# 可定制的 B 样条建模自适应正则化技术*

David Lenz[a,*], Raine Yeh[b], Vijay Mahadevan[a], Iulian Grindeanu[a], Tom Peterka[a]

[a]数学与计算机科学部，阿贡国家实验室，莱蒙特，*60439*，伊利诺伊州，美国[b]谷歌，纽约市，纽约州，美国

arXiv:2301.01209v1 [math.NA] 3 Jan 2023

**摘要**

B样条模型是一种通过函数逼近表示科学数据集的强大方法。然而，当待逼近的数据分布不均匀时，这些模型可能会出现伪振荡。传统上，人们通过模型正则化（即平滑）来尽量减少这些振荡；遗憾的是，有时无法在充分去除不需要的伪影的同时，避免将数据集的关键特征也平滑掉。本文提出一种模型正则化方法，既能保留数据集的重要特征，又能尽量减少人为振荡。我们的方法会自动调整定义域内平滑参数的强度，在约束不足的区域去除伪影，同时保持其他区域不变。所提出的方法有选择地引入基于一阶和二阶导数的正则化项，以维持模型精度并尽量减少数值伪影。我们在由科学模拟生成的一系列二维和三维数据集上验证了该方法的表现。此外，本文还重点介绍了一个关键调节参数，并详细展示了该参数的影响。本文是我们此前发表于 2022 年国际计算科学会议（ICCS） [1]的会议论文的扩展。

关键词：B 样条、正则化、函数逼近

## 1. 引言

由科学模拟或实验测量结果汇集而成的数据集，通常表示为位置-数值对列表，其中每个数据点都由一个测量值及其对应的位置组成。根据应用场景的不同，这些点位置可以构成结构化网格、非结构化网格或互不相连的点云。分析这些数据的方法往往只适用于特定的数据布局；通常，随着点位置的几何结构变得更加一般化（例如点云），数值分析技术也会变得更加复杂。即使是看似简单的任务，例如插值，在非结构化点云上也可能计算开销很大，而在高度非均匀的网格上则可能数值精度不高。避免这些挑战的一种方法是用数学函数表示数据集，然后分析该函数，而不是原始数据。这样可以大幅简化数据点之外的插值和微分过程，简化可视化任务，并使数据重采样几乎变得轻而易举。

本文重点探讨如何使用（张量积）B 样条表示科学数据集。B样条是一类高度光滑的函数，常用于精确几何建模 [2]，也是等几何分析（IGA）的基础 [3]。近期研究表明，大型复杂极大规模科学模拟产生的数据集可以通过 B样条有效建模 [4]。非均匀有理B样条（NURBS）是 B样条的一种推广，也得到了类似的研究结果 [5]。

B样条具有多种特性，使其适合作为数据的函数表示。B样条是高阶近似函数，求值、微分和积分 B 样条模型的运算速度快且数值稳定 [6]。尤为重要的是，微分和积分可以用闭式形式计算，且不会造成额外的精度损失，这与有限差分或黎曼和不同。因此，一旦计算出足够精确的样条来表示数据，分析函数模型往往比分析原始数据更有成效。

然而，计算最佳拟合B样条需要求解一个可能病态或秩亏的线性系统。导致这种病态性的常见原因是输入数据集中包含彼此相近的稀疏点块和稠密点块。若不采取额外措施，求解该系统可能会产生在数据点之间剧烈振荡、甚至在输入数据极为稀疏的区域发散的函数。这个问题可能难以自动检测，因为误差度量通常基于原始数据与模型之间的逐点误差定义。发生在远离输入数据处的伪振荡无法通过这些度量捕捉到。

为应对这些挑战，我们开发了一种将 B样条拟合到非结构化数据的新方法，该方法能够减少或消除振荡，同时保持数据集的关键特征不变。我们的方法通过添加一个强度可变的平滑参数，对 B 样条

*本研究由美国能源部科学办公室先进科学计算研究项目资助，合同编号为 DE-AC02-06CH11357，项目经理为 Margaret Lentz。

*通讯作者。电子邮箱： dlenz@anl.gov(David Lenz)

拟合问题进行正则化；该参数会根据输入数据集的特征自动调整。这一附加项会在数据集非常稀疏的区域平滑尖峰伪影，但在数据点密集的区域不进行任何平滑，从而保留这些区域的精度。此外，即使对于边界不规则的数据集，我们的方法也能创建定义明确的样条模型；无需了解边界信息，该方法会自动处理边界之外不包含数据点的区域。

本文其余部分安排如下。第2节回顾相关思想和方法。第 3 节介绍本文用于描述 B 样条的数学细节。第4节阐述本文的主要成果：一种用于 B样条模型自适应正则化的方法。随后，我们在第5节通过一系列数值示例展示该方法的性能。第6节概述未来研究方向，并在第7节给出结论。

## 2. 相关工作

创建 B样条模型来表示非结构化数据集，是散乱数据逼近（SDA）的一个具体示例。散乱数据逼近是一个广泛的研究领域，关注定义连续函数，以插值或逼近空间中散乱分布的输入。散乱数据逼近常用于图像重建问题，例如在医学 [7], 地震 [8],或天文 [9] 成像中，实验或物理约束不允许采集均匀间隔的样本。Francis等人 [10]编撰了散乱数据逼近方法的入门比较。

病态数值方法一直是散乱数据逼近领域面临的难题，因此人们提出了多种提高数值稳定性的方法。我们的方法与散乱数据逼近中的变分方法最为相似，这类方法通过最小化逼近函数导数的大小（或称“粗糙度”）来实现逼近。Duchon [11]的早期工作便是一个典型例子。从历史上看，粗糙度最小化是通过使用平滑样条实现的；Gu [12]的著作对平滑样条进行了全面阐述。应用平滑样条需要在精度与粗糙度最小化之间权衡，因为过度惩罚粗糙度往往会降低精度。因此，大量研究致力于对这种权衡进行恰当的参数化。Craven和Wahba [13] 提出了颇具影响力的“交叉验证”方法，Gu [14]对此作了进一步拓展。

本文采用的函数逼近基于全局张量积B样条，但已有许多其他基于样条的回归方法被提出。Wood [15]采用截断薄板样条，在保持薄板回归样条特有稳定性的同时提高了其效率。Lee等人 [16] 利用B样条层次结构拟合非结构化数据点，但并未明确处理稀疏点分布所导致的不稳定性。Francis等人 [10] 提出了一个两步过程，用于将具有可变点密度的非结构化点云重采样到非结构化网格上。尽管该方法不构造函数逼近，但其重采样方法表现良好。

我们的新型自适应正则化过程最早由第二作者在其学位论文 [17]中进行探索。该方法也直接受到 El-Rushaidat 等人 [18],的工作的启发；他们在将非结构化数据重采样到结构化网格的背景下提出了两级正则化过程
然而，他们的方法需要临时选定一套标准，以便在高、低正则化强度之间切换，同时还需要根据具体应用确定整体平滑程度。我们工作的一个显著贡献是提出了一种连续变化的正则化强度（而非两级强度），并能自动进行调整。

## 3. B 样条背景

本节将简要介绍描述科学数据的 B样条模型所需的基本定义和构造方法。de Boor [6]以及 Piegl 和 Tiller [19] 的著作对 B样条的基础理论作了详尽阐述。

### *3.1. B*样条曲线

次数为 $p$ 的单变量 B 样条曲线在 $\mathbb{R}^D$ 中是一个参数化曲线

$$C(u) = \sum_{j=0}^{n-1} N_{j,p}(u) P_j, \tag{1}$$

其中，每个 $N_{j,p}$ 都是次数为 $p$ 的分段多项式函数，而每个 $P_j \in \mathbb{R}^D$ 都是 $D$ 维空间中的一个“控制点”。

