (sparse pseudospectral approximation method, SPAM) 计算指标集对应的一般化 Smolyak 稀疏网格逼近函数; 然后利用自适应算法反复找出能最大程度地提高逼近精度的指标, 将其添加到指标集中, 并修正指标集对应的一般化 Smolyak 稀疏网格逼近函数, 直到迭代终止判据得到满足 (此时即认为算法已达到事先给定的逼近精度); 最后把所得到的逼近函数代替原函数关系, 进而计算输出变量 U 的各阶矩和概率密度函数.

2.1. 稀疏伪谱逼近法 (SPAM)

记 $f(\mathbf{z})$ 的逼近函数为 $\tilde{f}(\mathbf{z})$, 在稀疏伪谱逼近法 (SPAM) 中, $\tilde{f}(\mathbf{z})$ 为由各个张量积网格逼近函数线性组合而成的稀疏网格逼近函数:

$$\tilde{f}(\mathbf{z}) = \sum_{\mathbf{k} \in \mathcal{K}} c_{\mathbf{k}} \mathcal{S}_{\mathbf{k}}(f), \tag{2.2}$$

式中 $\mathbf{z}$ 为定义在 Ω 上的 d 维自变量; $\mathbf{k} = (k_1, \ldots, k_d)$ 和 $\mathbf{t} = (t_1, \ldots, t_d)$ 为 d 维指标 (简称为指标), 是由正整数构成的向量, k_i, t_i 分别是它们的第 i 个分量; $\mathcal{K}$ 为 d 维指标的集合 (简称为指标集); $\mathcal{S}_{\mathbf{k}}(f)$ 为与指标 $\mathbf{k}$ 对应的张量积网格逼近函数 (简称为逼近网格); $c_{\mathbf{k}}$ 为与指标 $\mathbf{k}$

对应的的网格系数. $\mathcal{S}_{\mathbf{k}}(f)$ 和 $c_{\mathbf{k}}$ 的表达式分别为:

$$\mathcal{S}_{\mathbf{k}}(f)=\sum_{n_1=0}^{s^{(1)}(k_1)}\cdots\sum_{n_d=0}^{s^{(d)}(k_d)}\mathcal{Q}_{\mathbf{k}}(\Phi_{\mathbf{n}}\cdot f)/\gamma_{\mathbf{n}}\cdot\Phi_{\mathbf{n}}(\mathbf{z}), \tag{2.3}$$

$$c_{\mathbf{k}}=\sum_{\mathbf{t}\in\mathcal{K},\forall i=1,\ldots,d,\ k_i\le t_i\le k_i+1}(-1)^{\|\mathbf{t}\|_1-\|\mathbf{k}\|_1}, \tag{2.4}$$

式中 $\Phi_{\mathbf{n}}(\mathbf{z})$ 为 Ω 上带权 $\rho(\mathbf{z})$ 的正交基函数 (简称为基函数); $\mathbf{n}=(n_1,\ldots,n_d)$ 为 $\Phi_{\mathbf{n}}(\mathbf{z})$ 的次数; $\gamma_{\mathbf{n}}$ 为 $\Phi_{\mathbf{n}}(\mathbf{z})$ 的带权 $\rho(\mathbf{z})$ 2- 范数的平方; $\mathbf{s}(\mathbf{k})=(s^{(1)}(k_1),\ldots,s^{(d)}(k_d))$ 为 $\mathcal{S}_{\mathbf{k}}(f)$ 中基函数的最高次数, 其第 i 个分量 $s^{(i)}(k_i)$ 为指标分量 k_i 的单调递增函数, 一般可取为线性函数 $s^{(i)}(k_i)=k_i-1$ 或指数函数 $s^{(i)}(k_i)=2^{k_i-1}-1$; $\mathcal{Q}_{\mathbf{k}}$ 是用于计算 $\mathcal{S}_{\mathbf{k}}(f)$ 中所有基函数系数的张量积积分算子, 它由各个输入维度 $i=1,\ldots,d$ 的分量 $\mathcal{Q}_{k_i}^{(i)}$ 通过张量积规则构成; 为保证 $\mathcal{Q}_{\mathbf{k}}$ 对 $\mathcal{S}_{\mathbf{k}}(f)$ 中所有 $\mathbf{0}\sim\mathbf{s}(\mathbf{k})$ 次基函数的系数计算的准确性, SPAM 规定各维度的 $\mathcal{Q}_{k_i}^{(i)}$ 的代数精度 $q^{(i)}(k_i)$ 必须随着 k_i 的增大而递增并且要大于等于 $2s^{(i)}(k_i)$; $\mathcal{Q}_{\mathbf{k}}(\Phi_{\mathbf{n}}\cdot f)$ 表示用 $\mathcal{Q}_{\mathbf{k}}$ 计算带权 $\rho(\mathbf{z})$ 内积 $\langle\Phi_{\mathbf{n}},f\rangle$ 而得的近似值; 记 $\mathcal{Q}_{\mathbf{k}}$ 所使用的所有积分点的集合为 $\Theta_{\mathbf{k}}$.

相应地, 上述逼近函数 $\tilde{f}(\mathbf{z})$ 也可表示成张量积差分网格逼近函数之和:

$$\tilde{f}(\mathbf{z})=\sum_{\mathbf{k}\in\mathcal{K}}\Delta_{\mathbf{k}}(f), \tag{2.5}$$

式中 $\Delta_{\mathbf{k}}(f)$ 为张量积差分网格逼近函数 (简称为差分逼近网格), 其与张量积网格逼近函数的关系为:

$$\begin{aligned}\Delta_{\mathbf{k}}(f)&=\sum_{\mathbf{t}\in I_k}(-1)^{\|\mathbf{k}\|_1-\|\mathbf{t}\|_1}\mathcal{S}_{\mathbf{t}}(f),\\ I_k&=\{\mathbf{t}\in\mathcal{K}:\forall i=1,\ldots,d,\ k_i-1\le t_i\le k_i\}.\end{aligned} \tag{2.6}$$

2.2. 自适应算法

自适应算法包括指标添加算法和终止判据两部分. 指标添加算法: 向指标集 $\mathcal{K}$ 中逐步添加最能提高逼近函数的逼近精度的指标. 终止判据: 根据逼近精度是否满足给定精度对算法进行终止判定.

指标添加算法为:

- 定义每个指标 $\mathbf{k}$ 的局部误差估计值 (local error estimate) 为:

$$\varepsilon_{\mathbf{k}}=\|\Delta_{\mathbf{k}}(f)\|_{\mathrm{L}^2_{\rho(\mathbf{z})}}=\sqrt{\int_\Omega(\Delta_{\mathbf{k}}(f))^2\rho(\mathbf{z})\,\mathrm{d}\mathbf{z}}, \tag{2.7}$$

  该式表示 $\Delta_{\mathbf{k}}(f)$ 的带权 $\rho(\mathbf{z})\,\mathrm{L}^2$ 范数, 为表述方便, 本文将其简称为 2- 范数, 并且将带权 $\rho(\mathbf{z})\,\mathrm{L}^2$ 范数符号和内积符号简记作 $\|\cdot\|_2$ 和 $\langle\cdot,\cdot\rangle$. 由逼近函数的差分表达式 (2.5) 可知, 逼近函数的精度取决于 $\mathcal{K}$ 中所有 $\mathbf{k}$ 对应的 $\Delta_{\mathbf{k}}(f)$ 项, 因此 $\varepsilon_{\mathbf{k}}$ 可表示在逼近函数中添加 $\Delta_{\mathbf{k}}(f)$ 项后, 逼近函数的精度的提高量 (或误差的减小量).

- 将指标集 $\mathcal{K}$ 分成两个互不相交的子集 $\mathcal{A},\mathcal{O}$, 从集合 $\mathcal{A}$ 中找到具有最大 $\varepsilon_{\mathbf{k}}$ 的指标 $\mathbf{k}$, 将其从 $\mathcal{A}$ 移入集合 $\mathcal{O}$ 中.

- 对于上述指标 $\mathbf{k}$ 的任意前邻指标, 若其所有后邻指标都属于 $\mathcal{O}$, 则将其添加到 $\mathcal{A}$ 和 $\mathcal{K}$ 中. 该过程可用公式表示为 $\mathcal{A} := \mathcal{A}\bigcup(\mathcal{F}(\mathbf{k})\bigcap\{\mathbf{p}:\mathcal{B}(\mathbf{p})\in\mathcal{O}\}), \mathcal{K} = \mathcal{A}\bigcup\mathcal{O}$, 其中 $\mathcal{F}(\mathbf{k})$ 表示 $\mathbf{k}$ 的所有前邻指标构成的集合, $\mathcal{B}(\mathbf{p})$ 表示指标 $\mathbf{k}$ 的所有后邻指标构成的集合. 因为所选指标 $\mathbf{k}$ 的 $\varepsilon_{\mathbf{k}}$ 最大, 故可认为该指标的前邻指标的局部误差估计值也很大, 将这些指标加入到 $\mathcal{K}$ 中, 能够最大程度地提高逼近精度.

终止判据为:

- 定义逼近函数的全局误差估计值 (global error estimate) 为:

$$\eta = \sum_{\mathbf{k}\in\mathcal{A}} \varepsilon_{\mathbf{k}} = \sum_{\mathbf{k}\in\mathcal{A}} \|\Delta_{\mathbf{k}}(f)\|_2 . \tag{2.8}$$

- 若 $\eta \le \mathrm{TOL}$(TOL 为误差槛值), 则终止判据满足, 此时可认为逼近函数的逼近精度已满足要求, 自适应算法结束.

2.3. A-SPAM 的计算步骤

1) 初始化: 令 $\mathcal{K}$ 和 $\mathcal{A}$ 为单指标集合 $\{1,\ldots,1\}$, 令 $\mathcal{O}$ 为空集; 为每个指标 $\mathbf{k}$ 选择 $\mathbf{s}(\mathbf{k})$ 和 $\mathcal{Q}_{\mathbf{k}}$.

2) 根据 2.2 节的指标添加算法添加指标到 $\mathcal{A}$ 和 $\mathcal{K}$ 中.

3) 根据 2.1 节的 SPAM 方法计算指标集 $\mathcal{K}$ 对应的逼近函数 (2.2), (2.5).

4) 根据 2.2 节的终止判据计算并判定是否终止迭代, 若终止判据未满足, 则转第 2) 步继续迭代, 否则转下一步.

5) 计算输出变量的各阶矩和概率密度.

