arXiv:2304.03222v1 [math.NA] 6 Apr 2023

一种用于近似海森矩阵的矩阵代数方法

Warren Hare* Gabriel Jarry–Bolduc† Chayne Planiden‡

2023年4月7日

摘要

这项工作提出了一种新颖的基于矩阵的方法，仅使用函数求值来构造近似海森矩阵。该方法比基于插值的方法需要更少的计算资源，并且易于在 MATLAB等基于矩阵的编程语言中实现。由于只需要函数求值，该方法适用于无导数算法。

对于结构合理的样本集，已证明该方法能够生成一阶精确的完整海森矩阵近似。在更特殊的结构下，已证明该方法能够达到二阶精度。本文还研究了未定情形，即采样点数量不足以进行完整插值的情况，并为由此得到的部分海森矩阵推导了误差界。

关键词：海森矩阵近似；无导数优化；N 阶精度；广义单纯形梯度；广义单纯形海森矩阵；广义中心单纯形 Hessian。
AMS 学科分类：主要分类 65K10、90C56。

1 引言

无导数优化（DFO）研究的是在不使用梯度或高阶导数信息的情况下，如何求得函数的最小值和极小点 [3]。近年来，无导数优化日益普及，并适用于梯度不可用、计算代价高昂或难以获取的情形 [3, 17]。无导数优化方法已用于光滑[9, 14, 22, 34, 39, 40, 43, 45] 和非光滑 [4, 7, 29, 31, 33, 36] 优化，最常见的形式是直接搜索 [1, 2, 6, 8, 23] 或基于模型 [30, 34, 35, 44, 45] 的方法，也有一些方法结合了两者 [5, 20]。基于模型的无导数优化方法采用数值分析技术来

*加拿大不列颠哥伦比亚大学奥肯那根校区数学系，基洛纳，不列颠哥伦比亚省 V1V 1V7。部分研究经费由加拿大 NSERC Discovery Grant 2018-03865 资助。warren.hare@ubc.ca，ORCID 0000-0002-4240-3903

†加拿大不列颠哥伦比亚大学奥肯那根校区数学系，基洛纳，不列颠哥伦比亚省 V1V 1V7。部分研究经费由加拿大自然科学与工程研究委员会（NSERC）Discovery Grant 2018-03865 资助。gabjarry@alumni.ubc.ca

‡澳大利亚伍伦贡大学数学与应用统计学院，伍伦贡，新南威尔士州 2500。研究经费由伍伦贡大学资助。chayne@uow.edu.au，ORCID 0000-0002-0412-8445

以可控的误差界近似梯度和海森矩阵。例如，单纯形梯度是一种近似梯度的技术。该方法选取一组在 $\mathbb{R}^n$ 中间隔适当的$n+1$采样点，称为单纯形，并利用这些点构造一个线性插值函数，在局部近似目标函数，然后计算仿射函数的梯度。对于梯度局部利普希茨连续的函数，研究表明单纯形梯度的误差界良好 [10]。

近期一项研究方向探讨了改进或推广单纯形梯度的方法，以便将其用于无导数优化 [15, 16, 18, 21, 25, 28, 38, 41]。基于 [16], 中的思路， [25] 将单纯形梯度推广，使其不再要求恰好使用 $n+1$个点；现在可以使用任意有限个间距适当的点，得到误差受控的近似。在 [28], 中，研究者采用类似方法考察了广义中心单纯形梯度。这种近似使用的点数是原来的两倍，但将误差控制阶数从 1 阶提升到了2 [24]阶。

类似地，研究者还探索了近似完整海森矩阵或部分海森矩阵的方法。在 [21], 中，作者提出了一种“单纯形海森矩阵”的构造思路：通过二次插值，利用（$n+1$）（$n+2$）/2 个适定采样点构造该矩阵。他们还指出，如果只需要海森矩阵的一部分（例如对角分量），则可以使用更少的点。这些思路在 [16] 中通过二次插值加以形式化，并借助拉格朗日插值多项式进行分析。 [19, 32]也讨论了如何近似海森矩阵的对角分量。

在这项工作中，我们通过引入广义单纯形海森矩阵（GSH）和广义中心单纯形Hessian（GCSH），继续发展这些工具。GSH与二次插值函数的海森矩阵密切相关。我们将看到，在某些情况下，这两种方法会得到相同的结果。

本文可视为对 [17, 21] 中介绍的与单纯形海森矩阵相关工作的扩展。我们基于矩阵代数概念，引入了一个用于近似函数 f的海森矩阵的显式公式。该公式具有多项优点。首先，无论所使用的采样点数量是多少，该公式都提供了一种定义良好的方法。事实上，只要用于构建采样点集的矩阵非空，GSH和 GCSH就是定义良好的。特别是，即使二次插值函数不存在或不唯一，GSH 也提供了一个用于近似海森矩阵的显式公式。然而，我们注意到，当GSH 计算中所使用的矩阵具有特定结构时，GSH 等价于海森矩阵的前向有限差分近似。

GSH 能够在合理假设下精确近似目标函数 f 的海森矩阵： $\mathbb{R}^n \to \mathbb{R}$。该技术使用一个关注点 x^0 和方向集，这些方向集存储在一组矩阵 $\{S, T_1, \ldots, T_m\}$ 中，矩阵的列向量与关注点相加，以构建采样点。将 Δ_S 定义为 S 的半径，将 Δ_T 定义为 T_i 的最大半径，我们证明，若 S 和所有 T_i 均为满行秩，且海森矩阵 $\nabla^2 f$ 存在并且利普希茨连续，则 GSH 对 $\nabla^2 f$ 的估计误差不超过 $\Delta_u := \max\{\Delta_S, \Delta_T\}$ 的倍数。就阶数为 N 的精度 [24], 而言，GSH 可提供一阶精确的完整海森矩阵近似（见本文定义 2.11）。此外，GCSH 可提供二阶精确的完整海森矩阵近似。文中还给出了欠定情形（见定义 2.9）的误差界，表明 GSH 及其中心化版本 GCSH 分别是适当部分海森矩阵的一阶和二阶精确近似。

本文给出的误差界与 [15, 16] 中的误差界有相似之处。在

这些论文中，作者给出了完全二次模型的海森矩阵的误差界 [15,定理 3], 二次回归模型的海森矩阵 [16, 定理 3.2], 以及二次未定模型的海森矩阵 [16,定理 5.12]。

GSH公式的第二个优点是，通过在定义中采用矩阵结构，海森矩阵近似可以在任何基于矩阵的编程语言中极其容易地实现。此外，在完成函数求值后，与二次插值相比，该矩阵结构所需的计算工作量（以浮点运算次数计）更少。实际上，当 S and所有 T_i 都是方阵时，计算GSH需要数量级为 $O(n^4)$的浮点运算次数。相比之下，使用一组适用于二次插值的采样点来求二次插值函数的海森矩阵，则需要数量级为 $O(n^6)$的浮点运算次数 [15]（另见 [17, §6.2]）。

本文的其余部分组织如下。第2节介绍符号说明和一些必要的定义，包括广义单纯形梯度、GSH 和 GCSH 的定义。第3节阐明 GSH 与 GCSH 之间的关系。第4节给出 GSH 和 GCSH 的若干误差界。误差界分为两类：所有方向矩阵 T_i 均相同的情形，以及所有矩阵 T_i 不一定相同的情形。对于这两种情形，误差界对欠定情形均有效。在第5节中，我们建立了用于计算 GSH 的极小确定集，并证明其对于二次插值是良态的。最后，第6节给出总结性评论，并建议未来在这一方向上的研究领域。

2 预备知识

在本文中，我们采用 [42] 中的标准记号。函数 f 的定义域记作 $\operatorname{dom} f$。矩阵 A 的转置记作 $A^\top$。我们在有限维空间 $\mathbb{R}^n$ 中工作，该空间配有内积 $x^\top y = \sum_{i=1}^n x_i y_i$ 及其诱导范数 $\|x\| = \sqrt{x^\top x}$。$\mathbb{R}^{n\times n}$ 中的单位矩阵记作 Id_n。我们用 $e_n^i \in \mathbb{R}^n$ 表示 $i \in \{1,\dots,n\}$，以表示 $\mathbb{R}^n$ 中的标准单位基向量，即 $\mathrm{Id}_n \cdot$ 的第 i^{th} 列。矩阵 $\operatorname{Diag}(v) = \operatorname{Diag}(v_1,\dots,v_n) =: M \in \mathbb{R}^{n\times n}$，其中 $v \in \mathbb{R}^n$，表示对角线元素为 $M_{i,i} = v_i$（对所有 $i \in \{1,\dots,n\}$.）的对角矩阵。$\mathbb{R}^n$ 中的零向量记作 0，$\mathbb{R}^{n\times m}$ 中的零矩阵记作 $0_{n\times m}$。矩阵 A 的第 i^{th} 行第 j^{th} 列元素记作 $A_{i,j}$.