B 样条基函数，记作$N_{j,p}$，定义在参数空间 $[0,1] \subset \mathbb{R}$ 上，该空间由一个非递减的 “ 节点 ” 序列划分：$t_0 \le t_1 \le \ldots\ldots \le t_{n+p} \in [0,1]$。每个基函数$N_{j,p}$在 $[t_j、\ t_{j+p+1}]$上是一个凸起函数，在其他位置为零。[1] 本文假设 B 样条的次数固定，并省略下标$p$ ，改用$j^{th}$ 函数表示为$N_j$。

为简化描述高维张量积样条时的记号，我们使用多重指标同时对多个维度中的量进行索引。多重指标 $\alpha = (\alpha^1, \ldots, \alpha^d)$ 是一个由非负指标组成的 $d$ 元组，其中 $\alpha$ 各分量之和记为 $|\alpha| = \sum_k \alpha_k$。

我们通常会考虑如下形式的多重指标指标集：

$$A = \{\text{all } \alpha \in \mathbb{N}^d \text{ such that } 0 \le \alpha^k < n_k \text{ for } 1 \le k \le d\}, \tag{2}$$

其中 $n_k$ 是预先定义的正数。我们对这些集合施加字典序，并将 $[\alpha]_A$ 记为 $\alpha$在 $A$的字典序中的索引。在接下来的章节中，我们考虑每一列对应一个多重指标的矩阵。在这种情况下，我们按字典序列出多重指标；因此，与第$j^{th}$ 列对应的多重指标 $\alpha$满足 $[\alpha]_A = j$。

[1]本文只考虑“夹持”节点序列；因此，前$p+1$ 个节点始终为 0，最后$p+1$ 个节点始终为 1。

### 3.2. 张量积$B$样条

张量积B样条是将B样条曲线自然扩展到更高维流形的一种形式，例如曲面、体积和超体积。在此，我们用 $d$ 表示张量积体积的维数，用 $D$ 表示环境空间的维数（例如，三维空间中的二维曲面对应于 $d = 2$、$D = 3$）。一个$d$维张量积B样条的参数空间为 $[0,1]^d$，该空间由 $d$ 个不同的节点向量 $\mathbf{t}_k = \{t^j{}_k\}^{n_k+}$、$_{p\,j=0}$、$k = 1$、.....、$d$划分。

给定一个元组 $u = (u^1, \ldots, u^d) \in [0,1]^d$，张量积基函数定义为

$$N_\alpha(u) = \prod_{k=1}^{d} N^k_{\alpha^k}(u^k). \tag{3}$$

其中，$\alpha$ 是如上所述的多重指标，$N^k_{\alpha^k}$ 是相对于节点向量$\mathbf{t}_k$的 $(\alpha^k)^{th}$基函数。每个维度中共有$n_k + p + 1$个节点，因此每个维度中有$n_k$ 个基函数。[2] 因此，张量积基函数的总数为 $n_{tot} = \prod^d{}_{k=1} n_k$，这也是张量积样条的控制点总数。

$\mathbb{R}^D$ 中的$d$维张量积B样条是如下形式的函数

$$C(u) = \sum_{\alpha \in A} N_\alpha(u) P_\alpha, \tag{4}$$

其中$A$ 是所有基函数的集合，且每个$\alpha$对应一个$P_\alpha \in \mathbb{R}^D$。

### 3.3. 最优控制点

给定一组节点向量和多项式次数，给定数据集的最佳拟合 B 样条由一个线性最小二乘最优化问题确定。令 $\{Q_i\}_{i=0}^{m-1}$为需要用 $\mathbb{R}^D$ 中的点拟合的点集，并用一个$d$维张量积样条进行近似。对于每个 $0 \le i < m$，令$v_i \in [0,1]^d$ 为与点$Q_i$对应的参数元组。最优控制点由以下最小二乘最优化问题确定：

$$\{\hat{P}_j\} = \underset{P_j}{\operatorname{argmin}} \sum_{i=0}^{m-1} \|Q_i - C(v_i)\|^2. \tag{5}$$

对公式（5）中的目标函数关于每个控制点求导，可将该最优化问题改写为标准形式。正规方程组可化为以下矩阵方程：

$$\mathbf{N}^T\mathbf{N}\mathbf{P} = \mathbf{N}^T\mathbf{Q}, \tag{6}$$

其中

$$\begin{aligned} \mathbf{N}_{ij} &= N_\alpha(v_i) \quad \text{where } [\alpha]_A = j \\ \mathbf{P}_{ij} &= P^j_\alpha, \qquad \text{where } [\alpha]_A = i \\ \mathbf{Q}_{ij} &= Q^j_i \end{aligned} \tag{7}$$

[2]为简单起见，此处假设 B 样条在各个维度上的次数相同，但实际应用中也可以根据需要采用不同次数。

上述方程中的上标表示向量$P_\alpha$ 和$Q_i$的分量索引。$\mathbf{N}$ 是一个 $m \times n_{tot}$ 矩阵，通常称为“B 样条配点矩阵”；$\mathbf{P}$ 是一个$n_{tot} \times D$ 矩阵，每一行包含一个控制点；$\mathbf{Q}$ 是一个$m \times D$ 矩阵，每一行包含一个输入点。

通常，矩阵$\mathbf{N}$ 非常稀疏，可以用迭代法或稀疏直接求解器求解该系统。然而，正如我们将在下一节中所示，当输入点$P_i$的样本密度随区域而变化时，该系统的条件很差。

## 4. 自适应正则化

使用张量积 B 样条对非结构化数据进行建模时，一个重大挑战是拟合过程的（病态）条件性。一般而言，张量积 B 样条模型可能会因输入数据稀疏区域的过拟合而产生强烈振荡（见图 1）。我们的自适应正则化方法能够为原本会秩亏的系统生成唯一解，并改善系统的整体条件性。在实际测试中，自适应正则化模型的振荡伪影较少，也没有表现出任何发散行为。与标准正则化技术不同，我们的方法不会不加区分地平滑模型，而只对需要平滑的区域进行正则化。

该技术采用空间变化的正则化强度，并根据输入数据点相对于 B样条节点的位置自动计算。一般而言，在几乎没有或完全没有输入数据的区域，正则化强度会增大；在输入点“饱和”的区域，正则化强度则会减小（甚至降至零）。当某个定义域区域内的正则化强度处处为零时，该区域不会进行平滑；因此，自适应正则化过程能够保留定义域中密集采样区域内存在的任何尖锐特征。

### 4.1. 二阶导数正则化

标准粗糙度最小化可表述为一个惩罚最小二乘最小化问题，与公式 (5) 类似。惩罚项使用一个新参数对每个点处二阶导数的大小进行加权，我们将该参数记为$\lambda > 0$。正则化样条的控制点定义如下：

$$\{\hat{P}_j\} = \underset{P_j}{\operatorname{argmin}} \left( \sum_{i=0}^{m-1} \|Q_i - C(v_i)\|^2 + \lambda^2 S(C) \right), \tag{8}$$

其中 $S(C)$ 用于近似$C$的二阶导数大小。

令$w_\alpha \in [0,1]^d$ 为使$N_\alpha$取最大值的参数。令 $\partial^\delta$ 表示一个偏导数，其中每个维度上的导数阶数由以下内容的分量给出：

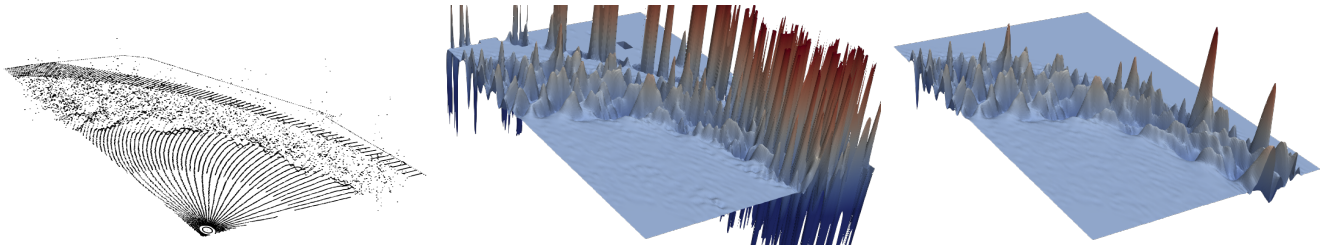

图 1：左：点密度不均匀的数据集。中：未经正则化的最佳拟合B样条模型。右：采用自适应正则化的B样条模型。中间的图像经过了裁剪；该模型中的尖峰伪影延伸到了帧外很远处。