   - 将逼近函数 (2.2) 写成基函数线性组合的形式:

$$\tilde{f}(\mathbf{z}) = \sum_{\Phi_{\mathbf{n}}\in\mathcal{H}} \hat{f}_{\mathbf{n}}\Phi_{\mathbf{n}}(\mathbf{z}), \tag{2.9}$$

   式中 $\hat{f}_{\mathbf{n}}$ 为基函数 $\Phi_{\mathbf{n}}(\mathbf{z})$ 的系数, $\mathcal{H}$ 为基函数的集合 (简称基函数集), 包含 $\tilde{f}(\mathbf{z})$ 中所有的基函数项.

   - 计算输出变量 U 的期望和方差:

$$\begin{aligned} \mathbb{E}[U] &\approx \mathbb{E}[\tilde{f}(\mathbf{Z})] = \hat{f}_{\mathbf{0}}, \\ \mathrm{Var}(U) &\approx \mathrm{Var}[\tilde{f}(\mathbf{Z})] = \sum_{\Phi_{\mathbf{n}}\in\mathcal{H}} (\hat{f}_{\mathbf{n}})^2\gamma_{\mathbf{n}} - (\hat{f}_{\mathbf{0}})^2, \end{aligned} \tag{2.10}$$

   然后用随机变量 $\tilde{f}(\mathbf{Z})$ 代替输出变量 U 进行蒙特卡罗抽样, 得到 U 的高阶矩和概率密度.

   注: 因为逼近函数 (2.9) 为简单的多项式函数, 而原函数 (2.1) 一般为较复杂的隐式函数, 所以对 $\tilde{f}(\mathbf{Z})$ 进行抽样得到 U 的高阶矩和概率密度所需要的计算时间, 要远小于直接对 $U = f(\mathbf{Z})$ 进行相应的抽样计算所需要的计算时间.

3. 适用于单输出量化问题的自适应稀疏伪谱逼近新方法 (NA-SPAM)

第 2 节所介绍的 A-SPAM 是有效的方法, 相比于其它 gPC 方法可获得具有更高逼近精度的逼近函数. 但它的终止判据不能够准确估计出逼近函数的误差, 因而可能会在逼近函数的实际误差仍大于给定的误差槛值时该判据因已满足而提前终止计算, 或者在逼近函数的实际误差已小于误差槛值时该该判据却未得到满足而仍进行不必要的计算, 最终导致 A-SPAM 所得逼近函数的逼近精度会偏离给定逼近精度.

本文从逼近误差的表达式出发, 首先分析现有全局误差估计值和相应的终止判据对于逼近误差估计不准确的原因, 然后基于引理、定义和假设提出了新的全局误差估计值和相应的新型终止判据, 基于新型终止判据提出了 NA-SPAM; 并以命题的形式证明了新型终止判据相比于现有终止判据, 能够更准确地估计出逼近误差.

3.1. A-SPAM 终止判据对于逼近误差估计不准确的原因分析

用 $\tilde{f}(\mathbf{z})$ 逼近 $f(\mathbf{z})$ 所产生的逼近误差可用两者之差的 2- 范数 $\left\|\tilde{f}(\mathbf{z})-f(\mathbf{z})\right\|_2$ 来表示. 逼近误差越小, 则 $\tilde{f}(\mathbf{z})$ 对 $f(\mathbf{z})$ 的逼近精度越高, 相应地表明 A-SPAM 的计算精度也越高.

原函数 (2.1) 可写成级数形式:

$$f(\mathbf{z})=\sum_{\mathbf{k}\in\mathbb{N}_1^d}\Delta_{\mathbf{k}}(f). \tag{3.1}$$

结合式 (2.5), 逼近误差可表示为:

$$\left\|\tilde{f}(\mathbf{z})-f(\mathbf{z})\right\|_2=\left\|\sum_{\mathbf{k}\in\mathcal{K}}\Delta_{\mathbf{k}}(f)-\sum_{\mathbf{k}\in\mathbb{N}_1^d}\Delta_{\mathbf{k}}(f)\right\|_2=\left\|\sum_{\mathbf{k}\in\mathcal{K}^{\mathrm{c}}}\Delta_{\mathbf{k}}(f)\right\|_2, \tag{3.2}$$

式中 $\mathcal{K}^{\mathrm{c}}$ 是 $\mathcal{K}$ 关于 d 维正整数集 $\mathbb{N}_1^d$ 的补集, 称为外部指标集 (为无穷集合).

若要精确计算式 (3.2), 则需要在与 $\mathcal{K}^{\mathrm{c}}$ 中所有指标 $\mathbf{k}$ 对应的积分点 $\Theta_{\mathbf{k}}$ 处对 $f(\mathbf{z})$(式 (2.1)) 进行抽样, 然后计算相应的 $\Delta_{\mathbf{k}}(f)$. $\mathcal{K}^{\mathrm{c}}$ 是无穷集合, 相应地精确计算式 (3.2) 的值需要无穷次抽样而无法实现, 因此只能采用估计方法来近似计算式 (3.2) 的值, 即利用已通过抽样算出的 $\Delta_{\mathbf{k}}(f),\mathbf{k}\in\mathcal{K}$ 来近似估计式 (3.2) 的值.

A-SPAM 用全局误差估计值 (式 (2.8)) 估计式 (3.2), 其估计精度偏低的一个原因为 [16]: 在第 2.2 节自适应算法的指标添加过程中, 需要找到 $\mathcal{A}$ 中具有最大 $\varepsilon_{\mathbf{k}}$ 的指标 $\mathbf{k}$ 并将其从 $\mathcal{A}$ 中移到 $\mathcal{O}$ 中, 若在 $\mathbf{k}$ 的所有前邻指标中没有满足所有后邻指标都属于 $\mathcal{O}$ 这一条件的指标, 那么本次操作就没有指标可添加, 由于 $\mathcal{K}=\mathcal{A}\bigcup\mathcal{O}$, 因此 $\mathcal{K}$ 不变, 相应地逼近函数式 (2.2) 在操作前后也不变, 其逼近误差应不变; 而由于该 $\mathbf{k}$ 已从 $\mathcal{A}$ 中移出而使得 $\mathcal{A}$ 发生了变化, 相应根据式 (2.8) 计算出的全局误差估计值在本次操作前后因 $\mathcal{A}$ 的变化而有所不同, 而实际上全局误差估计值应不变; 因此, 根据式 (2.8) 计算出的全局误差估计值在这种情况下无法真实地反映逼近函数对于原函数的逼近情况, 而存在一定的误差, 相应其估计精度偏低.

3.2. 引理、定义及假设条件

在提出本文的新型终止判据之前, 先给出如下的引理、定义及假设条件.

引理 1.

$$\sqrt{\sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{s}}}\|\Delta_{\mathbf{k}}(f)\|_2^2}-\left\|\sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{s}}}\Delta_{\mathbf{k}}(f)\right\|_2\approx 0, \tag{3.3}$$

式中 $\mathcal{K}_{\mathrm{s}}$ 为 $\mathcal{K}$ 的任意子集.

证明. 任意两个不相等的指标 $\mathbf{k},\mathbf{t}$ 对应的差分逼近网格可写成: $\Delta_{\mathbf{k}}(f)=\sum\limits_{\mathbf{n}\le\mathbf{s}(\mathbf{k})}B_{\mathbf{k},\mathbf{n}}\Phi_{\mathbf{n}}$, $\Delta_{\mathbf{t}}(f)=\sum\limits_{\mathbf{n}\le\mathbf{s}(\mathbf{t})}B_{\mathbf{t},\mathbf{n}}\Phi_{\mathbf{n}}$, 其中 $B_{\mathbf{k},\mathbf{n}},B_{\mathbf{t},\mathbf{n}}$ 为基函数 $\Phi_{\mathbf{n}}$ 的系数; 由式 (2.6) 以及式 (2.3) 可得 $B_{\mathbf{k},\mathbf{n}}$ 的计算式为:

$$B_{\mathbf{k},\mathbf{n}}=\sum_{\mathbf{p}\in I_{\mathbf{k}},\mathbf{s}(\mathbf{p})\ge\mathbf{n}}(-1)^{\|\mathbf{k}\|_1-\|\mathbf{p}\|_1}J_{\mathbf{p},\mathbf{n}}, \tag{3.4}$$

式中 $\mathbf{s}(\mathbf{p})$ 为 $\mathcal{S}_{\mathbf{p}}(f)$ 中基函数 $\Phi_{\mathbf{n}}$ 的最高次数; $J_{\mathbf{p},\mathbf{n}}$ 为 $\mathcal{S}_{\mathbf{p}}(f)$ 中基函数 $\Phi_{\mathbf{n}}$ 的系数, $J_{\mathbf{p},\mathbf{n}}=\mathcal{Q}_{\mathbf{p}}(\Phi_{\mathbf{n}}\ f)/\gamma_{\mathbf{n}}$. 当 $\mathbf{n}=\mathbf{s}(\mathbf{k})$ 时, 式 (3.4) 的求和号内只有 1 项, 故 $B_{\mathbf{k},\mathbf{s}(\mathbf{k})}=J_{\mathbf{k},\mathbf{n}}\neq 0$; 而当 $\mathbf{n}<\mathbf{s}(\mathbf{k})$ 时, (3.4) 的求和号内有偶数项, 且随着 $\mathbf{p}$ 的变化, $(-1)^{\|\mathbf{k}\|_1-\|\mathbf{p}\|_1}$ 的值正负交替出现. 因为与各个指标 $\mathbf{p}$ 对应的 $J_{\mathbf{p},\mathbf{n}}$ 的区别仅在于采用的数值积分不同, 所以可认为它们近似相等. 于是对于任意 $\mathbf{n}<\mathbf{s}(\mathbf{k})$, 有:

$$B_{\mathbf{k},\mathbf{n}}\approx J_{\mathbf{p},\mathbf{n}}\sum_{\mathbf{p}\in I_{\mathbf{k}},\mathbf{s}(\mathbf{p})\ge\mathbf{n}}(-1)^{\|\mathbf{k}\|_1-\|\mathbf{p}\|_1}=J_{\mathbf{p},\mathbf{n}}\cdot(1-1+\cdots+1-1)=0,$$

同理, 当 $\mathbf{n}=\mathbf{s}(\mathbf{t})$ 时, $B_{\mathbf{t},\mathbf{n}}\neq 0$, 对于任意 $\mathbf{n}<\mathbf{s}(\mathbf{t})$, 有 $B_{\mathbf{t},\mathbf{n}}\approx 0$. 因此, 由基函数的正交性, 当 $\mathbf{k}\neq\mathbf{t}$ 时, $\Delta_{\mathbf{k}}(f),\Delta_{\mathbf{t}}(f)$ 的内积为:

$$\langle\Delta_{\mathbf{k}}(f),\Delta_{\mathbf{t}}(f)\rangle=\sum_{\mathbf{n}\le\mathbf{s}(\mathbf{k}),\mathbf{n}\le\mathbf{s}(\mathbf{t})}B_{\mathbf{k},\mathbf{n}}B_{\mathbf{t},\mathbf{n}}\eta_{\mathbf{n}}\approx 0, \tag{3.5}$$