给定矩阵 $A \in \mathbb{R}^{n\times m}$，我们使用诱导矩阵范数

$$\|A\| = \|A\|_2 = \max\{\|Ax\|_2 : \|x\|_2 = 1\}$$

和弗罗贝尼乌斯范数

$$\|A\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}^2\right)^{\frac{1}{2}}.$$

我们分别用 $B(x^0,\Delta)$ 和 $\overline{B}(x^0,\Delta)$ 表示以 x^0 为中心、半径为 Δ 的开球和闭球。两个向量集 A 和 B 的闵可夫斯基和记作 $A \oplus B.$。也就是说

$$A \oplus B = \{a + b : a \in A, b \in B\}.$$

我们将二次函数 $Q\colon \mathbb{R}^n \to \mathbb{R}$ 定义为具有如下形式的函数： $Q(x) = \alpha_0 + \alpha^\top x + \frac{1}{2}x^\top \mathbf{H} x$，其中 $\alpha_0 \in \mathbb{R}$、$\alpha \in \mathbb{R}^n$，且 $\mathbf{H} = \mathbf{H}^\top \in \mathbb{R}^{n\times n}$.。仿射函数 $\mathcal{L}\colon \mathbb{R}^n \to \mathbb{R}$ 定义为任何可写成 $\mathcal{L}(x) = \alpha_0 + \alpha^\top x$ 形式的函数。注意，仿射函数和常函数 $C(x) = \alpha_0$ 也视为二次函数，其中 $\mathbf{H} = \mathbf{0}_{n\times n}$。

接下来，我们将介绍本文中将要使用的基本定义和符号。

定义2.1（二次插值的适定性）。 [15, 16, 17] 这组互异点 $\mathcal{Y} = \{y^0, y^1, \ldots, y^m\} \subset \mathbb{R}^n$ 满足 $m = \frac{1}{2}(n+1)(n+2) - 1$，当且仅当该方程组

$$\alpha_0 + \alpha^\top y^i + \frac{1}{2}(y^i)^\top \mathbf{H} y^i = 0, \quad i \in \{0, 1, \ldots, m\}, \tag{1}$$

对于 $\alpha_0 \in \mathbb{R}$、$\alpha \in \mathbb{R}^n$，$\mathbf{H} = \mathbf{H}^\top \in \mathbb{R}^{n\times n}$. 有唯一解

当一组采样点 $\mathcal{Y}$ 适用于二次插值时，这意味着存在一个经过 $\mathcal{Y}$. 中所有采样点的唯一二次模型。因此，通过 $\mathcal{Y}$ 定义的二次模型的海森矩阵是唯一的。请注意，包含$(n+1)(n+2)/2$ 个不同采样点的一组采样点不一定适用于二次插值。当且仅当与线性系统 (1) 相关联的矩阵是方阵且满秩时，这组采样点才适用于二次插值。例如，这组采样点

$$\left\{ \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}-1\\0\end{bmatrix}, \begin{bmatrix}0\\-1\end{bmatrix}, \begin{bmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix}, \begin{bmatrix}-\frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}}\end{bmatrix} \right\} \subset \mathbb{R}^2$$

不适用于二次插值，因为与系统 (1) 相关联的矩阵的秩为 $5 < 6$。集合

$$\left\{ \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}-1\\0\end{bmatrix}, \begin{bmatrix}0\\-1\end{bmatrix}, \begin{bmatrix}\frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}}\end{bmatrix} \right\} \subset \mathbb{R}^2$$

适用于二次插值，因为与系统 (1) 相关联的矩阵的秩为 6。当 $\mathcal{Y}$中的不同采样点数量少于 $(n+1)(n+2)/2$时，二次插值模型不再唯一。这种情形称为欠定情形 [17,Chapter 5]。在这种情形下，可以通过最小弗罗贝尼乌斯范数问题 [17, Chapter 5] 来定义二次模型（另见 [16, Section 5]）。当$\mathcal{Y}$ 中的不同采样点数量多于 $(n+1)(n+2)/2$，且不存在经过 $\mathcal{Y}$中所有点的二次函数时，可以通过最小二乘回归问题 [17,Chapter 4] 来确定二次模型（另见 [16,Sections 2,3 & 4]）。

定义 2.2（二次插值函数）。 [3, 定义 9。9] 设 $f\colon \operatorname{dom} f \subseteq \mathbb{R}^n \to \mathbb{R}$，且 $\mathcal{Y} = \{y^0, y^1, \ldots, y^m\} \subset \operatorname{dom} f$ 中的 $m = \frac{1}{2}(n+1)(n+2) - 1$ 对二次插值是适定的。则 f 在 $\mathcal{Y}$ 上的二次插值函数为

$$Q_f(\mathcal{Y})(x) = \alpha_0 + \alpha^\top x + \frac{1}{2}x^\top \mathbf{H} x,$$

其中，$(\alpha_0, \alpha, \mathbf{H} = \mathbf{H}^\top)$ 是以下方程的唯一解

$$\alpha_0 + \alpha^\top y^i + \frac{1}{2}(y^i)^\top \mathbf{H} y^i = f(y^i), \quad i \in \{0, 1, \ldots, m\}.$$

在下一个定义中，我们引入用于构造GSH的关键符号。此处，我们以矩阵形式表示一个向量集合，意指矩阵的每一列都是该集合中的一个向量。

定义 2.3（GSH 记号）。令 $f\colon \operatorname{dom} f \subseteq \mathbb{R}^n \to \mathbb{R}$，并令 $x^0 \in \operatorname{dom} f$ 为关注点。令

$$S = \begin{bmatrix} s^1 & s^2 & \cdots & s^m \end{bmatrix} \in \mathbb{R}^{n\times m} \quad and$$
$$T_i = \begin{bmatrix} t_i^1 & t_i^2 & \cdots & t_i^k \end{bmatrix} \in \mathbb{R}^{n\times k_i}, i \in \{1,\ldots,m\}$$

为包含于 $\mathbb{R}^n$ 中的方向集，并以矩阵形式表示。定义

$$T_{1:m} = \{T_1,\ldots,T_m\}.$$

假设 $x^0 \oplus T_i$，　　$x^0 + s^i$、$x^0 + s^i \oplus T_i$ 对于所有 $i \in \{1,\ldots,m\}$ 均包含于 $\operatorname{dom} f$ 中。定义

$$\Delta_S = \max_{i\in\{1,\ldots,m\}} \|s^i\|, \quad \Delta_{T_i} = \max_{j\in\{1,\ldots,k_i\}} \|t_i^j\|, \quad \Delta_T = \max_{i\in\{1,\ldots,m\}} \Delta_{T_i},$$

$$\widehat{S} = \frac{1}{\Delta_S} S, \tag{2}$$

and

$$\delta_s\left(x^0; T_i\right) = \begin{bmatrix} f(x^0 + t_i^1) - f(x^0) \\ f(x^0 + t_i^2) - f(x^0) \\ \vdots \\ f(x^0 + t_i^k) - f(x^0) \end{bmatrix} \in \mathbb{R}^{k_i}.$$

本文假设所有涉及的矩阵均具有非零秩。这确保矩阵半径非零，因此式（2）始终有定义。注意，在前述定义中，对于所有 $i \in \{1,\ldots,m\}$，$m \in \mathbb{N}$ 和 k_i 均可为任意正整数。

对于非方阵，矩阵逆的一种推广是伪逆。最广为人知的矩阵伪逆类型是穆尔－彭罗斯伪逆。

定义 2.4（穆尔－彭罗斯伪逆）。设 $A \in \mathbb{R}^{n\times m}$。$A$ 的穆尔－彭罗斯伪逆记为 $A^\dagger$，它是 $\mathbb{R}^{m\times n}$ 中满足以下四个方程的唯一矩阵：

(i) $AA^\dagger A = A$,
(ii) $A^\dagger AA^\dagger = A^\dagger$,
(iii) $(AA^\dagger)^\top = AA^\dagger$, (iv) $(A^\dagger A)^\top = A^\dagger A$.