多重指标 $\delta$.[3] 我们将$S(C)$定义为

$$S(C) = \sum_{\alpha \in A} \sum_{|\delta|=2} \left\| \partial^{\delta} C(w_{\alpha}) \right\|^2. \tag{9}$$

请注意，上述求和是对所有二阶导数求和，包括混合偏导数。

公式（8）可以像公式（5）一样转换为方程组。唯一额外的步骤是计算$S(C)$关于控制点的导数。其矩阵形式下的新方程组为

$$\begin{pmatrix} \mathbf{N}^T & \lambda \mathbf{M}^T \end{pmatrix} \begin{pmatrix} \mathbf{N} \\ \lambda \mathbf{M} \end{pmatrix} \mathbf{P} = \mathbf{N}^T \mathbf{Q} \tag{10}$$

其中

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{\delta_1} \\ \vdots \\ \mathbf{M}_{\delta_k} \end{pmatrix}, \quad \text{and } (\mathbf{M}_{\delta})_{i,j} = \partial^{\delta} N_{\beta}(w_{\alpha}), \tag{11}$$

对于 $[\alpha]_A = i$、 $[\beta]_A = j$。直观而言，$\mathbf{M}_{\delta}$ 的每一列都描述了单个 B 样条基函数的 $\partial^{\delta}$ 偏导数。矩阵$\mathbf{M}$ 由所有单独的$\mathbf{M}_{\delta}$ 矩阵拼接而成，其中 $|\delta| = 2$。$\mathbf{N}$、$\mathbf{P}$和$\mathbf{Q}$ 的定义与第 3 节相同。

我们的自适应正则化方案的创新改进在于：通过改变$\mathbf{M}$每一列的 $\lambda$ 大小，来修改上述方程组。由于该矩阵的每一列对应一个 B样条基函数和一个控制点，因此改变 $\lambda$ 的大小，就能分别调整施加于样条各个控制点上的平滑条件。得益于 B样条的局部支撑性质，对于给定区域内的控制点设置 $\lambda^{(2)}{}_j = 0$，可以“禁用”该区域内的正则化，同时仍允许在其他位置进行平滑。代数上，我们将标量参数 $\lambda$ 替换为对角矩阵 $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1^{(2)}, \ldots, \lambda_{n_{tot}}^{(2)})$，其中每个 $\lambda^{(2)}{}_j \geq 0$，并考虑新的线性系统

$$\begin{pmatrix} \mathbf{N}^T & (\mathbf{M}\mathbf{\Lambda})^T \end{pmatrix} \begin{pmatrix} \mathbf{N} \\ \mathbf{M}\mathbf{\Lambda} \end{pmatrix} \mathbf{P} = \mathbf{N}^T \mathbf{Q}. \tag{12}$$

[3]例如，$\partial^{(2,0)} f = \partial^2 f / \partial x_1^2$，以及 $\partial^{(0,2)} f = \partial^2 f / \partial x_2^2$，而 $\partial^{(1,1)} f = \partial^2 f / (\partial x_1 \partial x_2)$。

[4]每个 $\lambda_j$ 上的上标“(2)”表示该项应用于二阶导数的情形。

每个 $\lambda^{(2)}{}_i$ 的值会根据输入数据点与 B样条节点之间的相对位置自动计算。

为了更好地控制此函数，我们引入一个由用户指定的参数，称为“正则化阈值”，记作 $s^*$。调整正则化阈值可以改变判定标准，从而决定定义域中的哪些区域需要平滑，哪些区域不需要平滑。随着 $s^*$ 增大，平滑约束将应用于定义域中越来越大的区域。

令 $s_j$ 表示 $j^{th}$ 列和 $\mathbf{N}$，并令 $\widetilde{s}_j$ 表示 $j^{th}$ 列和 $\mathbf{M}$。给定 $s^* \geq 0$，我们将二阶导数局部正则化强度 定义为

$$\lambda_j^{(2)} = \frac{\max(s^* - s_j, 0)}{\widetilde{s}_j}. \tag{13}$$

因此，$\mathbf{\Lambda}$ 的定义方式使得 $\begin{pmatrix} \mathbf{N} \\ \mathbf{M}\mathbf{\Lambda} \end{pmatrix}$ 的每个列和都不小于 $s^*$。

以这种方式调整正则化强度 $\lambda_j$ 会带来若干重要结果。如果 $s^* = 0$，那么 $\mathbf{\Lambda} = 0$，最小化问题就会变成通常的最小二乘问题。当 $s^*$ 较小时，除非$\mathbf{N}$ 的第 $j^{th}$ 列和较小（这表明系统病态），否则 $\lambda_j$ 将为零。此处，自适应正则化只会平滑约束不足的控制点。

这一公式也解释了为什么自适应正则化方法能够保留清晰且密集采样的特征，同时平滑振荡伪影。在定义域中采样密集的区域，控制点会受到许多数据点的约束，因此$\mathbf{N}$ 中对应的列和相对较大。通过将 $s^*$ 设得足够小，该区域中的所有控制点的正则化强度都将为零；即 $\lambda_j = 0$。因此，该区域中的最佳拟合样条不会被人为平滑。

### *4.2.* 一阶导数正则化

上述自适应正则化框架以二阶导数为基础定义，但很容易扩展到其他阶的导数。我们发现，同时约束一阶和二阶导数，在对某些区域内完全没有点值的数据集进行建模时尤其有用。这种情况通常发生在数据表示的对象具有内部孔洞或不规则边界时。

一般而言，采用二阶导数的自适应正则化会在输入数据稀疏且拟合问题病态的区域生成曲率最小的模型，

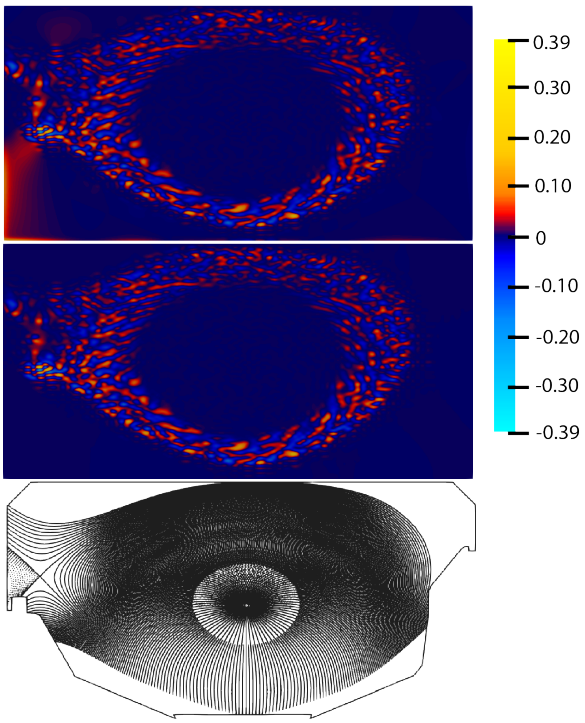


图2：在正则化方案中使用一阶导数项的效果。顶部：仅含二阶导数的XGC数据集正则化。中部：使用一阶和二阶导数的XGC数据集正则化。底部：XGC数据集的点分布。数据说明见第5节。

在没有数据点的区域，正则化模型仍可能表现出低频、低曲率振荡。作为这一现象的示例，图2展示了一个高度非均匀数据集[5]，以两种方式进行正则化后的B样条模型。仅含二阶导数的正则化（顶部）会在左边缘产生令人分心的“涂抹”效果；数据集在此处呈现出尖峰（注意其中的小块黄色区域），且这些尖峰紧邻一个没有数据来约束模型的区域。相比之下，同时包含一阶和二阶导数项的正则化（中部）可以减轻这一数值伪影。