根据范数平方展开公式得到:

$$\left\|\sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{s}}}\Delta_{\mathbf{k}}(f)\right\|_2^2=\sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{s}}}\|\Delta_{\mathbf{k}}(f)\|_2^2+\sum_{\mathbf{k},\mathbf{t}\in\mathcal{K}_{\mathrm{s}},\ \mathbf{k}\neq\mathbf{t}}\langle\Delta_{\mathbf{k}}(f),\Delta_{\mathbf{t}}(f)\rangle, \tag{3.6}$$

结合式 (3.5) 和 (3.6), 可得

$$\left\|\sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{s}}}\Delta_{\mathbf{k}}(f)\right\|_2\approx\sqrt{\sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{s}}}\|\Delta_{\mathbf{k}}(f)\|_2^2}, \tag{3.7}$$

因此式 (3.3) 得证. □

定义 1(边界指标). 若指标 $\mathbf{k}$ 满足 $\mathbf{k}\in\mathcal{K}$, 且 $\mathbf{k}$ 的前邻指标中至少有一个不在 $\mathcal{K}$ 中, 则称 $\mathbf{k}$ 为 $\mathcal{K}$ 的边界指标. 全体边界指标构成边界指标集.

定义 2(顶点指标). 若指标 $\mathbf{k}$ 满足 $\mathbf{k}\in\mathcal{K}$, 且 $\mathbf{k}$ 的所有前邻指标都不在 $\mathcal{K}$ 中, 则称 $\mathbf{k}$ 为 $\mathcal{K}$ 的顶点指标. 全体顶点指标构成顶点指标集.

定义 3(外边界指标). 若指标 $\mathbf{k}$ 满足 $\mathbf{k}\in\mathcal{K}^{\mathrm{c}}$, 且 $\mathbf{k}$ 的后邻指标中至少有一个在 $\mathcal{K}$ 中, 则称 $\mathbf{k}$ 为 $\mathcal{K}$ 的外边界指标. 全体外边界指标构成外边界指标集.

定义 4(外顶点指标). 若指标 $\mathbf{k}$ 满足 $\mathbf{k}\in\mathcal{K}^{\mathrm{c}}$, 且 $\mathbf{k}$ 的所有后邻指标都在 $\mathcal{K}$ 中, 则称 $\mathbf{k}$ 为 $\mathcal{K}$ 的外顶点指标. 全体外顶点指标构成外边界指标集, 记作 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$.

假设 1. 若 $\mathbf{k}>\mathbf{t}$, 则 $\|\Delta_{\mathbf{k}}(f)\|_2<\|\Delta_{\mathbf{t}}(f)\|_2$.

假设 2. 在估计逼近误差式 (3.2) 时, 可用 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 近似代替 $\mathcal{K}^{\mathrm{c}}$.

假设 3. 在计算 $\|\Delta_{\mathbf{k}}(f)\|_2,\mathbf{k}\in\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 时, 可将 $\min_{\mathbf{t}\in\mathcal{B}(\mathbf{k})}\|\Delta_{\mathbf{t}}(f)\|_2$ 作为它的近似值, 其中 $\mathcal{B}(\mathbf{k})$ 为 $\mathbf{k}$ 的所有后邻指标构成的集合.

上述假设的合理性分析如下:

1) 假设 1 是在估计逼近误差时的一个基本假设 [15]. 由级数式 (3.1) 的收敛性可知, 当 $\mathbf{k}\to\infty$ 时, $\Delta_{\mathbf{k}}(f)\to 0,\|\Delta_{\mathbf{k}}(f)\|_2\to 0$ 这表明在 $\mathbf{k}\to\infty$ 时, 总体上 $\|\Delta_{\mathbf{k}}(f)\|_2$ 的趋势必然是逐渐减小的, 即就总体趋势而言 $\mathbf{k}$ 越大, $\|\Delta_{\mathbf{k}}(f)\|_2$ 就越小, 因此假设 1 是合理的.

2) 设任意 $\mathbf{p}\in\mathcal{K}^{\mathrm{c}}\setminus\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$, 从 $\mathbf{p}$ 的后邻指标集中找一个指标 $\mathbf{p}^{(1)}\in\mathcal{K}^{\mathrm{c}}$, 再从 $\mathbf{p}^{(1)}$ 的后邻指标集中找一个指标 $\mathbf{p}^{(2)}\in\mathcal{K}^{\mathrm{c}}$, 重复进行该后邻查找操作, 直至得到 $\mathbf{p}^{(m)}$ 的后邻指标都不属于 $\mathcal{K}^{\mathrm{c}}$. 记 $\mathbf{k}=\mathbf{p}^{(m)}$, 根据定义可得 $\mathbf{k}\in\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$, 由于 $\mathbf{k}<\mathbf{p}$, 所以根据假设 1 可得 $\|\Delta_{\mathbf{k}}(f)\|_2>\|\Delta_{\mathbf{p}}(f)\|_2$, 相应地意味着与 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 中的指标对应的局部误差估计值大于与 $\mathcal{K}^{\mathrm{c}}\setminus\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 中的指标对应的局部误差估计值, 因而 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 中的指标相对于 $\mathcal{K}^{\mathrm{c}}\setminus\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 的指标, 对式 (3.2) 值的影响也更大, 在其值的组成中占主导地位, 因而在式 (3.2) 中可忽略 $\mathcal{K}^{\mathrm{c}}\setminus\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 而近似用 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 代替 $\mathcal{K}^{\mathrm{c}}$, 仍可保证对式 (3.2) 有足够的估计精度, 本文通过多个算例验证该假设是合理的, 具体见算例分析部分的算例 1 的第 2 部分.

3) 对任意 $\mathbf{k}\in\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$, 由于 $\mathbf{k}\notin\mathcal{K}$, 所以无法直接得到 $\|\Delta_{\mathbf{k}}(f)\|_2$, 需要用相应的已有计算结果来近似地获得其值, 即利用 $\mathcal{K}$ 中某个指标 $\mathbf{t}$ 对应的 $\|\Delta_{\mathbf{t}}(f)\|_2$ 来近似代替. 由外顶点指标的定义可知, $\mathcal{B}(\mathbf{k})\subset\mathcal{K}$ 并且在 $\mathcal{K}$ 的所有指标中, $\mathcal{B}(\mathbf{k})$ 中的指标与 $\mathbf{k}$ 的距离最近 (为 1), 自然地 $\mathcal{B}$ 中指标对应的 $\|\Delta_{\mathbf{t}}(f)\|_2,\mathbf{t}\in\mathcal{B}(\mathbf{k})$ 也最接近 $\|\Delta_{\mathbf{k}}(f)\|_2$. 对任意 $\mathbf{t}\in\mathcal{B}(\mathbf{k})$ 有 $\mathbf{t}<\mathbf{k}$, 由假设 1 可得 $\|\Delta_{\mathbf{t}}(f)\|_2>\|\Delta_{\mathbf{k}}(f)\|_2$, 所以 $\min_{\mathbf{t}\in\mathcal{B}(\mathbf{k})}\|\Delta_{\mathbf{t}}(f)\|_2$ 与 $\|\Delta_{\mathbf{k}}(f)\|_2$ 的值最接近, 相应地用它近似代替 $\|\Delta_{\mathbf{k}}(f)\|_2$ 最为合理, 因此假设 3 是合理的.

3.3. 新型终止判据及基于此的 NA-SPAM

基于上述引理、定义和假设条件, 本文提出如下新型全局误差估计值和相应的终止判据:

$$\begin{aligned}&\eta_{\mathrm{new}}=\sqrt{\sum_{\mathbf{t}\in\mathcal{K}_{\mathrm{E}}}b_{\mathbf{t}}\|\Delta_{\mathbf{t}}(f)\|_2^2}=\sqrt{\sum_{\mathbf{t}\in\mathcal{K}_{\mathrm{E}}}b_{\mathbf{t}}\varepsilon_{\mathbf{t}}^2},\\&\eta_{\mathrm{new}}\le\mathrm{TOL},\end{aligned}\tag{3.8}$$

式中 η_{new} 为与新型终止判据相对应的新的全局误差估计值; $\mathcal{K}_{\mathrm{E}}$ 为 $\mathcal{K}$ 的一个子集, 按如下方式获得: 对于多重集合 (即相同的元素可以重复出现的集合)$\mathcal{K}_{\mathrm{E}}'$:

$$\mathcal{K}_{\mathrm{E}}'=\{\arg\min_{\mathbf{t}\in\mathcal{B}(\mathbf{k})}\|\Delta_{\mathbf{t}}(f)\|_2:\mathbf{k}\in\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}\},\tag{3.9}$$

在上式中删除 $\mathcal{K}_{\mathrm{E}}'$ 中的重复元素并只保留其中一个之后所得到的集合即为 $\mathcal{K}_{\mathrm{E}}$; $b_{\mathbf{t}}$ 为 $\varepsilon_{\mathbf{t}}$ 的系数, 其值为指标 $\mathbf{t}$ 在 $\mathcal{K}_{\mathrm{E}}'$ 中出现的次数. 因为 $\mathcal{K}_{\mathrm{E}}'$ 完全取决于 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$, 进而完全取决于 $\mathcal{K}$, 当 $\mathcal{K}$ 和

逼近函数不变时, $\mathcal{K}'_{\mathrm{E}}$ 不变, $\mathcal{K}_{\mathrm{E}}$ 也不变, 因而由式 (23) 计算出的新的全局误差估计值 $b_{\mathbf{t}}$ 也不变, 相应地克服了现有终止判据在逼近函数不变时全局误差估计值却发生改变这一缺陷.

只要在第 2 节的终止判据中用式 (3.8) 代替式 (2.8) 计算全局误差估计值, 并相应地修改终止判据, 即得到基于新型终止判据的 NA-SPAM.

以下命题将证明式 (3.8) 对于逼近误差的估计相比于式 (2.8) 对于逼近误差的估计具有更高的精度.