注意，给定 $A \in \mathbb{R}^{n\times m}$，存在唯一的穆尔－彭罗斯伪逆 $A^\dagger \in \mathbb{R}^{m\times n}$。以下两个性质成立。

(i) 如果 A 具有列满秩 m，那么 $A^\dagger$ 是 A的左逆，即 $A^\dagger A = \mathrm{Id}_m$. 在这种情况下，

$$A^\dagger = (A^\top A)^{-1} A^\top. \tag{3}$$

(ii) 如果 A 具有满行秩 n，那么 $A^\dagger$ 是 A的右逆，即 $AA^\dagger = \mathrm{Id}_n$. 在这种情况下，$A^\dagger = A^\top (AA^\top)^{-1}$.

在介绍 GSH 之前，回顾广义单纯形梯度的定义是很有价值的。

定义 2.5（广义单纯形梯度）。 [25, 定义 2] 设 $f\colon \mathrm{dom}\, f \subseteq \mathbb{R}^n \to \mathbb{R}$，且令$x^0 \in \mathrm{dom}\ f$为关注点。令 $S \in \mathbb{R}^{n\times m}$满足 $x^0 \oplus S \subset \mathrm{dom}\, f.$。 f在 x^0处关于 S的广义单纯形梯度记为 $\nabla_s f(x^0;\ S)$，并定义为

$$\nabla_s f(x^0; S) = (S^\top)^\dagger \delta_s(x^0; S).$$

穆尔－彭罗斯伪逆允许 S 包含任意有限数量的列，而不局限于 $m = n$。当 $S \in \mathbb{R}^{n\times n}$ 且秩$S = n$时， $S^\top$ 的穆尔－彭罗斯伪逆就是 $S^\top$ 的逆，我们便得到单纯形梯度的定义[3, 定义 9.5]。注意，广义单纯形梯度是熟知的梯度前向有限差分近似 [12, 第4节1]的推广。事实上，如果 $S = H\,\mathrm{Id}_n$，其中$H = \mathrm{Diag}\begin{bmatrix} h_1 & h_2 & \dots & h_n \end{bmatrix}$， $h_i > 0$，那么我们便得到梯度的前向有限差分近似。

现在我们介绍本文的两个关键定义：广义单纯形海森矩阵和广义中心单纯形Hessian。它需要有限的方向集，以及用于近似海森矩阵的关注点。与广义单纯形梯度类似，GSH涉及穆尔－彭罗斯伪逆和差分矩阵。在这种情况下，差分矩阵由广义单纯形梯度之间的差构成。

为便于记号书写，在适当处我们使用定义 $T_{1:m} = \{T_1, \dots, T_m\}$。

定义 2.6（广义单纯形海森矩阵）。设 $f\colon \mathrm{dom}\ f \subseteq \mathbb{R}^n \to \mathbb{R}$，且 $x^0 \in \mathrm{dom}\, f$ 为关注点。设 $S = \begin{bmatrix} s^1 & s^2 & \cdots & s^m \end{bmatrix} \in \mathbb{R}^{n\times m}$ 和 $T_i \in \mathbb{R}^{n\times k_i}$，其中 $x^0 \oplus T_i$、$x^0 \oplus S$、$x^0 + s^i \oplus T_i$ 对所有 $i \in \{1, \dots, m\}$. 均包含于 $\mathrm{dom}\ f$。 f 在 x^0 处、关于 S 和 $T_{1:m}$ 的广义单纯形海森矩阵记作 $\nabla_s^2 f(x^0;\ S, T_{1:m})$，定义为

$$\nabla_s^2 f(x^0; S, T_{1:m}) = (S^\top)^\dagger \delta_s^2(x^0; S, T_{1:m}), \tag{4}$$

where

$$\delta_s^2(x^0; S; T_{1:m}) = \begin{bmatrix} (\nabla_s f(x^0 + s^1; T_1) - \nabla_s f(x^0; T_1))^\top \\ (\nabla_s f(x^0 + s^2; T_2) - \nabla_s f(x^0; T_2))^\top \\ \vdots \\ (\nabla_s f(x^0 + s^m; T_m) - \nabla_s f(x^0; T_m))^\top \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

在 $T_1 = T_2 = \cdots = T_m$ 情形下，我们使用 $\overline{T} := T_i$ 来简化记号，并写作 $\nabla_s^2 f(x^0;\ S;\ \overline{T})$，以强调这一特殊情形。

我们指出，$\nabla_s^2 f$、δ_s 和 δ_s^2 中的 s 指的是单纯形。现在我们引入广义中心单纯形 Hessian，这一点便显得很重要。δ_s^2 中的“平方”强调了我们是在对一阶对象求差（因此得到一个二阶对象）。

定义 2.7（广义中心单纯形 Hessian）。设 $f : \mathrm{dom}\, f \subseteq \mathbb{R}^n \to \mathbb{R}$，且 $x^0 \in \mathrm{dom}\ f$为关注点。设 $S \in \mathbb{R}^{n\times m}$和 $T_i \in \mathbb{R}^{n\times k_i}$满足 $x^0 \oplus S \oplus T_i, x^0 \oplus (-S) \oplus (-T_i)$ $x^0 \oplus (\pm S)$，并且对所有 $i \in \{1, \ldots, m\}.$， $x^0 \oplus (\pm T_i)$均包含于 $\mathrm{dom}\ f$。f在 x^0处、基于 S和 $T_{1:m}$的广义中心单纯形 Hessian 记为 $\nabla_c^2 f(x^0; S, T_{1:m})$，定义如下：

$$\nabla_c^2 f(x^0; S, T_{1:m}) = \frac{1}{2}\left(\nabla_s^2 f(x^0; S, T_{1:m}) + \nabla_s^2 f(x^0; -S, -T_{1:m})\right). \tag{5}$$

请注意，$\nabla_c^2 f$ 中的 c 表示中心。

GCSH 可视为海森矩阵中心有限差分近似（也称中点近似）的推广 [12, 第4节。1]。类似地，GSH 可视为海森矩阵前向有限差分近似的推广 [12, 第4节。1]。事实上，当 $S = T_1 = \cdots = T_m = H\,\mathrm{Id}_n$，其中 $H = \mathrm{Diag}\begin{bmatrix} h_1 & h_2 & \ldots & h_n \end{bmatrix}, h_i > 0$ 时，GSH（GCSH）等同于二阶导数的前向有限差分近似（中心有限差分近似）。此外，GSH 可视为导数方法背景下由泰勒定理给出的海森矩阵近似的无导数优化对应形式；该近似通过真实梯度之差来近似梯度（见 [37, 第8节。1]）。与前向有限差分近似（中心有限差分近似）相比，GSH（GCSH）的优点是限制较少。特别是，只要矩阵 S和 $T_1, \ldots, T_m$ 非空，其定义就是良好的。

注 2.8. 考虑 S 和 T_i 为方阵的情形，即对所有 $i \in \{1, \ldots, m\}.$， $m = n$ 和 $k_i = n$。GSH 的计算首先构造 $n + 1$ 个广义单纯形梯度。暂不考虑函数求值，计算单纯形梯度的主要开销是计算矩阵逆，需要 $O(n^3)$ 次浮点运算。因此，构造 $\delta_s^2(x^0; S; T_{1:m})$ 需要 $O(n^4)$ 次浮点运算。计算 S 的逆（需要 $O(n^3)$ 次浮点运算），并将其与 $\delta_s^2(x^0; S; T_{1:m})$ 相乘，需要 $O(n^3)$ 次浮点运算。因此，暂不考虑函数求值，计算 GSH 需要 $O(n^4)$ 次浮点运算。

相反，暂不考虑函数求值，计算二次插值需要 $O(n^6)$ 次浮点运算。

定义 2.9. 设 f: $\mathrm{dom}\ f \subseteq \mathbb{R}^n \to \mathbb{R}$，且令 $x^0 \in \mathrm{dom}\, f$为关注点。令 $S = \begin{bmatrix} s^1 & s^2 & \cdots & s^m \end{bmatrix} \in \mathbb{R}^{n\times m}$和 $T_i \in \mathbb{R}^{n\times k_i}$满足 $x^0 \oplus S \oplus T_i$、$x^0 \oplus (-S) \oplus (-T_i)$ $x^0 \oplus (\pm S)$，且$x^0 \oplus (\pm T_i)$ 对所有 $i \in \{1, \ldots, m\}.$均包含于 $\mathrm{dom}\ f$。假设所有矩阵的非零秩均不为零。我们定义以下四种情形，以刻画矩阵 $S \in \mathbb{R}^{n\times m}$ 和矩阵集合$T_{1:m} := \{T_1, \ldots, T_m\}$.