为了在正则化方案中加入一阶和二阶导数项，方程（12）中的最小二乘最优化问题修改为

$$\begin{pmatrix}\mathbf{N}^T & (\mathbf{M_2\Lambda_2})^T & (\mathbf{M_1\Lambda_1})^T\end{pmatrix}\begin{pmatrix}\mathbf{N}\\ \mathbf{M_2\Lambda_2}\\ \mathbf{M_1\Lambda_1}\end{pmatrix}\mathbf{P} = \mathbf{N}^T\mathbf{Q}, \qquad (14)$$

with

$$\mathbf{M_2} = \begin{pmatrix}\mathbf{M}_{\delta_1}\\ \vdots\\ \mathbf{M}_{\delta_k}\end{pmatrix}, \quad \text{where } |\delta_i| = 2,$$
$$\mathbf{M_1} = \begin{pmatrix}\mathbf{M}_{\delta'_1}\\ \vdots\\ \mathbf{M}_{\delta'_l}\end{pmatrix}, \quad \text{where } |\delta'_i| = 1, \qquad (15)$$

[5]所示数据是第5.1节中介绍的“XGC Fusion”数据。

矩阵 $\mathbf{M}_\delta$ 的定义与第4.1节相同。具体而言，每个 $\mathbf{N}_\delta$ 都是一个针对 $\delta$ 阶偏导数的正则化条件矩阵。$\mathbf{M_1}$ 是所有满足 $|\delta| = 1$ 的 $\mathbf{M}_\delta$ 的拼接；换言之，它是一个包含所有一阶偏导数正则化条件的矩阵。同样，$\mathbf{M_2}$ 是一个包含所有可能的二阶偏导数正则化项的矩阵。[6]

尽管一阶导数项有助于减轻不必要的视觉伪影，但通常不适合在整个定义域内施加这些约束。由于一阶导数正则化会惩罚非零梯度，因此，当信号不是常量时，这种条件与数据并不一致。数据集的点密度较低、但所建模的函数具有陡峭梯度，这种情况并不少见。在这种情形下，拟合算法无法在不过拟合数据的同时，既最小化梯度又准确地建模信号。因此，选择在定义域的哪些位置应用一阶导数正则化至关重要。

我们对此问题的解决方案只需对迄今所述的方法稍作修改。由于添加一阶导数项主要有助于减少没有输入数据的空间区域中的振荡，因此我们修改该方法，使这些项仅在此类区域中生效。为此，我们将至少包含一个数据点的区域中的一阶导数项局部正则化强度设为零。类比公式(13)，我们将一阶导数局部正则化强度 定义为

$$\mathbf{\Lambda_1} = \begin{pmatrix}\lambda_1^{(1)} & & \\ & \ddots & \\ & & \lambda_{n_{tot}}^{(1)}\end{pmatrix}, \quad \text{where} \quad \lambda_j^{(1)} = \begin{cases}0 & \text{if } s_j > 0\\ s^* & \\ \widetilde{s}_j & \text{if } s_j = 0\end{cases}, \qquad (16)$$

并且与之前一样，$s_j$是$j^{th}$列的$\mathbf{N}$列和，而 $\widetilde{\ }s_j$是$j^{th}$列的$\mathbf{M_1}$列和。

这个定义值得进一步讨论。回顾一下，矩阵$\mathbf{N}$ 是 B 样条配点矩阵（参见公式 (7)），其中每一列代表一个基函数，每一行代表一个输入点。因此，只有当基函数$j$的局部支撑中至少有一个输入点时，$j^{th}$ 列的$\mathbf{N}$ 列和才为正。若 $s_j = 0$，则必然没有输入数据可以约束$j^{th}$ 控制点。这正是图2所示的情形，此时一阶导数正则化的作用最为显著。最后需要指出的是，当$s_j = 0$时，公式 (16) 中的定义与公式 (13) 中先前给出的二阶导数正则化强度定义一致。

## 5. 结果

我们通过一系列数值实验展示了所提方法的有效性。首先，我们将自适应正则化与均匀正则化进行比较，后者的平滑参数由人工选定。接下来，我们研究如何从稀疏样本重建解析信号，并改变稀疏度。对于每个稀疏度，我们报告误差

[6]$\mathbf{M_2}$ 与第4.1节中给出的矩阵$\mathbf{M}$ 相同，只是此处重新标记了它，以匹配$\mathbf{M_1}$的记号。

以及未正则化模型和自适应正则化模型的条件数。随后，我们测试自适应正则化在某些区域没有数据的数据集上的表现。在这些问题中，我们构建一个外推至没有逐点约束区域的 B 样条模型，并检验自适应正则化方法能否产生合理的结果。

### *5.1.* 数据集

我们在一组具有不同特征的二维和三维点云上研究了自适应正则化方法的性能。其中一些数据集由解析函数采样得到，因此我们可以计算相对于真实值的逐点误差；另一些数据集则由科学实验和模拟生成。

二维 *Polysinc*。 Polysinc[7] 数据集是一个二维点云，对函数进行采样

$$f(x,y) = \operatorname{sinc}\left(x^2+y^2\right)\operatorname{sinc}\left(2(x-2)^2+(y+2)^2\right).$$

从盒状区域 $[-4\pi,4\pi]\times[-4\pi,4\pi]$ 中均匀采样 360,000 个点位置，但四个圆盘形区域的采样率 50× 较低。

二维 *Polysinc*（修改版）。 修改版 Polysinc 在非均匀分布点云上对上述函数 $f(x,y)$ 进行采样。该数据集由盒状区域$[-4\pi,4\pi]\times[-4\pi,4\pi]$中的 22,500 个位置组成，其中 (+, +) 象限的点密度最低，而 (–, –) 象限的点密度最高。图3展示了该数据集的采样位置。

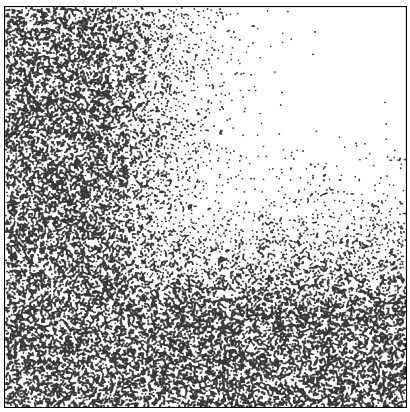

图3：修改后的多项式 sinc 数据集中的点分布。

*XGC Fusion*。 XGC 聚变数据集表示托卡马克聚变模拟中单个极向平面上静电势的归一化导数。该数据集包含 56,980 个点，具有不规则边界，由 XGC 代码 [20] 通过对陀螺动力学方程进行仿真生成。

*CMIP6* 气候。CMIP6 气候数据集表示南极洲周围一个投影框选区域内的海洋表面温度。该数据集包含 585,765 个点，中心有一个大空洞（代表南极洲），由耦合模式比较计划（CMIP6） [21] 模拟生成。

*sahex* 核。 sahex 核数据集源自对单个核反应堆部件进行的仿真，由 SHARP 工具包 [22]生成。三维数据由一个六棱柱界定，$z$ 维度的点密度比$x$ 和$y$维度更低。该数据集包含 63,048 个点。

### *5.2.* 自适应正则化与均匀正则化的比较

对整个模型应用统一的正则化强度可能会产生不理想的结果，因为在充分平滑振荡伪影的同时，也可能抹平尖锐特征。图4比较了我们的自适应正则化方案（其中 $s^*$ =6）与三种不同强度的均匀正则化。图4中的数据是 XGC 聚变数据集，其中环上有尖峰，而环内则是平坦区域。环外的数据稀疏或完全不存在。右侧的两幅图展示了采用均匀正则化的模型：正则化过弱会导致伪影（上图），过强则会抑制特征（下图）。左下方给出了我们找到的最佳均匀正则化强度，但即使在这个例子中，数据中的典型尖峰也被平滑削弱了。