对于同一个逼近函数 $\tilde{f}(\mathbf{z})$, 将其对 $f(\mathbf{z})$ 的实际逼近误差值 (式 (3.2)) 记为 ζ, 将现有全局误差估计值 (2.8) 记为 η_{old}, 将新全局误差估计值 (3.8) 记为 η_{new}. 则有如下命题成立:

命题 1.

$$|\eta_{\mathrm{new}}-\zeta|<|\eta_{\mathrm{old}}-\zeta|. \tag{3.10}$$

证明.

$$\begin{aligned}\eta_{\mathrm{new}}-\zeta=&\left(\left\|\sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}}\Delta_{\mathbf{k}}(f)\right\|_2-\left\|\sum_{\mathbf{k}\in\mathcal{K}^{\mathrm{c}}}\Delta_{\mathbf{k}}(f)\right\|_2\right)+\\&\left(\sqrt{\sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}}\|\Delta_{\mathbf{k}}(f)\|_2^2}-\left\|\sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}}\Delta_{\mathbf{k}}(f)\right\|_2\right)+\\&\left(\sqrt{\sum_{\mathbf{t}\in\mathcal{K}_{\mathrm{E}}}b_{\mathbf{t}}\|\Delta_{\mathbf{t}}(f)\|_2^2}-\sqrt{\sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}}\|\Delta_{\mathbf{k}}(f)\|_2^2}\right),\end{aligned} \tag{3.11}$$

$$\begin{aligned}\eta_{\mathrm{old}}-\zeta=&\left(\left\|\sum_{\mathbf{k}\in\mathcal{A}}\Delta_{\mathbf{k}}(f)\right\|_2-\left\|\sum_{\mathbf{k}\in\mathcal{K}^{\mathrm{c}}}\Delta_{\mathbf{k}}(f)\right\|_2\right)+\\&\left(\sum_{\mathbf{k}\in\mathcal{A}}\|\Delta_{\mathbf{k}}(f)\|_2-\left\|\sum_{\mathbf{k}\in\mathcal{A}}\Delta_{\mathbf{k}}(f)\right\|_2\right).\end{aligned} \tag{3.12}$$

将式 (3.11) 右端的三项分别记为 $\alpha_1,\alpha_2,\alpha_3$, 将式 (3.12) 右端的两项分别记为 β_1,β_2. α_1, α_3, β_1 分别为用 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 代替 $\mathcal{K}^{\mathrm{c}}$, 用 $\mathcal{K}_{\mathrm{E}}$ 代替 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$, 用 $\mathcal{A}$ 代替 $\mathcal{K}^{\mathrm{c}}$ 所产生的估计误差项, α_2,β_2 为将差分逼近网格和范数转化成范数和产生的估计误差项.

根据假设 2, 可近似用 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 代替 $\mathcal{K}^{\mathrm{c}}$, 故 $\alpha_1\approx 0$. 根据假设 3, 可近似用 $\min_{\mathbf{t}\in\mathcal{B}(\mathbf{k})}\|\Delta_{\mathbf{t}}(f)\|_2$ 代替 $\|\Delta_{\mathbf{k}}(f)\|_2,\mathbf{k}\in\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$, 即近似用 $\mathcal{K}_{\mathrm{E}}$ 代替 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$, 故 $\alpha_3\approx 0$. 由于 $\mathcal{A}\subset\mathcal{K}$, 所以 $\mathcal{A}\neq\mathcal{K}^{\mathrm{c}}$, $\beta_1\neq 0$.

α_2,β_2 分别为范数平方和近似和范数三角不等式近似的误差, 根据引理 (3.3) 可以直接得到 $\alpha_2\approx 0$.

由于 $-\beta_1$ 取决于集合 $\mathcal{K},\mathcal{A}$, 而 β_2 仅取决于集合 $\mathcal{A}$, 所以必然有 $-\beta_1\neq\beta_2$, 即 $\beta_1+\beta_2\neq 0$. 综上可得

$$\begin{aligned}|\eta_{\mathrm{new}}-\zeta|&=|\alpha_1+\alpha_2+\alpha_3|\approx 0,\\|\eta_{\mathrm{old}}-\zeta|&=|\beta_1+\beta_2|>0.\end{aligned}$$

因此, 式 (3.10) 左端项小于右端项, 命题成立. □

上述命题意味着 η_{new} 比 η_{old} 更接近 ζ , 即新型终止判据对于逼近误差的估计相比于现有判据更加准确.

因为 NA-SPAM 在逼近函数的计算方法和指标的添加算法上与 A-SPAM 相同, 仅在终止判据上有所不同, 所以 NA-SPAM 和 A-SPAM 的唯一区别在于何时终止迭代过程 (即指标集的扩展过程). 上述命题证明了新型终止判据的全局误差估计值相比于现有终止判据的全局误差估计值更接近于实际的逼近误差, 从而在给定相同的误差槛值 TOL(即给定相同的逼近精度) 时, 基于新型终止判据的 NA-SPAM 相比于 A-SPAM 能在逼近误差更接近误差槛值 TOL 时终止迭代过程, 从而能获得逼近精度更接近于给定逼近精度的逼近函数.

4. 适用于多输出量化问题的自适应稀疏伪谱新方法 (NA-SPAM-MOP)

对于单输出变量的随机量化问题, A-SPAM 是有效的方法; 但对于多输出变量的量化问题 (即 U 的维数大于 1), 目前相关文献未见讨论, 而大多工程问题却为多输出问题. 若将第 3 节的 NA-SPAM 直接推广到多输出量化问题, 则抽样结果无法得到充分利用而计算效率低.

本文基于指标集统一策略, 提出了适用于多输出量化问题的自适应稀疏伪谱新方法 (NA-SPAM-MOP), 该方法能提高抽样结果的利用率而提高计算效率, 相应克服了把 NA-SPAM 直接推广到多输出量化问题时计算效率低的缺点.

4.1. 将 NA-SPAM 直接推广到多输出量化问题的方法及缺点

将 NA-SPAM 直接推广到多输出量化问题的方法为: 将多输出量化问题 $\mathbf{U}=\boldsymbol{f}(\mathbf{Z})$ 转化为多个独立的单输出量化问题 $U^{(j)}=f^{(j)}(\mathbf{Z}), j=1,\ldots,N_{\mathrm{o}}$($N_{\mathrm{o}}$ 为输出变量总个数), 再分别使用 NA-SPAM 进行计算, 求解出各个单输出量化问题的指标集 $\mathcal{K}^{(j)}$、逼近网格 $\mathcal{S}_{\mathbf{k}}(f^{(j)})$、网格系数 $c_{\mathbf{k}}^{(j)}$、差分逼近网格 $\Delta_{\mathbf{k}}(f^{(j)})$ 以及逼近函数 $\tilde{f}^{(j)}(\mathbf{z})$:

$$\tilde{f}^{(j)}(\mathbf{z})=\sum_{\mathbf{k}\in\mathcal{K}^{(j)}}\Delta_{\mathbf{k}}(f^{(j)})=\sum_{\mathbf{k}\in\mathcal{K}^{(j)}}c_{\mathbf{k}}^{(j)}\mathcal{S}_{\mathbf{k}}(f^{(j)}). \tag{4.1}$$

这种做法简单, 但其存在如下缺点: 在计算一个输出变量时, 不能够充分地利用其他输出变量的抽样结果来提高该输出变量的计算精度, 从而使得 NA-SPAM 的计算效率低、计算时间长. 具体原因为: 计算 $\mathcal{S}_{\mathbf{k}}(f^{(1)}),\ldots,\mathcal{S}_{\mathbf{k}}(f^{(N_{\mathrm{o}})})$ 时所用的数值积分算子 $\mathcal{Q}_{\mathbf{k}}$ 和积分点集 $\Theta_{\mathbf{k}}$ 是相同的, 与之相应的输出变量 $U^{(j)}, j=1,\ldots,N_{\mathrm{o}}$ 的取值集合记为 $\Xi_{\mathbf{k}}^{(j)}=\{f^{(j)}(\mathbf{Z}):\mathbf{Z}\in\Theta_{\mathbf{k}}\}$. 由于在该做法中, 各输出变量 $U^{(j)}, j=1,\ldots,N_{\mathrm{o}}$ 的求解是独立的, 所以在求解各输出变量时得到的指标集一般不会相同, 不妨设 $\mathcal{K}^{(1)}\neq\mathcal{K}^{(2)}$, 则必定存在一个指标 $\mathbf{k}$, 满足 $\mathbf{k}\in\mathcal{K}^{(1)}$ 且 $\mathbf{k}\notin\mathcal{K}^{(2)}$. 在计算 $\mathcal{S}_{\mathbf{k}}(f^{(1)})$ 时, 需要计算输入变量取值集合 $\Theta_{\mathbf{k}}$ 所对应的 $U^{(1)}$ 的取值集合 $\Xi_{\mathbf{k}}^{(1)}$, 因为在多输出量化问题中给定一组输入变量的值就可同时计算出各输出变量的值, 所以在计算出 $\Xi_{\mathbf{k}}^{(1)}$ 的同时, 也计算出了 $\Theta_{\mathbf{k}}$ 所对应的 $U^{(2)}$ 的取值集合 $\Xi_{\mathbf{k}}^{(2)}$, 但 $\mathbf{k}\notin\mathcal{K}^{(2)}$ 导致 $\tilde{f}^{(2)}(\mathbf{z})$ 中不包含 $\Xi_{\mathbf{k}}^{(2)}$ 这一项, 使得在计算 $\tilde{f}^{(2)}(\mathbf{z})$ 时不能够利用 $\Xi_{\mathbf{k}}^{(2)}$ 这一抽样结果, 于是造成了抽样信息 $\Xi_{\mathbf{k}}^{(2)}$ 无法得到利用而降低了抽样信息的利用效率.

4.2. 基于指标集统一策略的 NA-SPAM-MOP

考虑到上述将 NA-SPAM 直接推广到多输出量化问题的方法存在的缺点, 本文基于指标集统一策略提出了适用于多输出问题量化的 NA-SPAM-MOP. 该方法的基本思想为: 采用指标集的统一策略使得在指标集自适应扩展过程中各输出变量的逼近函数的指标集始终保持相同, 这样就避免了上述方法因各逼近函数的指标集不相同而存在抽样信息无法得到充分利用

的缺点. 本文所提出的指标集统一策略包括如下两部分内容: 1) 统一各输出变量对应的局部误差估计值; 2) 统一各输出变量对应的全局误差估计值.