- 欠定：如果 S 为非方阵且列满秩，则称 GSH (GCSH) 是 S-欠定的。如果集合 $T_{1:\,m}$ 中的所有矩阵都列满秩，且至少有一个矩阵为非方阵，则称其为 $T_{1:\,m}$-欠定的。

- 判定：如果 S 是方阵且满秩，则称 GSH (GCSH) 是 S-判定的。它是$T_{1:\,m}$-判定的，当且仅当集合 $T_{1:\,m}$ 中的所有矩阵都是方阵且满秩。

• 超定：如果 S 是非方阵且满行秩，则称 GSH（GCSH）为 S-超定。若集合$T_{1:m}$ 中的所有矩阵均为满行秩，且至少一个为非方阵，则称其为 $T_{1:m}$-超定。

• 不定：如果 GSH (GCSH) 不属于前述三种情形中的任何一种，则称其为 S-不定。若集合 $T_{1:\ m}$ 不属于前述三种情形中的任何一种，则称其为 $T_{1:\ m}$-不定。

注意， S-欠定 GSH (GCSH) 的定义意味着 span $S \neq \mathbb{R}^n$，当且仅当 $SS^\dagger \neq \mathrm{Id}_n$ 时，这一条件成立。类似地，$T_{1:m}$-欠定 GSH (GCSH) 的定义意味着对于某个 $\bar{i} \in \{1, \ldots, m\}$，span $T_{\bar{i}} \neq \mathbb{R}^n$，当且仅当对于 $\bar{i}$， $T_{\bar{i}} T_{\bar{i}}^\dagger \neq \mathrm{Id}_n$ 时，这一条件成立。

如上所述，将欠定、定、超定和不定定义后，可得到用于分类 GSH (GCSH) 的16种不同情形，这些情形均在第4节中进行了研究。

以下命题给出了广义单纯形梯度 $\nabla_s f(x^0; S)$ 与真实梯度 $\nabla f(x^0)$ 之间绝对误差的界。注意，假设 $f \in \mathcal{C}^2$ on$B(x^0, \overline{\Delta})$ 意味着 ∇f 在 $\overline{B}(x^0, \Delta)$ 上是利普希茨连续的，其中 $\Delta < \overline{\Delta}$.

我们指出，类似于命题2.10的结果已经使用各种不同的技术 [21, 41, 25]被多次证明。在 [15, 16]（另见 [17] 和 [21, 定理3.1]）中，使用拉格朗日插值多项式对单纯形梯度进行了分析。在 [41]，（另见 [3, 25]）中，使用矩阵条件数对单纯形梯度的误差界进行了分析。在命题2.10中，我们以这种第二种形式给出了误差界，因为进一步的结果将以相同的方式研究广义单纯形海森矩阵。

命题2.10（广义单纯形梯度的误差界）。设 $S = \begin{bmatrix} s^1 & s^2 & \cdots & s^m \end{bmatrix} \in \mathbb{R}^{n \times m}$ 的半径为 $\Delta_S > 0$。设 $\widehat{S} = S/\Delta_S$，并令 f: dom $f \subseteq \mathbb{R}^n \to \mathbb{R}$ 在 $B(x^0, \overline{\Delta})$ 上为 $\mathcal{C}^2$，其中 x^0 为关注点，且 $\overline{\Delta} > \Delta_S$. 。记 $L_{\nabla f} \geq 0$ 为 ∇f 在 $\overline{B}(x^0, \Delta_S)$ 上的利普希茨常数。则

$$\begin{aligned} \| \operatorname{Proj}_S \nabla_s f(x^0; S) - \operatorname{Proj}_S \nabla f(x^0) \| &= \| \nabla_s f(x^0; S) - \operatorname{Proj}_S \nabla f(x^0) \| \\ &\leq \frac{\sqrt{m}}{2} L_{\nabla f} \left\| (\widehat{S}^\top)^\dagger \right\| \Delta_S, \end{aligned}$$

其中 Proj$_S(\cdot) = (S^\top)^\dagger S^\top(\cdot)$。

命题2.10中给出的误差界仅要求 S 具有非零秩。因此，它涵盖了所有可能的情况： S-欠定、 S-定、 S-超定以及S-不定。在第4节中，我们将为 GSH 和 GCSH 推导与命题2.10中形式相似的误差界。为此，接下来定义对子空间对的投影。

给定矩阵 $S \in \mathbb{R}^{n \times m}$ 和 $T_i \in \mathbb{R}^{n \times k_i}$，将矩阵 $H \in \mathbb{R}^{n \times n}$ 投影到 S 和 $T_{1:m}$ 上所得的投影记为 Proj$_{S,\ T_{1:m}} H$ ，定义为

$$\operatorname{Proj}_{S, T_{1:m}} H = \sum_{i=1}^{m} (S^\top)^\dagger e_m^i (e_m^i)^\top S^\top H T_i T_i^\dagger.$$

在 $T_1 = T_2 = \cdots = T_m =: \overline{T}$ 的情况下，将 H 投影到 S 和 $\overline{T}$ 上所得的投影记为 $\mathrm{Proj}_{S,\overline{T}} H$，并可化简为

$$\begin{aligned}\mathrm{Proj}_{S,\overline{T}} H &= \sum_{i=1}^{m} (S^\top)^\dagger e_m^i (e_m^i)^\top S^\top H \overline{TT}^\dagger \\ &= (S^\top)^\dagger \left(\sum_{i=1}^{m} e_m^i (e_m^i)^\top \right) S^\top H \overline{TT}^\dagger \\ &= (S^\top)^\dagger \mathrm{Id}_m S^\top H \overline{TT}^\dagger \\ &= (S^\top)^\dagger S^\top H \overline{TT}^\dagger.\end{aligned}$$

我们通过定义本文多次使用的 N 阶精度来结束本节，以描述所提出的海森矩阵近似技术的质量。

定义 2.11（N 阶精度）。设 $P \in \mathbb{R}^{n\times n}$。给定 $f \in \mathcal{C}^2, x^0 \in \mathrm{dom}\, f$ 和 $\overline{\Delta} > 0$，我们称 $\{\tilde{f}_\Delta\}_{\Delta\in(0,\overline{\Delta}]}$ 为一类在 x^0 处对 f 进行建模、由 Δ 参数化的模型，若其在 x^0 处对 P 提供 N 阶精度˜，则存在一个标量 $\kappa(x^0)$，使得对于任意 $\Delta \in (0,\overline{\Delta}]$，模型 f_Δ 满足

$$\|P - \nabla^2 \tilde{f}_\Delta(x^0)\| \le \kappa(x^0)\Delta^N.$$

在第4节定义误差界时，前一定义中的矩阵 P 将等于 $\mathrm{Proj}_{S,\ T_{1:m}} \nabla^2 f(x^0)$。我们也将 $\mathrm{Proj}_{S,\ T_{1:m}} \nabla^2 f(x^0)$称为部分海森矩阵。注意，当 S为满行秩且对于所有 i， T_i 均为满行秩时，$\mathrm{Proj}_{S,\ T_{1:m}} \nabla^2 f(x^0)$等于真实的海森矩阵$\nabla^2 f(x^0)$。我们将看到，在适当的假设下，可以对部分海森矩阵 $\mathrm{Proj}_{S,\ T_{1:m}} \nabla_s^2 f(x^0;\ S, T_{1:m})$获得 1 阶精度或 2 阶精度。首先，下一节将阐明 GSH 与 GCSH 之间的关系。