### *5.3* 分析信号的准确性

为了量化采用自适应正则化的 B 样条模型的精度，我们考察具有高度非均匀输入数据集的振荡多项式 sinc 函数（图5）。我们展示了两个 B 样条模型：一个未经过正则化拟合，另一个采用我们的自适应正则化方法（$s^* = 1$）。两个模型的次数均为四次，并具有一个 $300 \times 300$ 控制点网格。未经正则化时，模型在样本密度较低的区域会发散；采用自适应正则化后，即使在样本密度较低的区域，模型也能准确表示信号。图5第二行给出了误差剖面的俯视图。仔细比较真实值（左上）和自适应正则化样条（右上）可以发现，尽管稀疏区域经过了平滑处理，但样条模型并未对密集区域进行人为平滑。特别是，我们的正则化过程保留了信号中鲜明的振荡和峰值。

输入数据的稀疏度会显著影响 B 样条模型的精度。表1列出了空洞中不同稀疏程度下各模型的误差。为准确观察模型在该区域的表现，我们测量了两个空洞周围方框内的误差（见图5）。当空洞内外的点密度相同时（稀疏度为= 1.0），两个模型的误差都很小。随着空洞变得更加稀疏，未正则化模型的误差增加了四个数量级，而自适应正则化模型的误差基本保持不变。

数据稀疏度也会影响最小二乘最小化问题的条件数。表1给出了各稀疏度下矩阵$\mathbf{N}$ （“无正则化”）和 $\begin{pmatrix}\mathbf{N}\\ \mathbf{M}\Lambda\end{pmatrix}$ （“正则化”）的条件数。随着稀疏度增加， $\begin{pmatrix}\mathbf{N}\\ \mathbf{M}\Lambda\end{pmatrix}$ 的条件数仍然较低且稳定，而$\mathbf{N}$ 的条件数起初较高

[7]我们考虑未归一化的 sinc 函数：$\operatorname{sinc}(x) = \sin(x)/x$，其中 $\operatorname{sinc}(0) = 1$。

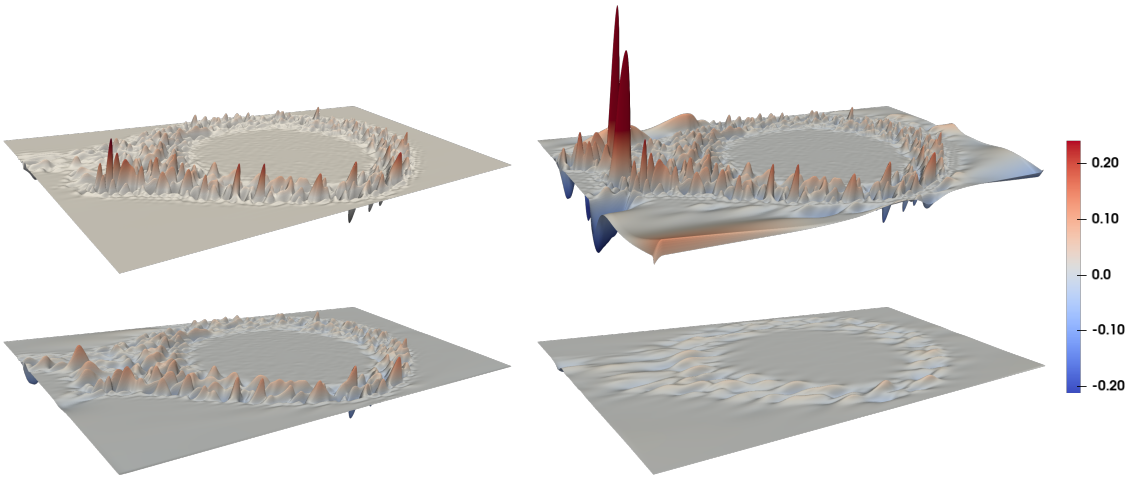


图4：均匀正则化与 自适应正则化。按顺时针方向从左上角开始：自适应正则化、采用 $\lambda = 10^{-6}$ 的均匀正则化、采用 m 正则化的均匀正则化，采用 未用10r 等的均匀的正则化。

表1：模型误差随稀疏度的变化。在两个空洞附近（见图5），分别计算自适应正则化模型和未正则化模型的最大误差与$L^2$（平均）误差。表格底部列出了两个最小化问题的条件数。

| 稀疏性 | 0.02 | 0.08 | 0.16 | 0.32 | 0.64 | 1.00 |
|---|---|---|---|---|---|---|
| 最大误差（正则化） | 3.25e-2 | 2.89e-2 | 2.71e-2 | 2.39e-2 | 1.20e-2 | 1.16e-2 |
| 最大误差（无正则化） | 1.29e2 | 6.27e2 | 2.36e-1 | 3.65e-2 | 1.20e-2 | 1.16e-2 |
| $L^2$ 误差（正则化） | 1.93e-3 | 1.53e-3 | 1.13e-3 | 7.04e-4 | 5.56e-4 | 5.44e-4 |
| $L^2$ 误差（无正则化） | 2.17e0 | 4.68e0 | 3.42e-3 | 7.35e-4 | 5.56e-4 | 5.44e-4 |
| 条件数（正则化） | 177 | 980 | 289 | 198 | 121 | 189 |
| 条件数（无正则化） | inf | inf | 6.97e4 | 2.58e3 | 1.57e3 | 3.74e3 |

最终变为无穷大。条件数使用 MATLAB 的 svds 例程计算。

### *5.4.* 正则化阈值对模型质量的影响

自适应正则化方法定义中的一个基本参数是正则化阈值，$s^*$。该参数决定定义域中的哪些区域将进行正则化，同时也为问题设定了最大的正则化强度。根据应用需求和数据集的特征，“最佳”$s^*$ 值会有所不同。不过，在本文考察的所有示例中，我们发现，$s^* \in [0, 10]$ 这一取值生成的模型最为忠实，且数值伪影最少。

为了详细说明正则化阈值对模型质量的影响，我们研究了在六种不同的$s^*$取值下，修改后的多项式 sinc 数据集模型的精度和局部正则化强度。使用包含一阶和二阶导数项的完整自适应正则化方法创建模型。该 B样条模型的次数为 3，共有 6,400 个控制点（$80 \times 80$ 网格）。

图6显示了正则化参数取值为 $s^* \in \{0.5、1.0、2.0、4.0、8.0、16.0\}$ 时的自适应正则化过程结果。真实函数的参考图像请参见图5。我们观察到，正则化阈值为 $s^* = 0.5$ 显然不合适，因为模型出现了较大的伪峰，并且振荡超出了可见帧范围。在这种情况下，模型欠约束。相反，与 $s^* = 8$和 $s^* = 16$ 对应的模型可能过度平滑，即使在点密度较高的区域，多重正弦函数的显著特征也被模糊了（点分布请参见图3）。在此示例中，$s^*$ 取值介于 1 和 4 之间时，在精度和良态性之间实现了最佳权衡。

表2列出了六个正则化模型对应的解析误差。为了计算这些误差指标，我们在高分辨率的规则网格（$400 \times 400$）上对 B样条模型进行采样，并将模型的输出与多项式 sinc 函数给出的真实函数值进行比较。同样可以看到，最严重的$L^\infty$ 误差出现在两端。由于这六个模型都未能充分捕捉定义域边缘处低振幅、高频率的振荡，因此$L^2$ 误差没有显著变化。约束不足的模型往往会产生高度局部化的伪峰，因此$L^\infty$误差通常比$L^2$误差更能反映模型质量；后者对小邻域内较大的残差不够敏感。表2中观察到的行为差异与这一点相符。