1) 统一各输出变量对应的局部误差估计值. 仿照式 (2.7), 定义与第 $j=1,\dots,N_{\rm o}$ 个输出变量对应的、指标 $\mathbf{k}$ 的局部误差估计值为:

$$\varepsilon_{\mathbf{k}}^{(j)}=\left\|\Delta_{\mathbf{k}}(f^{(j)})\right\|_2 \tag{4.2}$$

对各输出变量对应的 $\varepsilon_{\mathbf{k}}^{(j)}$ 取加权平均, 就得到统一的局部误差估计值 $\varepsilon_{\mathbf{k}}$:

$$\varepsilon_{\mathbf{k}}=\frac{1}{N_{\rm o}}\sum_{j=1}^{N_{\rm o}}\frac{\varepsilon_{\mathbf{k}}^{(j)}}{\varepsilon_{\rm ref}^{(j)}} \tag{4.3}$$

$$\varepsilon_{\rm ref}^{(j)}=\sqrt{\sum_{\mathbf{k}\in\mathcal{K}}(\varepsilon_{\mathbf{k}}^{(j)})^2}=\sqrt{\sum_{\mathbf{k}\in\mathcal{K}}\left\|\Delta_{\mathbf{k}}(f^{(j)})\right\|_2^2} \tag{4.4}$$

2) 统一各输出变量对应的全局误差估计值: 根据新型终止判据 (3.11), 各输出变量对应的全局误差估计值为:

$$\eta^{(j)}=\sqrt{\sum_{\mathbf{t}\in\mathcal{K}_E^{(j)}}b_{\mathbf{t}}^{(j)}\cdot\varepsilon_{\mathbf{t}}^{(j)}\varepsilon_{\mathbf{t}}^{(j)}} \tag{4.5}$$

式中 $\mathcal{K}_E^{(j)}$ 和 $b_{\mathbf{t}}^{(j)}$ 分别为根据新型终止判据来估计 $\tilde{f}^{(j)}$ 对 $f^{(j)}(\mathbf{z})$ 的逼近误差时所用的集合和系数, 其计算方法与式 (3.11) 中的 $\mathcal{K}_E$ 和 $b_{\mathbf{t}}$ 相同.

对各输出变量的 $\eta^{(j)}$ 取加权平均, 就得到统一的全局误差估计值和终止判据:

$$\begin{aligned}&\eta=\frac{1}{N_{\rm o}}\sum_{j=1}^{N_{\rm o}}\eta^{(j)}/\varepsilon_{\rm ref}^{(j)}\\&\eta\le{\rm TOL}\end{aligned} \tag{4.6}$$

在式 (4.3) 和式 (4.6) 中, $1/\varepsilon_{\rm ref}^{(j)}$ 分别是第 j 个输出变量对应的局部误差估计值的加权系数和全局误差估计值的加权系数, 该系数的作用及合理性说明如下:

1) 由式 (4.4) 可知, $\varepsilon_{\rm ref}^{(j)}$ 为第 j 个输出变量 $U^{(j)}$ 的指标集 $\mathcal{K}^{(j)}$ 中所有 $\varepsilon_{\mathbf{k}}^{(j)}$ 的平方和的根 (相当于由所有 $\varepsilon_{\mathbf{k}}^{(j)},\mathbf{k}\in\mathcal{K}^{(j)}$ 构成的向量的欧氏长度), 因此 $1/\varepsilon_{\rm ref}^{(j)}$ 具有归一化作用, 能合理均衡不同大小的 $\varepsilon_{\mathbf{k}}^{(j)}$ 对统一的误差估计值 $\varepsilon_{\mathbf{k}}$ 的影响, 使各个输出变量的局部误差估计值均能在 $\varepsilon_{\mathbf{k}}$ 中得到合理的体现.

2) 根据范数平方展开近似式 (3.7) 可得:

$$\varepsilon_{\rm ref}^{(j)}\approx\left\|\sum_{\mathbf{k}\in\mathcal{K}}\Delta_{\mathbf{k}}(f^{(j)})\right\|_2=\left\|\tilde{f}^{(j)}\right\|_2$$

该式意味着 $\varepsilon_{\rm ref}^{(j)}$ 同时也是 $U^{(j)}$ 的逼近函数的 2- 范数 $\left\|\tilde{f}^{(j)}\right\|_2$ 的近似值, $1/\varepsilon_{\rm ref}^{(j)}$ 即为该近似值的倒数, 因此该倒数也意味着利用各输出变量的逼近函数值作为相应的各局部误差估计值的加权值, 能使各个输出变量的局部误差估计值均能在统一的局部误差估计值 $\varepsilon_{\mathbf{k}}$ 中得到合理的体现. 上述 $\varepsilon_{\rm ref}^{(j)}$ 的作用及合理性的解释同样适用于统一的全局误差估计值 η.

基于上述统一的局部误差估计值和统一的全局误差估计值, 只要对第 2.2 节 A-SPAM 的指标添加算法和终止判据分别进行如下修改即可得到 NA-SPAM-MOP.

修改 1: 用式 (4.3) 代替式 (2.7) 计算指标 $\mathbf{k}$ 的局部误差估计值 $\varepsilon_{\mathbf{k}}$, 就可得到 NA-SPAM-MOP 在多输出变量情形下的指标添加算法. 由于修改后的 $\varepsilon_{\mathbf{k}}$ 和指标添加算法与具体的输出维度无关, 所以各输出变量的指标集总是添加相同的指标, 只要设置各输出变量的初始指标集都为 $\{(1,\ldots,1)\}$, 就可保持各输出变量的指标集始终相同, 从而实现指标集的统一.

修改 2: 用式 (4.5),(4.6) 代替式 (3.12) 计算全局误差估计值, 就可得到 NA-SPAM-MOP 在多输出变量情形下的终止判据.

4.3. NA-SPAM-MOP 计算流程

基于 4.2 节对于 A-SPAM 的修改, 可以得到 NA-SPAM-MOP 的计算流程, 如图 1 所示.

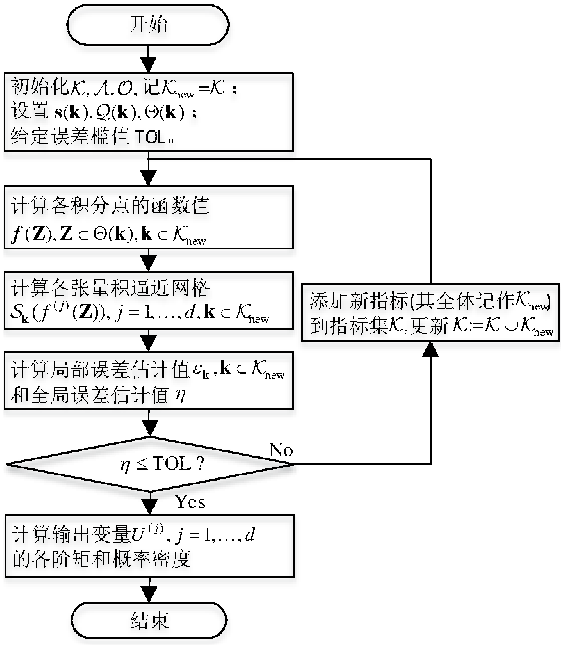

图 1 NA-SPAM-MOP 的流程图

5. 算例分析

本文算法均采用 C++ 编写, 计算环境为: CPU Intel Core i3-4160 3. 60GHz 的台式机.

在本文算例中 $\mathbf{s}(\mathbf{k})$ 的每个分量 $s^{(i)}(k_i)=k_i-1$, $\mathcal{Q}_{\mathbf{k}}$ 的每个分量 $\mathcal{Q}_{k_i}^{(i)}$ 为 $s^{(i)}(k_i)+1$ 点高斯积分规则.

5.1. NA-SPAM 与 A-SPAM 的逼近误差估计值的对比

本算例验证新型终止判据相比于现有终止判据对逼近误差的估计更准确, 进而验证在给定相同的逼近误差槛值时由 NA-SPAM 得到的逼近函数相比于 A-SPAM 得到的逼近函数具有更接近误差槛值的逼近误差. 同时, 本算例还验证了新型终止判据所依赖的假设 2 的合理

性. 在研究数值积分、插值以及多项式逼近方法的计算效率时, 通常选取 Genz 函数族中的函数作为算法的测试函数 [17]. Genz 函数族包含振荡型、乘积峰型、角落峰型、高斯型、连续型和不连续型这 6 类, 其中最常使用的为高斯型. 因此本算例使用高斯型测试函数作为输出变量与输入随机变量 $\mathbf{Z}$ 的函数, 其具体表达式为:

$$f(\mathbf{Z}) = d \cdot \exp\left(-\sum_{i=1}^{d} v_i^2 (Z_i - o_i)^2\right), \tag{5.1}$$

式中 d 为随机变量 $\mathbf{Z}$ 的维数, v_i, o_i 为函数参数.

设定 $d = 5$, 让参数 v_i, o_i 在区间 $[0, 1]$ 上均匀地随机产生; 设各输入随机变量相互独立, Z_1, Z_2 分别服从区间 $[-1, 1]$ 和 $[-0.5, 0.5]$ 上的均匀分布, Z_3, Z_4 分别服从均值为 0、标准差为 1 和均值为 0、标准差为 0. 7 的正态分布, Z_5 服从参数为 1 的指数分布. 逼近误差 (3.2) 可看成随机变量 $\tilde{f}(\mathbf{Z}) - f(\mathbf{Z})$ 的 2 阶矩的平方根, 可用 10^5 次蒙特卡罗抽样得到.

随机产生一组参数 v_i, o_i , 就得到一个函数 $f(\mathbf{Z})$, 分别使用 A-SPAM、NA-SPAM 自适应迭代地逼近该函数并取其中 9 次迭代结果, 就可得到逼近函数在 9 次迭代时的状态, 相应地可以得到逼近函数在 9 次迭代时的逼近误差、A-SPAM 的现有判据及 NA-SPAM 的新型判据分别在 9 次迭代时各自对逼近误差的估计值 (分别为式现有全局误差估计值 (3.12) 和新的全局误差估计值 (3.11)). 本文共产生 100 组参数, 得到 100 个函数, 进而可得到 100 个逼近函数、900 个逼近误差、现有判据和新型判据对 900 个逼近误差的估计值.