3 GSH 与 GCSH 之间的关系

本节中，我们研究 GSH 与 GCSH 之间的关系。首先，我们给出一个计算 GCSH 的公式，其形式与 GSH 的公式 (4) 类似。

在引理 3.1 中，我们引入记号 δ_c^2。这是 δ_s^2 的中心化版本；因此，单纯形的 s 被替换为中心化的 c。

引理 3.1。设 $f\colon \mathrm{dom}\ f \subseteq \mathbb{R}^n \to \mathbb{R}$，且 $x^0 \in \mathrm{dom}\, f$ 为关注点。设 $S = \begin{bmatrix} s^1 & \cdots & s^m \end{bmatrix} \in \mathbb{R}^{n\times m}$，并设 $T_i \in \mathbb{R}^{n\times k_i}$ 满足 $x^0 \oplus S \oplus T_i$、$x^0 \oplus (-S) \oplus (-T_i)$ $x^0 \oplus (\pm S)$ $x^0 \oplus (\pm T_i)$ 对所有 $i \in \{1,\dots,m\}$. 均包含于 $\mathrm{dom}\ f$ 中。则

$$\nabla_c^2 f(x^0; S, T_{1:m}) = (S^\top)^\dagger \delta_c^2(x^0; S, T_{1:m})$$

其中

$$\delta_c^2(x^0; S, T_{1:m}) = \frac{1}{2} \begin{bmatrix} \left(\delta_s(x^0 + s^1; T_1) + \delta_s(x^0 - s^1; -T_1) - \delta_s(x^0; T_1) - \delta_s(x^0; -T_1)\right)^\top T_1^\dagger \\ \vdots \\ \left(\delta_s(x^0 + s^m; T_m) + \delta_s(x^0 - s^m; -T_m) - \delta_s(x^0; T_m) - \delta_s(x^0; -T_m)\right)^\top T_m^\dagger \end{bmatrix}.$$

证明。我们有

$$
\begin{aligned}
&\nabla_c^2 f(x^0; S, T_{1:m}) \\
=&\frac{1}{2}\left(\nabla_s^2 f(x^0; S, T_{1:m}) + \nabla_s^2 f(x^0; -S, -T_{1:m})\right) \\
=&\frac{1}{2}(S^\top)^\dagger \left(\delta_s^2(x^0; S, T_{1;m}) - \delta_s^2(x^0; -S, -T_{1:m})\right) \\
=&(S^\top)^\dagger \frac{1}{2}\begin{bmatrix} \left(\nabla_s f(x^0+s^1; T_1) - \nabla_s f(x^0; T_1)\right)^\top - \left(\nabla_s f(x^0-s^1; -T_1) - \nabla_s f(x^0; -T_1)\right)^\top \\ \vdots \\ \left(\nabla_s f(x^0+s^m; T_m) - \nabla_s f(x^0; T_m)\right)^\top - \left(\nabla_s f(x^0-s^m; -T_m) - \nabla_s f(x^0; -T_m)\right)^\top \end{bmatrix} \\
=&(S^\top)^\dagger \frac{1}{2}\begin{bmatrix} \left(\delta_s(x^0+s^1; T_1) - \delta_s(x^0; T_1)\right)^\top T_1^\dagger - \left(\left(\delta_s(x^0-s^1; -T_1) - \delta_s(x^0; -T_1)\right)^\top (-T_1)^\dagger\right) \\ \vdots \\ \left(\delta_s(x^0+s^m; T_m) - \delta_s(x^0; T_m)\right)^\top T_m^\dagger - \left(\left(\delta_s(x^0-s^m; -T_m) - \delta_s(x^0; -T_m)\right)^\top (-T_m)^\dagger\right) \end{bmatrix} \\
=&(S^\top)^\dagger \frac{1}{2}\begin{bmatrix} \left(\delta_s(x^0+s^1; T_1) + \delta_s(x^0-s^1; -T_1) - \delta_s(x^0; T_1) - \delta_s(x^0; -T_1)\right)^\top T_1^\dagger \\ \vdots \\ \left(\delta_s(x^0+s^m; T_m) + \delta_s(x^0-s^m; -T_m) - \delta_s(x^0; T_m) - \delta_s(x^0; -T_m)\right)^\top T_m^\dagger \end{bmatrix} \\
=&(S^\top)^\dagger \delta_c^2(x^0; S, T_{1:m}).
\end{aligned}
$$

□

注意

$$
\begin{aligned}
&\left[\delta_s(x^0+s^i; T_i) + \delta_s(x^0-s^i; -T_i) - \delta_s(x^0; T_i) - \delta_s(x^0; -T_i)\right]_{j_i} \\
=&\frac{1}{2}\big(f(x^0+s^i+t_i^{j_i}) - f(x^0+s^i) + f(x^0-s^i-t_i^{j_i}) - f(x^0-s^i) \\
&- f(x^0+t_i^{j_i}) + f(x^0) - f(x^0-t_i^{j_i}) + f(x^0)\big) \\
=&\frac{1}{2}\big(f(x^0+s^i+t_i^{j_i}) + f(x^0-s^i-t_i^{j_i}) \\
&- f(x^0+s^i) - f(x^0-s^i) - f(x^0+t_i^{j_i}) - f(x^0-t_i^{j_i}) + 2f(x^0)\big)
\end{aligned} \tag{6}
$$

对于 $j_i \in \{1, \ldots, k_i\}$ 和 $i \in \{1, \ldots, m\}$.，我们将使用定理 4.3 中的 (6)。

接下来，我们证明 GCSH 实际上是 GSH 的一种特殊情形，其中方向矩阵具有特定形式。这个结果类似于 [28, 命题 2.10], 中的结果，该结果表明，当用于计算广义单纯形梯度的方向矩阵具有 $A = \begin{bmatrix} S & -S \end{bmatrix}$ 的形式时，广义中心梯度等于广义单纯形梯度，其中 $S \in \mathbb{R}^{n\times m}$. 是某个矩阵。我们将使用以下引理。

引理 3.2。令 $A = \begin{bmatrix} S & -S \end{bmatrix} \in \mathbb{R}^{n\times 2m}$ 对于某个矩阵 $S \in \mathbb{R}^{n\times m}$.，则

$$
A^\dagger = \frac{1}{2}\begin{bmatrix} S^\dagger \\ -S^\dagger \end{bmatrix}.
$$

证明。这可由基本代数得到。 □

命题 3.3。令 $f\colon \operatorname{dom} f \subseteq \mathbb{R}^n \to \mathbb{R}$，并令 $x^0 \in \operatorname{dom} f$ 为关注点。令 $A = \begin{bmatrix} a^1 & a^2 & \cdots & a^{2m} \end{bmatrix} = \begin{bmatrix} S & -S \end{bmatrix} \in \mathbb{R}^{n\times 2m}$ 对于某个 $S = \begin{bmatrix} s^1 & \cdots & s^m \end{bmatrix} \in \mathbb{R}^{n\times m}$，并令 $B_i = T_i \in \mathbb{R}^{n\times k_i}$，且对于所有 $i \in \{1,\dots,m\}$. 都有 $B_{m+i} = -T_i$。假设 $x^0 \oplus S \oplus T_i$、$x^0 \oplus (-S) \oplus (-T_i)$ $x^0 \oplus (\pm S)$ $x^0 \oplus (\pm T_i)$ 对于所有 $i \in \{1,\dots,m\}$. 都包含在 $\operatorname{dom} f$ 中，则

$$\nabla_s^2 f(x^0; A, B_{1:2m}) = \nabla_c^2 f(x^0; S, T_{1:m}).$$

证明。我们有

$$\begin{aligned}
&\nabla_s^2 f(x^0; A, B_{1:2m}) \\
&= \begin{bmatrix} S^\top \\ -S^\top \end{bmatrix}^\dagger \begin{bmatrix} \left(\nabla_s f(x^0 + a^1; B_1) - \nabla_s f(x^0; B_1)\right)^\top \\ \vdots \\ \left(\nabla_s f(x^0 + a^{2m}; B_{2m}) - \nabla_s f(x^0; B_{2m})\right)^\top \end{bmatrix} \\
&= \frac{1}{2}\begin{bmatrix} (S^\top)^\dagger & -(S^\top)^\dagger \end{bmatrix} \begin{bmatrix} \left(\nabla_s f(x^0 + a^1; B_1) - \nabla_s f(x^0; B_1)\right)^\top \\ \vdots \\ \left(\nabla_s f(x^0 + a^{2m}; B_{2m}) - \nabla_s f(x^0; B_{2m})\right)^\top \end{bmatrix} \\
&= \frac{1}{2}(S^\top)^\dagger \begin{bmatrix} \left(\nabla_s f(x^0 + s^1; T_1) - \nabla_s f(x^0; T_1)\right)^\top \\ \vdots \\ \left(\nabla_s f(x^0 + s^m; T_m) - \nabla_s f(x^0; T_m)\right)^\top \end{bmatrix} + \frac{1}{2}(-S^\top)^\dagger \begin{bmatrix} \left(\nabla_s f(x^0 - s^1; -T_1) - \nabla_s f(x^0; -T_1)\right)^\top \\ \vdots \\ \left(\nabla_s f(x^0 - s^m; -T_m) - \nabla_s f(x^0; -T_m)\right)^\top \end{bmatrix} \\
&= \frac{1}{2}\left(\nabla_s^2 f(x^0; S, T_{1:m}) + \nabla_s^2 f(x^0; -S, -T_{1:m})\right) \\
&= \nabla_c^2 f(x^0; S, T_{1:m}).
\end{aligned}$$