最后，图7和图8展示了局部正则化强度随位置变化的曲线。图7显示了一阶导数局部正则化强度，除右上角外均为零，与$s^*$的值无关。这正是我们的方法所预期的，因为 $\mathbf{\Lambda_1}$ 被定义为在附近至少有一个输入点时取零。这里，

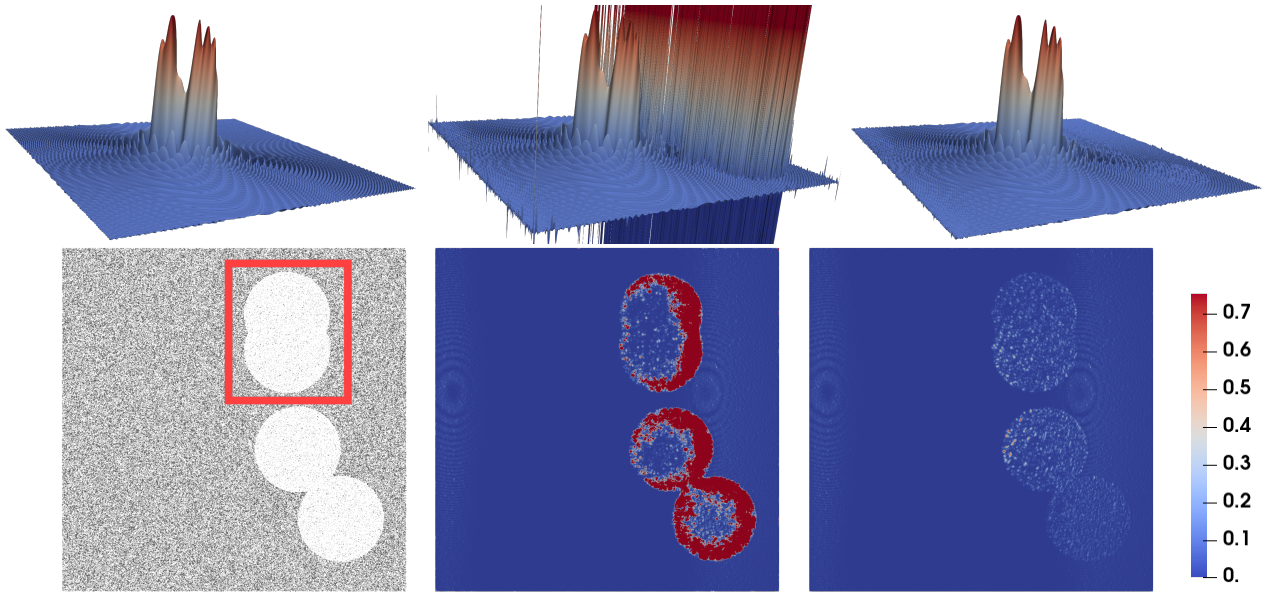


图5：上排：合成多项正弦信号（左）、无正则化的模型（中）、采用自适应正则化的模型（右）。下排：输入分布俯视图（左）、无正则化的误差剖面（中）、采用自适应正则化的误差剖面（右）。误差计算的关注区域以红色标示于左下方。

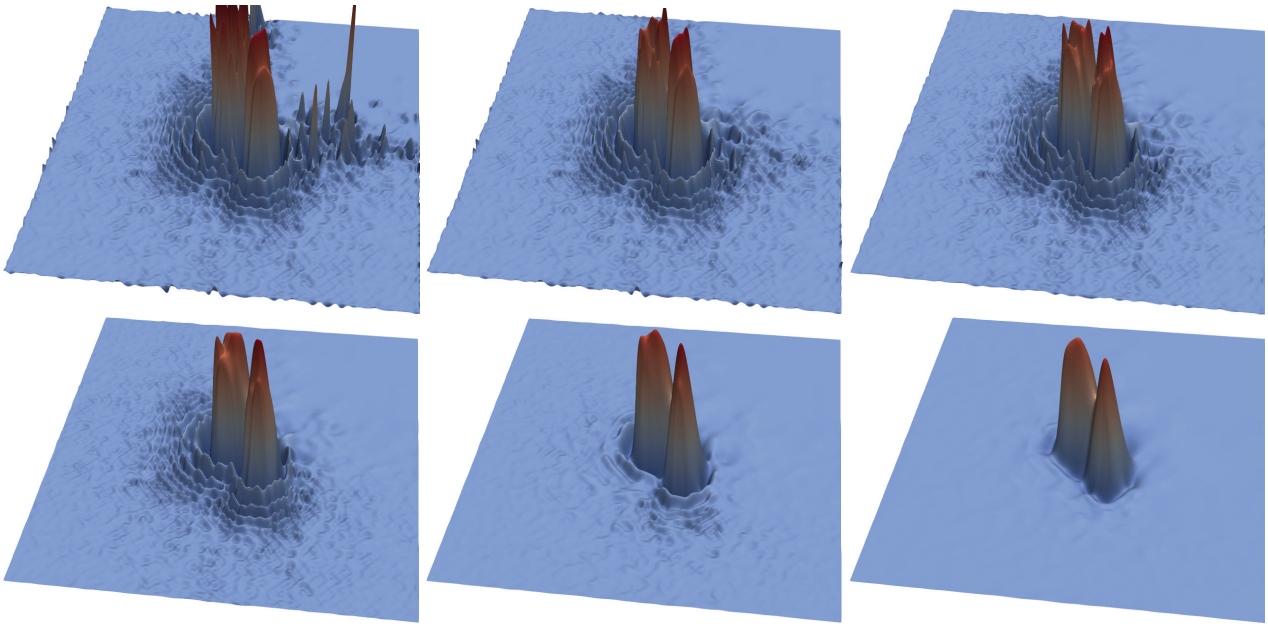

图6：修改后的多项式 sinc 数据集的 B 样条模型，正则化阈值有六种不同取值$s^*$。上排从左至右：$s^* = 0.5$、$s^* = 1.0$ 和 $s^* = 2.0$。下排从左至右：$s^* = 4.0$、$s^* = 8.0$ 和 $s^* = 16.0$。

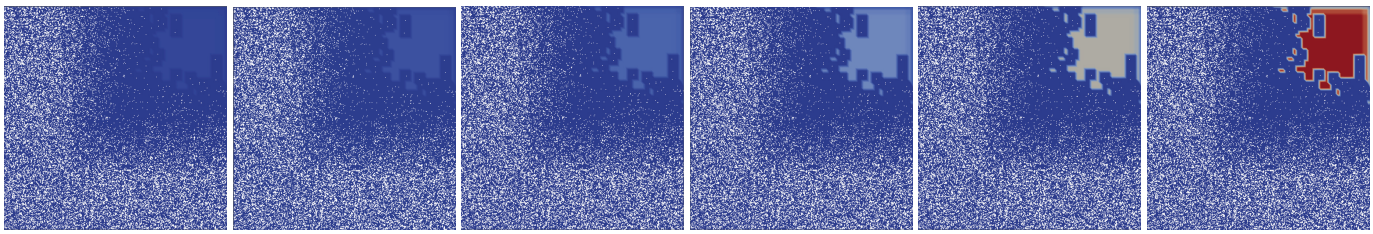

图7：图6中修改后的多项式 sinc 数据集的 B 样条模型的一阶导数局部正则化强度，正则化阈值有六种不同取值$s^*$。点位置以白色叠加显示。上排从左至右：$s^* = 0.5$、$s^* = 1.0$ 和 $s^* = 2.0$。下排从左至右：$s^* = 4.0$、$s^* = 8.0$ 和 $s^* = 16.0$。