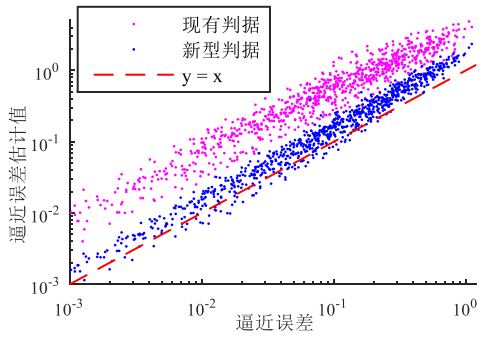

图 2 现有判据和新型判据对逼近误差的估计值的对比

图 2 为 NA-SPAM 与 A-SPAM 各自得到的所有逼近误差及其估计值的散点图, 图中共 1800 个点, 每个点的横坐标为逼近误差值, 纵坐标为该逼近误差的估计值, 红色参考线 $y = x$ 表示逼近误差的估计值与逼近误差值相等, 散点越接近该参考线, 则散点对应的估计越准确. 从图 2 可以看出新型判据的散点相较现有判据的散点, 与参考线的偏离程度更小且更加集中于该参考线附近, 这表明新型判据比现有判据对逼近误差的估计更准确. 进而发现现有判据的所有散点和新型判据 90.7% 的散点都位于参考线的上方, 即这些散点的纵坐标值大于横坐标值, 它们对应的逼近误差估计值大于相应要求的逼近误差值, 所以逼近误差的估计值绝大多数情况下是偏大的.

图 3 和图 4 为 100 个函数中某两个函数的逼近误差及其估计值, 图 5 为 100 个函数的逼近误差及其估计值的均值, 每幅图中的 9 个正方形点分别对应着逼近函数在这 9 次迭代时的状态, 其横坐标表示计算时间, 纵坐标表示逼近误差值, 与这些点的横坐标相同的其他点对应

着现有判据或者新型判据误差估计, 其纵坐标表示对相应的逼近误差的估计值. 从图 3 、图 4 以及图 5 可以看出, 新型判据对逼近误差的估计值曲线比现有判据对逼近误差的估计值曲线更接近于逼近误差实际值曲线, 因此新型判据对逼近误差的估计更准确. 在图 4 中给定相同的误差槛值时, 例如 TOL = 0.03(图中黑色横向虚线), NA-SPAM 和 A-SPAM 为了使它们的逼近误差估计值分别等于误差槛值, 需要分别计算至点 b 和点 c, 即点 b 和点 c 分别表示了在给定 TOL = 0.03 时 NA-SPAM 和 A-SPAM 的计算时间和逼近误差. 由于点 b 与误差槛值线的距离小于点 c 与误差槛值线的距离, 所以 NA-SPAM 的逼近误差相比于 A-SPAM 的逼近误差更接近于给定的误差槛值 (即图中点 a). 并且点 b 和点 c 位于误差槛值线下方, 所以分别与点 c 和点 b 相对应的 A-SPAM 和 NA-SPAM 的逼近误差小于误差槛值, 即它们都已满足给定的精度要求. 又由于点 b 在点 c 的左边, 所以在满足给定的相同的精度要求时, NA-SPAM 的计算时间小于 A-SPAM 的计算时间.

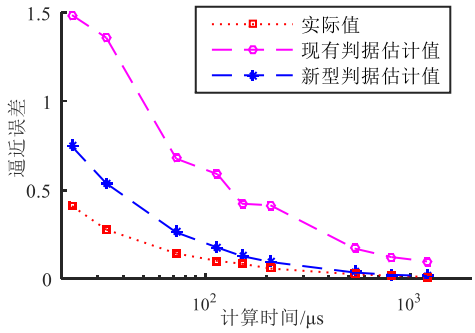

图 3 对第 1 个函数的逼近误差及其估计值

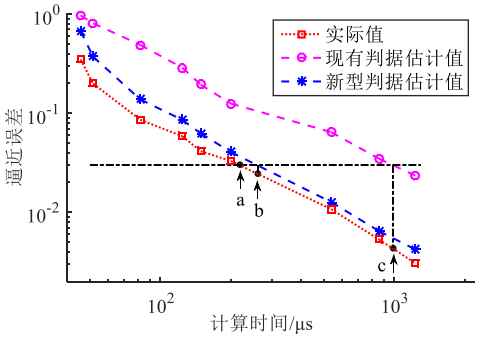

图 4 对第 2 个函数的逼近误差及其估计值

3.3 节中的命题证明了新型终止判据对逼近误差的估计相较于现有终止判据更准确, 该命题的成立部分取决于该命题的假设 1、2、3, 其中假设 1 为广泛采用的假设, 在工程中大多能够成立, 假设 3 基于假设 1 而得到, 因此也是能成立的. 假设 2 由于将逼近误差表达式 (3.2) 中的无穷集合 $\mathcal{K}^{\mathrm{c}}$ 用有限集合 $\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}$ 来近似代替, 所以其合理性需要单独验证. 假设 2 的合理性

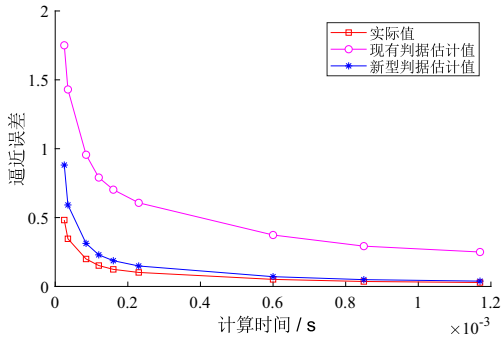

图 5 100 个函数的逼近误差及其估计值的均值

可以用代替操作后得到的逼近误差的近似值与逼近误差的真实值之比来表征:

$$r = \left\| \sum_{\mathbf{k}\in\mathcal{K}_{\mathrm{V}}^{\mathrm{c}}} \Delta_{\mathbf{k}}(f) \right\|_2 \Big/ \left\| \sum_{\mathbf{k}\in\mathcal{K}^{\mathrm{c}}} \Delta_{\mathbf{k}}(f) \right\|_2$$

r 的值越接近 1, 则假设 2 越合理.

若函数 $f(\mathbf{Z})$ 仍为 (5.1), 输入随机变量的分布与之前算例的相同 (若输入随机变量个数 $d > 5$ 且 $d \leq 10$, 则令 $Z_{i+5}, i = 1, \ldots, 5$ 的分布与 Z_i 相同). 给定 d 和误差槛值 TOL, 随机生成一组参数 v_i, o_i 则可得到一个函数 $f(\mathbf{Z})$ 和一组 NA-SPAM 的计算结果, 从而算出一个 r 值. 随机生成 100 组参数, 就得到 100 个不同性态的函数, 从而得到 100 个不同情形下 r 值 $(r_1, \ldots, r_{100})$, 取其平均值作为假设 2 合理性的表征量:

$$r_{\mathrm{aver}} = \frac{1}{100} \sum_{j=1}^{100} r_j$$

与不同的 d 和 TOL 对应的 r_{aver} 值示于表 1 中. 可以看出在 d 和 TOL 取不同的值时, r_{aver} 的值始终稍大于 1, 没有偏离 1 过多, 相应证明了假设 2 的合理性.

表 1 不同的 d 和 TOL 值对应的 r_{aver} 值

	TOL= 0.1	TOL= 0.07	TOL= 0.05	TOL= 0.02
$d = 2$	1. 029 8	1. 026 7	1. 026 6	1. 026 7
$d = 5$	1. 084 2	1. 078 4	1. 023 6	1. 041 0
$d = 8$	1. 125 3	1. 280 4	1. 190 5	1. 184 2

5.2. NA-SPAM-MOP 与 NA-SPAM 在求解多输出量化问题时的对比

本算例验证 NA-SPAM-MOP 相比于直接推广到多输出量化问题的 NA-SPAM, 在同等计算量的条件下具有更高的计算精度. 本算例所用的第一个测试函数为如下多输入多输出函数:

$$f^{(j)}(\mathbf{Z}) = d \cdot \exp\left(-\sum_{i=1}^{d} v_{i,j}^2 (Z_i - o_{i,j})^2 \right), \quad j = 1, \ldots, N_{\mathrm{o}}, \tag{5.2}$$

其中 $d=5, N_{\mathrm{o}}=5$, 各输入随机变量的分布与 5.1 节相同.

本文将 NA-SPAM 直接推广到多输出量化问题的方法称为独立计算的 NA-SPAM. 在独立计算的 NA-SPAM 中, 每个单输出量化问题都采用新型终止判据 (式 (3.11)), 第一个测试函数的误差槛值 TOL 为 0.05, 为方便比较, 设定 NA-SPAM-MOP 的计算量与独立计算的 NA-SPAM 的计算量相同 (即 NA-SPAM-MOP 达到独立计算的 NA-SPAM 的计算量时, 无论终止判据是否满足, NA-SPAM-MOP 都将停止).

随机产生一组参数 $v_{i,j}, o_{i,j}$, 就得到一组函数 $f^{(j)}(\mathbf{Z}), j=1,\ldots,N_{\mathrm{o}}$ 和相应的计算结果, 将独立计算的 NA-SPAM 和 NA-SPAM-MOP 的第 j 个输出变量的逼近函数的逼近误差分别记为 $e_1^{(j)}, e_2^{(j)}$, 如图 6 所示. 将 NA-SPAM-MOP 与独立计算的 NA-SPAM 的逼近误差比值分别记为 $e_{2,\mathrm{r}}=\sum_{j=1}^{N_{\mathrm{o}}} e_2^{(j)}/e_1^{(j)}$, 随机产生 20 组参数, 就得到 20 组函数和相应的逼近误差比值, 如图 7 所示.

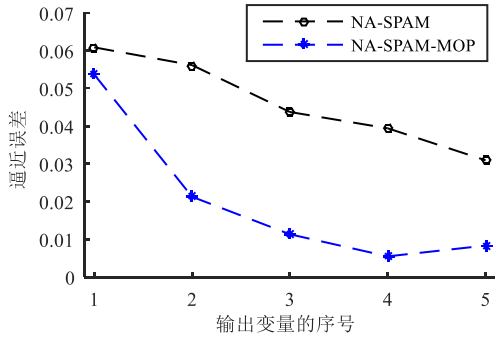

图 6 独立计算的 NA-SPAM 和 NA-SPAM-MOP 的逼近误差

从图 6 可以看出, 当计算量相同时, NA-SPAM-MOP 对各个输出变量的逼近误差相比于独立计算的 NA-SPAM 大为减小; 从图 7 也可以看出 20 组结果中的逼近误差比值 $e_{2,\mathrm{r}}$ 总是大幅小于 1, 即在 20 个不同的问题中 NA-SPAM-MOP 的逼近误差总是比相同计算量的独立计算的 NA-SPAM 小的多.