□

现在我们可以考察 GSH 和 GCSH 的误差界了。

4 误差界

本节给出 GSH 和 GCSH 的误差界。这些误差界分为两类：一般情形，其中所有矩阵 T_i 不一定相等；以及特殊情形，其中所有矩阵 $T_i =: \overline{T}$ 相等。GSH 的误差界表明，它对完整海森矩阵 $\nabla^2 f(x^0)$ 的近似具有一阶精度；若 GSH 是 S-欠定或 $T_{1:m}$-欠定的，则其对部分海森矩阵 $\operatorname{Proj}_{S,\, T_{1:m}} \nabla^2 f(x^0)$ 的近似具有一阶精度。类似地，GCSH 的误差界表明，它对完整海森矩阵 $\nabla^2 f(x^0)$ 的近似具有二阶精度；若 GCSH 是 S-欠定或 $T_{1:m}$-欠定的，则其对部分海森矩阵 $\operatorname{Proj}_{S,\, T_{1:m}} \nabla^2 f(x^0)$ 的近似具有二阶精度。

在介绍误差界之前，我们注意到，在某些情况下，GSH(GCSH)在 S 和 $T_{1:m}$ 上的投影等于GSH(GCSH)。

命题 4.1. 设 f: dom $f \subseteq \mathbb{R}^n \to \mathbb{R}$ 且 $x^0 \in$ dom f.。设 $S = \begin{bmatrix} s^1 & \cdots & s^m \end{bmatrix} \in \mathbb{R}^{n\times m}$ 和$T_i \in \mathbb{R}^{n\times k_i}$.。假设 $x^0 \oplus (\pm S)$ $x^0 \oplus (\pm T_i)$ $x^0 \oplus S \oplus T_i$ 和 $x^0 \oplus -S \oplus -T_i$ 对所有 i. 都包含在dom f 中。那么以下结论成立。

(i) 如果对于所有 $i \in \{1, \ldots, m\}$, S列满秩或 T_i 满行秩，则

$$\mathrm{Proj}_{S,T_{1:m}} \nabla_s^2 f(x^0; S, T_{1:m}) = \nabla_s^2 f(x^0; S, T_{1:m}) \tag{7}$$

and

$$\mathrm{Proj}_{S,T_{1:m}} \nabla_c^2 f(x^0; S, T_{1:m}) = \nabla_c^2 f(x^0; S, T_{1:m}). \tag{8}$$

(ii) 如果 $T_1 = T_2 = \cdots = T_m =: \overline{T}$，那么

$$\mathrm{Proj}_{S,\overline{T}} \nabla_s^2 f(x^0; S, \overline{T}) = \nabla_s^2 f(x^0; S, \overline{T}) \tag{9}$$

and

$$\mathrm{Proj}_{S,\overline{T}} \nabla_c^2 f(x^0; S, \overline{T}) = \nabla_c^2 f(x^0; S, \overline{T}). \tag{10}$$

证明。假设 S是列满秩。由于 $T_i^\dagger T_i T_i^\dagger = T_i^\dagger$ 根据穆尔—彭罗斯伪逆的性质 (ii)，我们得到

$$\begin{aligned}
\mathrm{Proj}_{S,T_{1:m}} \nabla_s^2 f(x^0; S, T_{1:m}) &= \sum_{i=1}^m (S^\top)^\dagger e_m^i (e_m^i)^\top S^\top \nabla_s^2 f(x^0; S, T_{1:m}) T_i T_i^\dagger \\
&= \sum_{i=1}^m (S^\top)^\dagger e_m^i (e_m^i)^\top S^\top (S^\top)^\dagger \delta_s^2(x^0; S, T_{1:m}) T_i T_i^\dagger \\
&= (S^\top)^\dagger \sum_{i=1}^m \mathrm{Diag}(e_m^i) \delta_s^2(x^0; S, T_{1:m}) T_i T_i^\dagger \\
&= (S^\top)^\dagger \delta_s^2(x^0; S, T_{1:m}) \\
&= \nabla_s^2 f(x^0; S, T_{1:m}).
\end{aligned}$$

式（8）同理可得。

现在假设 T_i 对于所有 $i \in \{1, \ldots, m\}$. 都是满行秩的，然后我们知道 $T_i T_i^\dagger = \mathrm{Id}_n$ 对于所有 $i \in \{1, \ldots, m\}$. 我们得到

$$\begin{aligned}
\mathrm{Proj}_{S,T_{1:m}} \nabla_s^2 f(x^0; S, T_{1:m}) &= \sum_{i=1}^m (S^\top)^\dagger e_m^i (e_m^i)^\top S^\top \nabla_s^2 f(x^0; S, T_{1:m}) T_i T_i^\dagger \\
&= \sum_{i=1}^m (S^\top)^\dagger e_m^i (e_m^i)^\top S^\top \nabla_s^2 f(x^0; S, T_{1:m}) \\
&= (S^\top)^\dagger \left(\sum_{i=1}^m e_m^i (e_m^i)^\top \right) S^\top \nabla_s^2 f(x^0; S, T_{1:m}) \\
&= (S^\top)^\dagger S^\top (S^\top)^\dagger \delta_s^2(x^0; S, T_{1:m}) \\
&= (S^\top)^\dagger \delta_s^2(x^0; S, T_{1:m}) = \nabla_s^2 f(x^0; S, T_{1:m}).
\end{aligned}$$

方程（8）同理可得。

最后，假设 $T_1 = T_2 = \cdots = T_m =: \overline{T}$。我们有

$$
\begin{aligned}
\mathrm{Proj}_{S,\overline{T}} \nabla_s^2 f(x^0; S, \overline{T}) &= (S^\top)^\dagger S^\top \nabla_s^2 f(x^0; S, \overline{T}) \overline{T}\overline{T}^\dagger \\
&= (S^\top)^\dagger S^\top (S^\top)^\dagger \delta_s^2(x^0; S, \overline{T}) \overline{T}\overline{T}^\dagger \\
&= (S^\top)^\dagger S^\top (S^\top)^\dagger \begin{bmatrix} \left(\nabla_s f(x^0 + s^1; \overline{T}) - \nabla_s f(x^0; \overline{T})\right)^\top \\ \vdots \\ \left(\nabla_s f(x^0 + s^m; \overline{T}) - \nabla_s f(x^0; \overline{T})\right)^\top \end{bmatrix} \overline{T}\overline{T}^\dagger \\
&= (S^\top)^\dagger S^\top (S^\top)^\dagger \begin{bmatrix} \left((\overline{T}^\top)^\dagger \delta_s(x^0 + s^1; \overline{T}) - (\overline{T}^\top)^\dagger \delta_s(x^0; \overline{T})\right)^\top \\ \vdots \\ \left((\overline{T}^\top)^\dagger \delta_s(x^0 + s^m; \overline{T}) - (\overline{T}^\top)^\dagger \delta_s(x^0; \overline{T})\right)^\top \end{bmatrix} \overline{T}\overline{T}^\dagger \\
&= (S^\top)^\dagger S^\top (S^\top)^\dagger \begin{bmatrix} \left(\delta_s(x^0 + s^1; \overline{T}) - \delta_s(x^0; \overline{T})\right)^\top \\ \vdots \\ \left(\delta_s(x^0 + s^m; \overline{T}) - \delta_s(x^0; \overline{T})\right)^\top \end{bmatrix} \overline{T}^\dagger \overline{T}\overline{T}^\dagger \\
&= (S^\top)^\dagger \begin{bmatrix} \left(\delta_s(x^0 + s^1; \overline{T}) - \delta_s(x^0; \overline{T})\right)^\top \\ \vdots \\ \left(\delta_s(x^0 + s^m; \overline{T}) - \delta_s(x^0; \overline{T})\right)^\top \end{bmatrix} \overline{T}^\dagger
\end{aligned}
$$

根据Moore–Penrose伪逆的性质 (ii)。因此，$\mathrm{Proj}_{S,\,\overline{T}} \nabla_s^2 f(x^0;\, S, \overline{T}) = \nabla_s^2 f(x^0;\, S, \overline{T})$。式 (10) 同理可得。