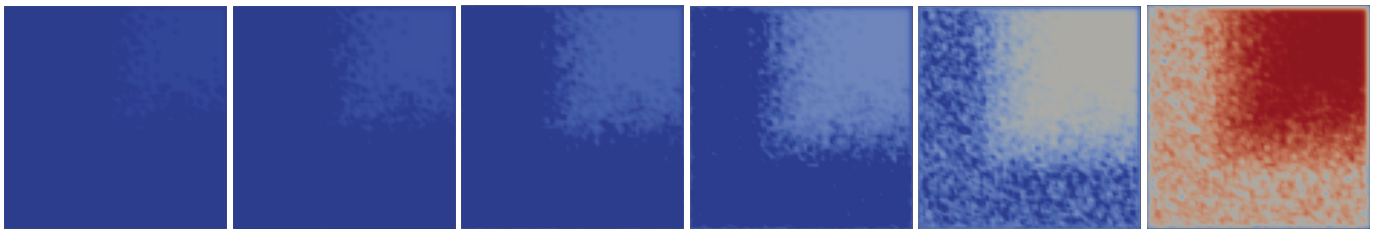

图8：二阶导数局部正　　则化强度，对应于修改后的多项式 sinc 数据集在图6中展示的 B 样条模型，其正则化阈值有六个不同取值正则化阈值 $s^*$。顶排，从左到右，　左到右：$s^* = 0.5$、$s^* = 1.0$ 和 $s^* = 2.0$。底排，从左到右：$s^* = 4.0$、$s^* = 8.0$ 和 $s^* = 16.0$。

表2：解析误差随正则化阈值 $s^*$ 的变化，对应图6所示模型。误差通过比较模型值与规则网格上各点的真实函数值计算得出。$L^2$ 误差也称为均方根误差，$L^\infty$ 误差则称为最大逐点误差。

| $s^*$ | 0.5 | 1.0 | 2.0 | 4.0 | 8.0 | 16.0 |
|---|---|---|---|---|---|---|
| $L^2$ 误差 | 0.639 | 0.265 | 0.246 | 0.267 | 0.344 | 0.438 |
| $L^\infty$ 误差 | 22.3 | 6.01 | 3.93 | 3.40 | 3.88 | 5.12 |

随着 $s^*$ 增大，局部正则化强度会在固定区域内逐渐提高。相比之下，二阶导数局部正则化强度（如图7所示）在 $s^*$ 较小时接近于零，但会随着 $s^*$ 增大而逐渐扩展至整个定义域。因此，我们可以通过增大 $s^*$ 来控制定义域中有多少部分受到正则化，其中点密度最低的区域会“最先”受到正则化。最后，我们在此指出，图6中观察到的精度不足可以通过考察二阶导数局部正则化强度来直接解释。当 $s^* = 0.5$ 且模型出现伪峰时，可以看到这些峰出现在点密度较低、局部正则化强度为零的区域。相比之下，当 $s^* = 8$ 和 $s^* = 16$ 且模型过度平滑时，我们发现，即使在点密度最高且无需正则化的左下角，局部正则化强度也非常高。

### *5.5.* 向无约束区域外推

当定义域中的大块区域没有任何数据点来约束最佳拟合 B 样条问题时，最小二乘最小化将是病态的，所得模型可能出现剧烈振荡。然而，在科学和工业应用中，包含空白区域或“空洞”的数据集十分常见。例如，一些气候模型会测量海洋温度或陆地温度，但不会同时测量两者。工业模拟通常对边界不规则的物体进行建模，而物理模拟的数据则由探测器的位置决定。

尽管在后续分析中通常会省略空白区域，但了解并控制 B 样条模型在空白区域中的行为仍然很重要。孔洞边界附近的剧烈振荡会使模型远离该边界处的导数发生偏移。此外，如果模型在空白区域表现出不可预测的行为，那么尝试计算模型的简单统计量（最小值、最大值、平均值）时可能会产生偏差。

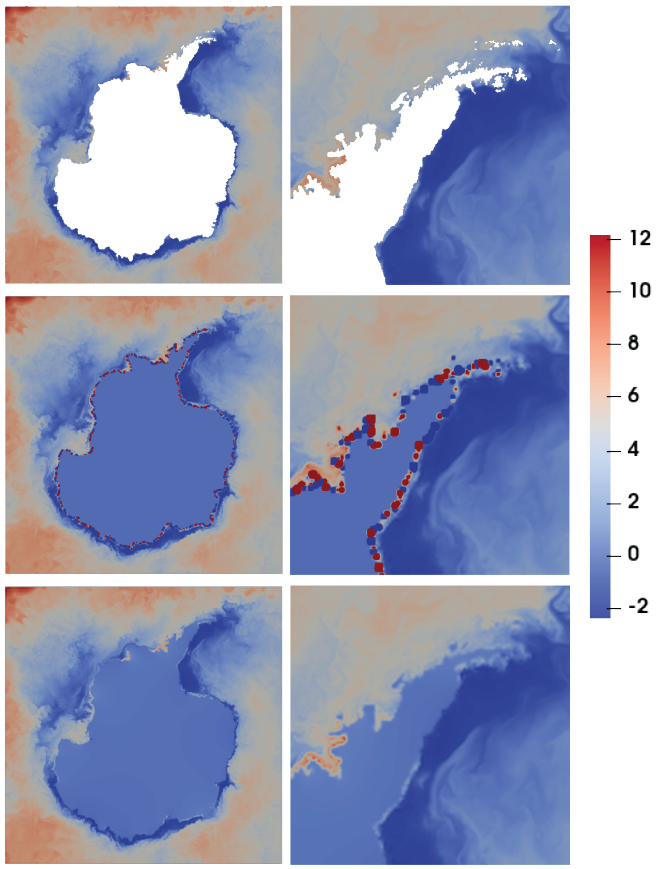


图 9：南极洲周边的海洋温度模拟。上排：六边形网格上的输入数据（左）、未进行正则化的 B 样条模型（中）、采用自适应正则化的 B 样条模型（右）。下排：上排图像的细节视图。

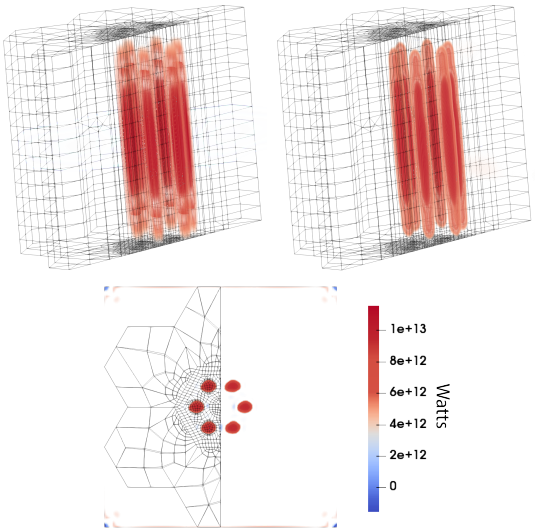


图 10：核反应堆仿真中的功率产生（瓦特）。输入数据（左）、样条模型（中）、样条模型俯视图（右）。

图 9 展示了海洋温度仿真的一个场景。该数据集以南极洲为中心，这里没有给出温度值。图中展示了两个模型，它们的阶数均为 2，并使用了一个 400 ×400 控制点网格。中间的未正则化模型在海岸线附近于 $\pm 10^8$ 之间振荡（而输入数据的范围为 -2 到 10）。相比之下，正则化模型（使用$s^* = 5$）在海岸线处平滑过渡，在陆地上趋于近似恒定的值。正则化采用了一阶和二阶导数约束，具体说明见第4节末尾。

接下来，我们考察一个表示核反应堆某个部件所产生功率的三维数据集（图 10）。数据位于一个六棱柱内，但 B 样条模型定义在该棱柱的包围盒上。因此，盒子的角落处没有任何数据。若不进行正则化，最小二乘最小化不会收敛，因此图 10 仅展示我们的正则化模型（其中 $s^* = 10$）。自适应正则化方法（图 10，中间和右侧）能够准确地拟合内部的六个“针状结构”，并且在角落区域也有明确的定义。角落处可观察到一些伪影，但其影响不足以波及模型内部。未进行正则化时，系统的条件数为无穷大；采用自适应正则化后，条件数为 $2.18 \times 10^5$。该模型通过同时约束一阶和二阶导数生成。