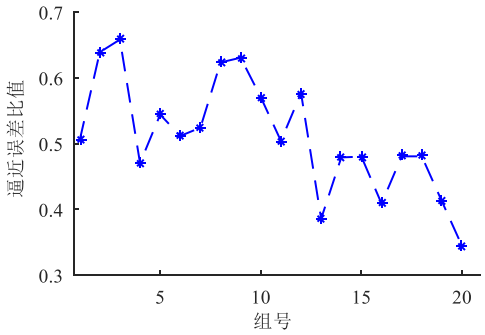

图 7 NA-SPAM-MOP 的逼近误差相对于独立计算的 NA-SPAM 的逼近误差的比值

第二个测试函数为如下含两个随机参数 c_1 和 c_2(作为输入随机变量) 的偏微分方程问题:

$$\frac{\partial u_1}{\partial t} = \frac{\partial}{\partial x}\Big[c_1\Big(\frac{\partial u_1}{\partial x} + \exp(-x)\Big) + c_2 u_1\Big] \tag{5.3}$$

$$\text{s.t. } u_1(x,0) = -\sin(\pi x),\ u_1(0,t) = u_1(1,t) = 0,\ 0 \le x \le 1 \tag{5.4}$$

$$\frac{\partial u_2}{\partial t} = \frac{\partial}{\partial x}\Big[c_1 u_2 + c_2\Big(\frac{\partial u_2}{\partial x} + \exp(x)\Big)\Big] \tag{5.5}$$

$$\text{s.t. } u_2(x,0) = \sin(\pi x),\ u_1(0,t) = u_1(1,t) = 0,\ 0 \le x \le 1 \tag{5.6}$$

函数的输出为:

$$f^{(1)}(c_1,c_2) = 10u_1(0.4,0.8),\ f^{(2)}(c_1,c_2) = 10u_2(0.4,0.8) \tag{5.7}$$

其中 c_1 服从区间 $[0.4,0.6]$ 上的均匀分布, c_2 服从均值为 0.5, 标准差为 0.1 的正态分布.

本测试函数中误差槛值设为 0.001, 同样设定 NA-SPAM-MOP 的计算量与独立计算的 NA-SPAM 的计算量相同. 独立计算的 NA-SPAM 和 NA-SPAM-MOP 的两个输出变量的逼近误差如图 8 所示.

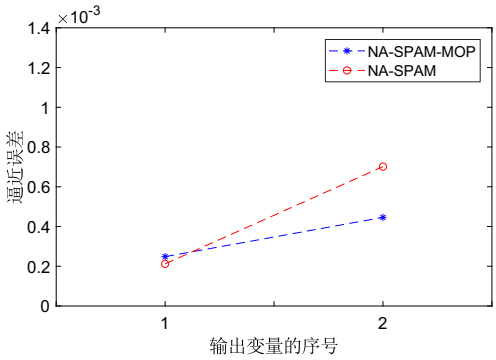

图 8 独立计算的 NA-SPAM 和 NA-SPAM-MOP 的逼近误差 (PDE)

由图 8 可知, 对于第一个输出变量, 独立计算的 NA-SPAM 和 NA-SPAM-MOP 计算出来的的逼近误差几乎相等, 而对于第二个输出变量, NA-SPAM-MOP 计算出来的的逼近误差要明显低于独立计算的 NA-SPAM. 这表明单独考虑各个输出变量的逼近误差来扩展指标集时, 由于不同输出变量的指标集不尽相同, 所以相同计算量 (抽样数) 下, 将单输出 NA-SPAM 直接推广到多输出 NA-SPAM 的方法的指标利用效率低, 总体逼近误差偏高; 而采用指标集统一策略的 NA-SPAM-MOP 更能同时考虑各输出维度的逼近误差, 故总体逼近误差更低.

由上述算例可以看出, NA-SPAM-MOP 的指标集统一策略较直接把 NA-SPAM 推广到多输出问题的方法, 由于能同时考虑到各输出变量的逼近误差, 从而选择出使各输出变量的逼近误差更小的指标, 因此在相同计算量下, NA-SPAM-MOP 的精度更高, 相应地具有更高的计算效率.

6. 结 论

本文提出了适用于多输出问题的自适应稀疏伪谱逼近新方法. 本文首先提出了 A-SPAM 的新型终止判据, 并以命题的形式证明了所提新型终止判据的有效性; 然后提出了基于此新型

终止判据的自适应稀疏伪谱逼近新方法 (NA-SPAM). 由于新型终止判据能够更准确地估计出逼近函数的当前精度, 所以 NA-SPAM 得到的逼近函数的精度更接近于预期精度. 进而, 针对多输出问题, 本文基于指标集的统一策略, 提出了适用于多输出问题的自适应稀疏伪谱逼近新方法 (NA-SPAM-MOP). 指标集的统一策略使抽样信息得到充分利用, 从而使得基于指标集统一策略的 NA-SPAM-MOP 在求解多输出问题时, 相较直接把 NA-SPAM 推广到多输出问题的方法具有更高的计算效率.

7. 附 录

7.1. 正交基函数的定义

设 $\mathbf{z}=(z_1,\ldots,z_d)$ 为定义在 d 维方形区域 $\Omega=\Omega^{(1)}\times\cdots\times\Omega^{(d)}$ 上的变量, 其中 $\Omega^{(i)}$ 为 z_i 的定义域. $h(\mathbf{z})$ 和 $r(\mathbf{z})$ 为任意两个关于 $\mathbf{z}$ 的函数, $\rho(\mathbf{z})=\rho_1(z_1)\times\cdots\times\rho_d(z_d)$ 为权函数. 则 $h(\mathbf{z})$ 与 r $(\mathbf{z})$ 的带权 $\rho(\mathbf{z})$ 内积和 $h(\mathbf{z})$ 的带权 $\rho(\mathbf{z})$ 2- 范数分别定义为:

$$\langle h(\mathbf{z}),r(\mathbf{z})\rangle=\int_{\Omega}h(\mathbf{z})r(\mathbf{z})\rho(\mathbf{z})\,\mathrm{d}\mathbf{z} \tag{7.1}$$

$$\|h(\mathbf{z})\|_2=\sqrt{\langle h(\mathbf{z}),h(\mathbf{z})\rangle} \tag{7.2}$$

Ω 上带权 $\rho(\mathbf{z})$ 的正交基函数系 $\{\Phi_{\mathbf{n}}(\mathbf{z})\}_{\mathbf{n}=\mathbf{0}}^{\infty}$ 定义为:

$$\langle\Phi_{\mathbf{m}}(\mathbf{z}),\Phi_{\mathbf{n}}(\mathbf{z})\rangle=\begin{cases}\gamma_{\mathbf{n}}, & \mathbf{m}=\mathbf{n},\\ 0, & \mathbf{m}\neq\mathbf{n}\end{cases} \tag{7.3}$$

$$\Phi_{\mathbf{n}}(\mathbf{z})=\prod_{i=1}^{d}\phi_{i,n_i}(z_i) \tag{7.4}$$

式中 $\mathbf{m}=(m_1,\ldots,m_d),\mathbf{n}=(n_1,\ldots,n_d)$ 分别为正交基函数 $\Phi_{\mathbf{m}}(\mathbf{z}),\Phi_{\mathbf{n}}(\mathbf{z})$ 的次数; $\gamma\mathbf{n}$ 为 $\Phi_{\mathbf{n}}(\mathbf{z})$ 带权 $\rho(\mathbf{z})$ 2- 范数的平方; $\phi_{i,n_i}(z_i)$ 为单变量 z_i 的 n_i 次正交多项式, 它满足如下正交条件:

$$\int_{\Omega^{(i)}}\phi_{i,m_i}(z_i)\phi_{i,n_i}(z_i)\rho_i(z_i)\mathrm{d}z_i=\begin{cases}\gamma_{i,n_i}, & m_i=n_i,\\ 0, & m_i\neq n_i\end{cases} \tag{7.5}$$

结合式 (7.3) 和 (7.5) 可得 $\gamma_{\mathbf{n}}=\gamma_{1,n_1}\times\cdots\times\gamma_{d,n_d}$.

7.2. 各类指标和指标集的定义

指标 (index): $\mathbf{k}=(k_1,\ldots,k_d)$ 为 d 维多重指标 (简称指标), k_i 为指标 $\mathbf{k}$ 的第 i 个分量.

指标的大小: 指标 $\mathbf{k}$ 大于指标 $\mathbf{t}$ 是指 $\mathbf{k}$ 的每个分量 k_i 都大于等于 $\mathbf{t}$ 的对应分量 t_i , 且存在某个维度 j , 在该维度上有 $k_j>t_j$.

后邻指标: 指标 $\mathbf{k}$ 的后邻指标为 $\mathbf{k}-\mathbf{e}_i,i=1,\ldots,d(\mathbf{k}-\mathbf{e}_i\in\mathbb{N}_1^d)$, 其中 $\mathbf{e}_i$ 为单位指标, 它的第 i 个分量为 1, 其余分量都为 0.

前邻指标: 指标 $\mathbf{k}$ 的前邻指标为 $\mathbf{k}+\mathbf{e}_i,i=1,\ldots,d$.

指标间的距离: 指标 $\mathbf{k}$ 与指标 $\mathbf{t}$ 间的距离定义为 $\sum_{i=1}^{d}|k_i-t_i|$. 与指标 $\mathbf{k}$ 距离最小的指标为 $\mathbf{k}$ 的前邻指标和后邻指标, 最小距离为 1.

指标集 (index set): 由指标构成的集合称为指标集, 记为 $\mathcal{K}$. 在传统的 Smolyak 稀疏网格中, 指标集 $\mathcal{K}$ 只能取如下受限制的形式:

$$\mathcal{K}=\{\mathbf{k}:l+1\leq\|k\|_1\leq l+d\} \tag{7.6}$$

式中 l 为稀疏网格的阶次. 而在 A-SPAM 所用的一般化 Smolyak 稀疏网格中, $\mathcal{K}$ 可以为任意准许集.

准许集 (admissible set): 对于指标集 $\mathcal{K}$ 中的任意指标 $\mathbf{k}$, 若 $\mathbf{k}$ 的所有后邻指标都属于 $\mathcal{K}$, 则称 $\mathcal{K}$ 为准许集.

7.3. Smolyak 稀疏网格的构建过程

张量积算子 $\mathcal{L}_{\mathbf{k}}$ 定义为 [13]:

$$\mathcal{L}_{\mathbf{k}}=\mathcal{L}_{k_1}^{(1)}\otimes\cdots\otimes\mathcal{L}_{k_d}^{(d)} \tag{7.7}$$

式中 $\mathbf{k}=(k_1,\ldots,k_d)$ 为 d 重指标, $\mathcal{L}_{k_i}^{(i)},i=1,\ldots,d$ 为第 i 个维度的 k_i 阶单维算子, $\otimes$ 为张量积运算符.