□

在 th e 以下定理以及本文的其余部分，我们使用记号 n

$$
\begin{aligned}
\widehat{T}_i &= T_i / \Delta_{T_i}, \quad i \in \{1, \ldots, m\}, \\
\Delta_u &= \max\{\Delta_S, \Delta_{T_1}, \ldots, \Delta_{T_m}\}, \\
\Delta_l &= \min\{\Delta_S, \Delta_{T_1}, \ldots, \Delta_{T_m}\}, \\
\Delta_T &= \max\{\Delta_{T_1}, \ldots, \Delta_{T_m}\} \\
\widehat{T} &= \widehat{T}_i \quad \text{such that} \quad \left\|\widehat{T}_i^\dagger\right\| \quad \text{is maximal}, \quad i \in \{1, \ldots, m\}, \\
k &= \max\{k_1, \ldots, k_m\}, \\
H &= \nabla^2 f(x^0).
\end{aligned}
$$

定理 4.2（GSH的误差界）。设 $f\colon \mathrm{dom}\, f \subseteq \mathbb{R}^n \to \mathbb{R}$ 是 $\mathcal{C}^3$ 在 $B(x^0, \overline{\Delta})$ 上，其中$x^0 \in \mathrm{dom}\, f$是关注点，且 $\overline{\Delta} > 0$。分别记 $L_{\nabla^2 f} \geq 0$为 $\nabla^2 f$在 $\overline{B}(x^0, \overline{\Delta})$上的利普希茨常数。设 $S = \begin{bmatrix} s^1 & s^2 & \cdots & s^m \end{bmatrix} \in \mathbb{R}^{n \times m}$且 $T_i = \begin{bmatrix} t_i^1 & t_i^2 & \cdots & t_i^{k_i} \end{bmatrix} \in \mathbb{R}^{n \times k_i}$对所有 $i \in \{1, \ldots, m\}$. 成立。假设 $B(x^0, \Delta_{T_i}) \subset B(x^0, \overline{\Delta})$ 且 $B(x^0 + s^i, \Delta_{T_i}) \subset B(x^0, \overline{\Delta})$ 对所有$i \in \{1, \ldots m\}$.成立

(i) 若对于所有 $i \in \{1, \ldots, m\}$， S 是列满秩矩阵或 T_i 是满行秩矩阵，则

$$\begin{aligned} \|\operatorname{Proj}_{S,T_{1:m}} \nabla_s^2 f(x^0; S, T_{1:m}) - \operatorname{Proj}_{S,T_{1:m}} \nabla^2 f(x^0)\| &= \|\nabla_s^2 f(x^0; S, T_{1:m}) - \operatorname{Proj}_{S,T_{1:m}} \nabla^2 f(x^0)\| \\ &\le 4m\sqrt{k} L_{\nabla^2 f} \|(\widehat{S}^\top)^\dagger\| \|\widehat{T}^\dagger\| \left(\frac{\Delta_u}{\Delta_l}\right)^2 \Delta_u. \end{aligned} \tag{11}$$

(ii) 如果 $T_1 = T_2 = \cdots = T_m =: \overline{T}$，则

$$\begin{aligned} \|\operatorname{Proj}_{S,\overline{T}} \nabla_s^2 f(x^0; S, \overline{T}) - \operatorname{Proj}_{S,\overline{T}} \nabla^2 f(x^0)\| &= \|\nabla_s^2 f(x^0; S, \overline{T}) - \operatorname{Proj}_{S,\overline{T}} \nabla^2 f(x^0)\| \\ &\le 4\sqrt{mk} L_{\nabla^2 f} \frac{\Delta_u}{\Delta_l} \left\|(\widehat{S}^\top)^\dagger\right\| \left\|\widehat{T}^\dagger\right\| \Delta_u. \end{aligned} \tag{12}$$

证明。第（i）项和第（ii）项中的等式分别由（7）和（9）得到。我们有

$$\begin{aligned} &\|\nabla_s^2 f(x^0; S, T_{1:m}) - \operatorname{Proj}_{S,T_{1:m}} H\| \\ &= \left\|(S^\top)^\dagger \delta_s^2(x^0; S, T_{1:m}) - \sum_{i=1}^m (S^\top)^\dagger e_m^i (e_m^i)^\top S^\top H T_i T_i^\dagger\right\| \\ &\le \frac{1}{\Delta_S} \left\|(\widehat{S}^\top)^\dagger\right\| \left\|\delta_s^2(x^0; S, T_{1:m}) - \sum_{i=1}^m (s^i)^\top H T_i T_i^\dagger\right\| \\ &= \frac{1}{\Delta_S} \left\|(\widehat{S}^\top)^\dagger\right\| \left\|\sum_{i=1}^m (e_m^i)^\top \delta_s^2(x^0; S, T_{1:m}) - (s^i)^\top H T_i T_i^\dagger\right\| \\ &\le \sum_{i=1}^m \frac{1}{\Delta_S \Delta_{T_i}} \left\|(\widehat{S}^\top)^\dagger\right\| \left\|\widehat{T}_i^\dagger\right\| \left\|\big(\delta_s(x^0 + s^i; T_i) - \delta_s(x^0; T_i)\big)^\top - (s^i)^\top H T_i\right\|. \end{aligned} \tag{13}$$

设 $u_i = (\delta_s(x^0 + s^i;\ T_i) - \delta_s(x^0;\ T_i))^\top \in \mathbb{R}^{k_i}$ 对所有 $\{1, \ldots, m\}$. 接下来我们为每个$\big[u_i - (s^i)^\top H T_i\big]_{j_i}$， $j_i \in \{1, \ldots, k_i\}$. 求一个界。注意

$$\big[u_i - (s^i)^\top H T_i\big]_{j_i} = f(x^0 + s^i + t_i^{j_i}) - f(x^0 + s^i) - f(x^0 + t_i^{j_i}) + f(x^0) - (s^i)^\top H t_i^{j_i}.$$

每个函数值 $f(x^0 + s^i + t_i^{j_i})$ $f(x^0 + s^i)$ 和 $f(x^0 + t_i^{j_i})$ 都可以写成关于 x^0 的二阶泰勒展开，并加上余项 $R_2(\cdot)$。因此，我们得到

$$\begin{aligned} &\big|f(x^0 + s^i + t_i^{j_i}) - f(x^0 + s^i) - f(x^0 + t_i^{j_i} + f(x^0) - (s^i)^\top H T_i^{j_i}\big| \\ &\le |R_2(x^0 + s^i + t_i^{j_i})| + |R_2(x^0 + s^i)| + |R_2(x^0 + t_i^{j_i})| \\ &\le \frac{1}{6} L_{\nabla^2 f} \|s^i + t_i^{j_i}\|^3 + \frac{1}{6} L_{\nabla^2 f} \|s^i\|^3 + \frac{1}{6} L_{\nabla^2 f} \|t_i^{j_i}\|^3 \\ &\le \frac{1}{2} L_{\nabla^2 f} (\Delta_S + \Delta_{T_i})^3. \end{aligned} \tag{14}$$

将 (14) 代入 (13)，得到

$$\begin{aligned}\|\nabla_s^2 f(x^0;S,T_{1:m})-\mathrm{Proj}_{S,T_{1:m}}H\| &\le \sum_{i=1}^m \frac{1}{\Delta_S\Delta_{T_i}}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\widehat{T}_i^\dagger\right\|\frac{\sqrt{k_i}}{2}L_{\nabla^2 f}(\Delta_S+\Delta_{T_i})^3\\ &\le \sum_{i=1}^m \frac{1}{\Delta_l^2}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\widehat{T}^\dagger\right\|\frac{\sqrt{k}}{2}L_{\nabla^2 f}(2\Delta_u)^3\\ &= 4m\sqrt{k}L_{\nabla^2 f}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\widehat{T}^\dagger\right\|\left(\frac{\Delta_u}{\Delta_l}\right)^2\Delta_u.\end{aligned}$$

现在，假设 $T_1=T_2=\cdots=T_m=:\ \overline{T}$. 我们有

$$\begin{aligned}\|\nabla_s^2 f(x^0;S,\overline{T})-\mathrm{Proj}_{S,\overline{T}}H\| &= \left\|(S^\top)^\dagger\delta_s^2(x^0;S,\overline{T})-(S^\top)^\dagger S^\top H\overline{T}\overline{T}^\dagger\right\|\\ &\le \frac{1}{\Delta_S}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\delta_s^2(x^0;S,\overline{T})-S^\top H\overline{T}\overline{T}^\dagger\right\|\\ &\le \frac{1}{\Delta_S\Delta_T}\|(\widehat{S}^\top)^\dagger\|\|\widehat{T}^\dagger\|\|\|(D_s^2)^\top-S^\top H\overline{T}\|, \qquad (15)\end{aligned}$$