## 6. 未来工作

在构建我们的方法时，我们对输入数据密度较低或为零的区域内 B 样条的一阶和二阶导数施加了约束。不过，这些附加约束也可以基于不同阶数的导数，或者完全与导数无关。例如，另一种约束是惩罚偏离已知基准值的程度。另一种做法是，对超出给定范围（例如输入数据原始边界）的 B 样条值施加惩罚。我们指出，张量积B样条的自适应正则化过程与所施加的人工约束类型彼此独立。根据具体应用，不同的约束可能更为适用，而我们的方法允许改用这些约束。

未来研究的另一个方向是根据输入数据的局部性质进一步调整正则化强度。目前，整个定义域都采用相同的正则化阈值。这种定义使该方法非常易于使用，但也可以采用更灵活或更复杂的方案。例如，可以分别为一阶导数正则化和二阶导数正则化设置不同的正则化阈值，从而有望更好地平衡这两类约束。此外，本文所述方法会将正则化强度从 0 调整至式（13）中定义的值 $s^* \sqcap s_i$。对于预期需要更多用户交互的应用，可以调整该方法，使其在定义域的不同子区域内实现定制的局部平滑最小值或最大值水平。当已知数据（或数据的子区域）存在噪声，或对平滑处理高度敏感时，这种灵活性尤其有用。

## 7. 结论

由于拟合问题存在病态性，使用张量积B样条对非结构化数据集建模可能十分困难。一般而言，点密度变化较大的数据集或存在无数据区域的数据集会加剧这一问题，以至于必须进行人工平滑。然而，对整个模型进行平滑可能会抹去数据中的尖锐特征。

我们提出了一种适用于B样条模型的正则化过程，它通过在定义域内调整正则化强度来保留特征。我们的方法会根据输入点密度自动改变平滑强度，并且只依赖一个由用户指定的参数，我们称之为正则化阈值。我们发现，自适应正则化的效果优于常见的均匀正则化方案；后者可能导致某些区域过度平滑，而另一些区域平滑不足。我们还证明，我们的方法能够对点密度极其稀疏的数据集拟合B样条模型，并且在没有数据点的区域也能保持定义良好。总的来说，对B样条模型进行自适应正则化，能够为原本难以拟合的数据集生成平滑且准确的模型。

## 参考文献

[1] D. Lenz, R. Yeh, V. Mahadeven, I. Grindeanu, T. Peterka, 《科学数据B样条模型的自适应正则化》，载于：D. Groen、C. de Mu- latier、M. Paszynski、V. V. Krzhizhanovskaya、J. J. Dongarra、P. M. A. Sloot（编），《计算科学——ICCS 2022》，Springer International Publishing，Cham，2022年，第150–163页。doi:https://doi.org/10.1007/ 978-3-031-08751-6 11。

[2] H. Lin、T. Maekawa、C. Deng，《几何迭代方法及其应用综述》，《计算机辅助设计》95（2018）40–51。doi:https: //doi.org/10.1016/j.cad.2017.10.002。[3] T. Hughes、J. Cottrell、Y. Bazilevs，《等几何分析：CAD、有限元、NURBS、精确几何与网格细化》，《计算机辅助设计与应用力学工程方法》194（2005）4135–4195。doi:https://doi.org/10.1016/j.cma.2004.10.008。[4] T. Peterka、Y. Nashed、I. Grindeanu、V. Mahadevan、R. Yeh、X. Trixoche，《科学数据多元函数逼近基础》，载于：2018年IEEE大数据分析与可视化研讨会论文集，2018年，第61–71页。doi:https://doi.org/10.1109/LDAV.2018.8739195。[5] Y. S. Nashed、T. Peterka、V. Mahadevan、I. Grindeanu，《科学数据的有理逼近》，载于：J. M. F. Rodrigues、P. J. S. Cardoso、J. Mon- teiro、R. Lam、V. V. Krzhizhanovskaya、M. H. Lees、J. J. Dongarra、P. M. Sloot（编），《计算科学——ICCS 2019》，Springer，2019年，第18–31页。doi:https://doi.org/10.1007/978-3-030-22734-0 2。[6] C. De Boor，《样条实用指南》，第27卷，应用数学科学，修订版，Springer-Verlag New York，2001年。[7] M. Arigovindan、M. Suhling、C. Jansen、P. Hunziker、M. Unser，《基于回波多普勒数据的完整运动与流场恢复》，《IEEE医学成像汇刊》26（2006）31–45。doi:https://doi.org/10.1109/TMI. 2006.884201。[8] A. Duijndam、M. Schonewille、C. Hindriks，《沿一个空间方向非规则采样的带限信号重建》，《地球物理学》64（1999）524–538。doi:https://doi.org/10.1190/1.1444559。[9] R. Vio、T. Strohmer、W. Wamsteker，《非规则采样时间序列的重建》，《太平洋天文学会出版物》112（2000）74。doi:https://doi.org/10.1086/316495。[10] B. Francis、S. Viswanath、M. Arigovindan，《通过规则网格加权平滑进行散乱数据逼近》，S¯adhan¯a 43（2018）1–16。doi:https: //doi.org/10.1007/s12046-017-0765-y。[11] J. Duchon，《Sobolev空间中使旋转不变半范数最小化的样条》，载于：W. Schempp、K. Zeller（编），《多元函数构造理论》，Springer Berlin Heidelberg，1977年，第85–100页。doi:https://doi.org/10.1007/BFb0086566。[12] C. Gu，《平滑样条方差分析模型》，第297卷，Springer，2013年。doi:https://doi.org/10.1007/978-1-4757-3683-0。[13] P. Craven、G. Wahba，《用样条函数平滑含噪数据》，《数值数学》31（1978）377–403。doi:https://doi.org/10.1007/ BF01404567。[14] C. Gu，《非高斯数据的交叉验证》，《计算与图形统计学杂志》1（1992）169–179。doi:https://doi.org/10.1080/ 10618600.1992.10477012。[15] S. N. Wood，《薄板回归样条》，《英国皇家统计学会杂志：B辑（统计方法论）》65（2003）95–114。doi:https: //doi.org/10.1111/1467-9868.00374。[16] S. Lee、G. Wolberg、S. Shin，《基于多层B样条的散乱数据插值》，《IEEE可视化与计算机图形学汇刊》3（1997）228–244。doi:https://doi.org/10.1109/2945.620490。[17] R. Yeh，《用于精确基于B样条的数据逼近的高效节点优化》，博士论文，普渡大学研究生院，2020年。[18] D. El-Rushaidat、R. Yeh、X. M. Tricoche，《在直线网格上对非结构化数据集进行精确并行重建》，《可视化杂志》（2021）1–20。doi:https://doi.org/10.1007/s12650-020-00740-0。[19] L. Piegl、W. Tiller，《NURBS书》，第2版，Springer-Verlag，1997年。doi:https://doi.org/10.1007/978-3-642-59223-2。[20] S. Ku、R. Hager、C.-S. Chang、J. Kwon、S. E. Parker，《用于托卡马克边缘等离子体回旋动力学模拟的新型混合拉格朗日数值格式》，《计算物理学杂志》315（2016）467–475。doi:https: //doi.org/10.1016/j.jcp.2016.03.062。[21] V. Eyring、S. Bony、G. A. Meehl、C. A. Senior、B. Stevens、R. J. Stouffer、K. E. Taylor，《第六阶段耦合模式比较计划（CMIP6）实验设计与组织概述》，《地球科学模型开发》9（2016）1937–1958。doi:https://doi.org/10.5194/ gmd-9-1937-2016。[22] Y. Yu、E. She mon、V. S. Mahadevan、R. O. Rahaman，《SHARP多物理场教程》，技术报告ANL/NE-16/1，阿贡国家实验室（ANL），美国伊利诺伊州莱蒙特，2016年。doi:https://doi.org/10. 2172/1250465。