张量积差分算子 $\Delta_{\mathbf{k}}$ 定义为 [13]:

$$\Delta_k=\Delta_{k_1}^{(1)}\otimes\cdots\otimes\Delta_{k_d}^{(d)}=(\mathcal{L}_{k_1}^{(1)}-\mathcal{L}_{k_1-1}^{(1)})\otimes\cdots\otimes(\mathcal{L}_{k_d}^{(d)}-\mathcal{L}_{k_d-1}^{(d)}) \tag{7.8}$$

式中 $\Delta_{k_i}^{(i)},i=1,\ldots,d$ 为第 i 个维度的 k_i 阶单维差分算子, 其定义为 $\Delta_{k_i}^{(i)}=\mathcal{L}_{k_i}^{(i)}-\mathcal{L}_{k_i-1}^{(i)},\Delta_0^{(i)}=\mathcal{L}_0^{(i)}=0$. 上式按乘法的分配律展开即可得到张量积差分算子关于各张量积算子 $\mathcal{L}_{\mathbf{k}}$c.

稀疏网格算子定义为 [13]:

$$A(\mathcal{K},d,\mathcal{L})=\sum_{\mathbf{k}\in\mathcal{K}}\Delta_{\mathbf{k}} \tag{7.9}$$

式中 $\mathcal{K}$ 为指标集. 根据张量积差分算子和张量积算子的关系, 上式也可写成

$$A(\mathcal{K},d,\mathcal{L})=\sum_{\mathbf{k}\in\mathcal{K}}c_{\mathbf{k}}\mathcal{L}_{\mathbf{k}} \tag{7.10}$$

式中 $c_{\mathbf{k}}$ 为指标 $\mathbf{k}$ 的网格系数.

不同类型的张量积算子 $\mathcal{L}_{\mathbf{k}}$ 具有不同的具体表达式. 当 $\mathcal{L}_{\mathbf{k}}$ 为张量积积分算子 (记作 $\mathcal{Q}_{\mathbf{k}}$) 时, 其表达式为:

$$\mathcal{Q}_{\mathbf{k}}(f)=(\mathcal{Q}_{k_1}^{(1)}\otimes\cdots\otimes\mathcal{Q}_{k_d}^{(d)})(f)=\sum_{q_1=1}^{a^{(1)}(k_1)}\cdots\sum_{q_d=1}^{a^{(d)}(k_d)}f(z_1^{(q_1)},\ldots,z_d^{(q_d)})\cdot\prod_{i=1}^{d}w_{q_i}^{(i)} \tag{7.11}$$

$$\mathcal{Q}_{k_i}^{(i)}(h)=\sum_{q_i=1}^{a^{(i)}(k_i)}h(z_i^{(q_i)})\cdot w_{q_i}^{(i)},i=1,\cdots,d \tag{7.12}$$

式中 $f(\mathbf{z})$ 是关于 $\mathbf{z}$ 的任意函数; $h(z_i)$ 为变量 z_i 的任意函数; $\mathcal{Q}_{k_i}^{(i)}$ 为第 i 个维度的 k_i 阶单维数值积分算子, $a^{(i)}(k_i)$ 为其积分点数, $z_i^{(q_i)}$ 和 $w_{q_i}^{(i)}$ 分别为其积分点和对应的积分权重.

对于单维数值积分算子; $\mathcal{Q}_{k_i}^{(i)}$, 其代数精度的定义为: 记数值积分式 (7.12) 对应的准确积分值为 $\mathcal{I}(h)$, 若 $h(z_i)$ 为任意次数小于等于 m 的多项式函数时, 总有 $\mathcal{Q}_{k_i}^{(i)}(h)=\mathcal{I}(h)$, 而当 $h(z_i)$ 为某个次数为 $m+1$ 的多项式函数时, $\mathcal{Q}_{k_i}^{(i)}(h)\neq\mathcal{I}(h)$, 则称数值积分算子 $\mathcal{Q}_{k_i}^{(i)}$ 的代数精度为 m.

当 $\mathcal{L}_{\mathbf{k}}$ 为张量积逼近算子 (记作 $\mathcal{S}_{\mathbf{k}}$) 时, 其表达式为:

$$\mathcal{S}_{\mathbf{k}}(f) = (\mathcal{S}_{k_1}^{(1)} \otimes \cdots \otimes \mathcal{S}_{k_d}^{(d)})(f) = \sum_{n_1=0}^{s^{(1)}(k_1)} \cdots \sum_{n_d=0}^{s^{(d)}(k_d)} \mathcal{Q}_{\mathbf{k}}(\Phi_{\mathbf{n}} \cdot f)/\gamma_{\mathbf{n}} \cdot \Phi_{\mathbf{n}}(\mathbf{z}) \tag{7.13}$$

$$\mathcal{S}_{k_i}^{(i)}(h) = \sum_{n_i=0}^{s^{(i)}(k_i)} \mathcal{Q}_{k_i}^{(i)}(\phi_{i,n_i} \cdot h)/\gamma_{i,n_i} \cdot \phi_{i,n_i}(z_i),\ i = 1, \cdots, d \tag{7.14}$$

式中 $\mathcal{S}_{k_i}^{(i)}$ 为第 i 个维度的 k_i 阶单维多项式逼近算子, $\mathcal{Q}_{\mathbf{k}}(\Phi_{\mathbf{n}} \cdot f)$ 为使用 $\mathcal{Q}_{\mathbf{k}}$ 计算内积 $\langle \Phi_{\mathbf{n}}(\mathbf{z}), f(\mathbf{z}) \rangle$ 所得的数值积分值.

在稀疏网格算子表达式 (7.9) 和 (7.10) 中, 分别用 $\mathcal{S}_{\mathbf{k}}$ 和 $\mathcal{S}_{k_i}^{(i)}$ 代替 $\mathcal{L}_{\mathbf{k}}$ 和 $\mathcal{L}_{k_i}^{(i)}$, 就可得到稀疏网格逼近算子 $A(\mathcal{K}, d, \mathcal{S})$.

参 考 文 献

[1] Mackay D J C. Introduction to Monte Carlo Methods [G]. In Jordan M I, editor, Learning in Graphical Models, Learning in Graphical Models, pages 175–204. Springer Netherlands, Dordrecht, 1998.

[2] Liu W K, Belytschko T, Mani A. Random field finite elements [J]. International journal for numerical methods in engineering, 1986, 23(10): 1831–1845.

[3] Zhang D. Stochastic methods for flow in porous media: coping with uncertainties[M]. Academic press, 2001.

[4] Xiu D, Karniadakis G E. Modeling uncertainty in flow simulations via generalized polynomial chaos [J]. Journal of Computational Physics, 2003, 187(1): 137–167.

[5] Xiu D, Karniadakis G E. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos [J]. Computer Methods in Applied Mechanics and Engineering, 2002, 191(43): 4927–4948.

[6] Xiu D. Numerical Methods for Stochastic Computations–A Spectral Method Approach[M]. Princeton University Press, 2010.

[7] Lin G, Elizondo M, Lu S, Wan X. Uncertainty Quantification in Dynamic Simulations of Large-scale Power System Models using the High-Order Probabilistic Collocation Method on Sparse Grids [J]. International Journal for Uncertainty Quantification, 2014, 4(3).

[8] Tang J, Ni F, Ponci F, Monti A. Dimension-Adaptive Sparse Grid Interpolation for Uncertainty Quantification in Modern Power Systems: Probabilistic Power Flow [J]. IEEE TRANSACTIONS ON POWER SYSTEMS, 2016.

[9] Sun Y, Mao R, Li Z, Tian W. Constant Jacobian Matrix-Based Stochastic Galerkin Method for Probabilistic Load Flow [J]. Energies, 2016, 9(3): 153.

[10] Eldred M S, Burkardt J. Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification [J]. AIAA paper, 2009, 976(2009): 1–20.

[11] Xiu D. Efficient collocational approach for parametric uncertainty analysis [J]. Commun. Comput. Phys, 2007, 2(2): 293–309.

[12] Constantine P G, Eldred M S, Phipps E T. Sparse pseudospectral approximation method [J]. Computer Methods in Applied Mechanics and Engineering, 2012, 229-232: 1–12.

[13] Conrad P R, Marzouk Y M. Adaptive Smolyak pseudospectral approximations [J]. SIAM Journal on Scientific Computing, 2013, 35(6): A2643–A2670.

[14] Smolyak S A. Quadrature and interpolation formulas for tensor products of certain classes of functions [C]. In Dokl. Akad. Nauk SSSR, volume 4 of Dokl. Akad. Nauk SSSR, page 123, 1963. 240–243.

[15] Gerstner T, Griebel M. Dimension-adaptive tensor-product quadrature [J]. Computing, 2003, 71(1): 65–87.

[16] Winokur J G. Adaptive Sparse Grid Approaches to Polynomial Chaos Expansions for Uncertainty Quantification[D]. PhD thesis, Duke University, 2015.

[17] Novak E, Ritter K. High dimensional integration of smooth functions over cubes [J]. Numerische Mathematik, 1995, 75(1): 79–97.

A NEW ADAPTIVE SPARSE PSEUDOSPECTRAL APPROXIMATION METHOD

Lin Jikeng Yuan Kaiming Shen Danfeng

(*College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*)

Luo Pingping

(*College of Electrical Engineering, Shanghai Electrical Power University, Shanghai 200090, China*)

Liu Yangsheng

(*College of Electronics and Information Engineering,Tongji University, Shanghai 201804, China*)

Abstract

The adaptive sparse pseudospectral approximation method is a highly efficient method of polynomial chaos, but there exist two defects. One defect is that its termination criterion cannot estimate its approximation error accurately, and the other is that it' s only applicable to single-output problems. A new adaptive sparse pseudospectral approximation method applicable to multi-output problems is proposed in this paper. First, a new termination criterion that can estimate the approximation error more accurately than the original is put forward and its correctness is ensured by a proposition which is strictly proved in the paper. Based on the new criterion, a new adaptive sparse approximation pseudospectral method for single-output problems is thus proposed whose approximation error is closer to the required error than the original method. Then, a new adaptive sparse pseudospectral approximation method for multi-output problems is proposed by the strategy of unifying the index sets corresponding to all output variables and using the new termination criterion. The proposed method is computationally more efficient than the methods of extending the adaptive sparse pseudospectral approximation method for single-output to the multi-output problems directly. Several mathematical cases demonstrate the effectiveness and validity of the proposed method.

Keywords: adaptive sparse pseudospectral approximation method, termination criterion, approximation error, single-output, multi-output

2010 Mathematics Subject Classification: 65D15, 65Y20