其中

$$(D_s^2)^\top=\begin{bmatrix}\left(\delta_s(x^0+s^1;\overline{T})-\delta_s(x^0;\overline{T})\right)^\top\\ \vdots\\ \left(\delta_s(x^0+s^m;\overline{T})-\delta_s(x^0;\overline{T})\right)^\top\end{bmatrix}\in\mathbb{R}^{m\times k}.$$

每个条目 $\left[(D_s^2)^\top-S^\top H\overline{T}\right]_{i,j}$都有在（14）中得到的前一个界限。由于

$$\|(D_s^2)^\top-S^\top H\overline{T}\|\le\|(D_s^2)^\top-S^\top H\overline{T}\|_F,$$

将 (14) 代入 (15)，得到不等式

$$\begin{aligned}\|\nabla_s^2 f(x^0;S,\overline{T})-\mathrm{Proj}_{S,\overline{T}}H\| &\le \frac{\sqrt{mk}}{2}L_{\nabla^2 f}\frac{(\Delta_S+\Delta_T)^3}{\Delta_S\Delta_T}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\widehat{T}^\dagger\right\|\\ &\le 4\sqrt{mk}L_{\nabla^2 f}\frac{\Delta_u}{\Delta_l}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\widehat{T}^\dagger\right\|\Delta_u.\end{aligned}$$

□

之前的误差界表明，GSH是部分海森矩阵的一阶精确近似；当GSH是 S-确定\超定且 $T_{1:m}$-确定\超定时，GSH 也是完整海森矩阵的一阶精确近似（等价地，$(S^\top)^\dagger S^\top=\mathrm{Id}_n$ 且$T_iT_i^\dagger=\mathrm{Id}_n$ 对所有 i均成立）。比值 $\frac{\Delta_u}{\Delta_l}$表明，矩阵的半径应取为相同的量级。如果其中一个半径减小，这意味着所有半径都应以相同的幅度减小。最后，误差界中出现 m和 k 大致表明，采用 S-超定或 $T_{1:m}$-超定并无优势。关于这一主题的更多细节可参见 [11, 27]。

当所有矩阵 T_i 都相等时，请注意，条目 (ii) 涵盖了定义 2.9 中的全部 16 种情况。当所有矩阵 T_i 都不相等时，前一定理中的条目 (i) 涵盖了 16 种情况中的 12 种。它未涵盖 S-不定\超定和 $T_{1:m}$-不定 \欠定的情况。这些情况留作未来研究方向，并在结论中讨论。

定理 4.3（GCSH 的误差界）。设 $f\colon \operatorname{dom} f \subseteq \mathbb{R}^n \to \mathbb{R}$ 是 $B(x^0, \overline{\Delta})$ 上的 $\mathcal{C}^4$，其中 $x^0 \in \operatorname{dom} f$ 是关注点，且 $\overline{\Delta} > 0$。记 $L_{\nabla^3 f}$ 为 $\overline{B}(x^0, \overline{\Delta})$ 上 $\nabla^3 f$ 的利普希茨常数。设 $S = \begin{bmatrix} s^1 & s^2 & \cdots & s^m \end{bmatrix} \in \mathbb{R}^{n\times m}$，$T_i = \begin{bmatrix} t_i^1 & t_i^2 & \cdots & t_i^{k_i} \end{bmatrix} \in \mathbb{R}^{n\times k_i}$ 满足：对于所有 $i \in \{1, \ldots, m\}$，$\overline{B}(x^0 \pm s^i, \Delta_{T_i}) \subset B(x^0, \overline{\Delta})$。

(i) 若对于所有 $i \in \{1, \ldots, m\}$，S 列满秩或 T_i 满行秩，则

$$\begin{aligned} \|\operatorname{Proj}_{S,T_{1:m}} \nabla_c^2 f(x^0; S, T_{1:m}) - \operatorname{Proj}_{S,T_{1:m}} \nabla^2 f(x^0)\| &= \|\nabla_c^2 f(x^0; S, T_{1:m}) - \operatorname{Proj}_{S,T_{1:m}} \nabla^2 f(x^0)\| \\ &\leq 2m\sqrt{k} L_{\nabla^3 f} \left(\frac{\Delta_u}{\Delta_l}\right)^2 \left\|(\widehat{S}^\top)^\dagger\right\| \left\|\left(\widehat{T}\right)^\dagger\right\| \Delta_u^2. \end{aligned} \tag{16}$$

(ii) 若 $T_1 = T_2 = \cdots = T_m =: \overline{T} \in \mathbb{R}^{n\times k}$，则

$$\begin{aligned} \left\|\operatorname{Proj}_{S,\overline{T}} \nabla_c^2 f(x^0; S, \overline{T}) - \operatorname{Proj}_{S,\overline{T}} \nabla^2 f(x^0)\right\| &= \left\|\nabla_c^2 f(x^0; S, \overline{T}) - \operatorname{Proj}_{S,\overline{T}} \nabla^2 f(x^0)\right\| \\ &\leq 2\sqrt{mk} L_{\nabla^3 f} \frac{\Delta_u}{\Delta_l} \left\|(\widehat{S}^\top)^\dagger\right\| \left\|\left(\widehat{T}\right)^\dagger\right\| \Delta_u^2. \end{aligned} \tag{17}$$

证明。等式 条目（i）和（ii）中的等式分别由（8）和（10）得出。我们有 e

$$\begin{aligned} &\left\|\nabla_c^2 f(x^0; S, T_{1:m}) - \operatorname{Proj}_{S,T_{1:m}} H\right\| \\ &= \left\|(S^\top)^\dagger \delta_c^2(x^0; S, T_{1:m}) - \sum_{i=1}^m (S^\top)^\dagger e_m^i (e_m^i)^\top S^\top H T_i T_i^\dagger\right\| \\ &\leq \frac{1}{\Delta_S} \left\|(\widehat{S}^\top)^\dagger\right\| \left\|\delta_c^2(x^0; S, T_{1:m}) - \sum_{i=1}^m (s^i)^\top H T_i T_i^\dagger\right\| \\ &= \frac{1}{\Delta_S} \left\|(\widehat{S}^\top)^\dagger\right\| \left\|\sum_{i=1}^m (e_m^i)^\top \delta_c^2(x^0; S, T_{1:m}) - (s^i)^\top H T_i T_i^\dagger\right\| \\ &\leq \sum_{i=1}^m \frac{1}{\Delta_S \Delta_{T_i}} \left\|(\widehat{S}^\top)^\dagger\right\| \left\|\widehat{T_i}^\dagger\right\| \left\|\frac{1}{2}\left(\delta_s(x^0 + s^i; T_i) + \delta_s(x^0 - s^i; -T_i) - \delta_s(x^0; T_i) - \delta_f^s(x^0; -T_i)\right)^\top - (s^i)^\top H T_i\right\|. \end{aligned} \tag{18}$$

令 $w_i = \frac{1}{2}(\delta_s(x^0 + s^i; T_i) + \delta_s(x^0 - s^i; -T_i) - \delta_s(x^0; T_i) - \delta_s(x^0; -T_i))^\top \in \mathbb{R}^{k_i}$. 接下来我们为每个元素 $\left[w_i - (s^i)^\top H T_i\right]_{j_i}, j_i \in \{1, \ldots, k_i\}$. 求一个界。利用 (6)，我们知道

$$\begin{aligned} &\left[w_i - (s^i)^\top H T_i\right]_{j_i} \\ =&\frac{1}{2}\left(f(x^0 + s^i + t_i^{j_i}) + f(x^0 - s^i - t_i^{j_i}) - f(x^0 + s^i) - f(x^0 - s^i) - f(x^0 + t_i^{j_i}) - f(x^0 - t_i^{j_i}) + 2f(x^0)\right) \\ &- (s^i)^\top H t_i^{j_i}. \end{aligned} \tag{19}$$

每个函数值 $f(x^0+s^i+t_i^{j_i})$ $f(x^0-s^i-t_i^{j_i})$ $f(x^0+s^i)$ $f(x^0-s^i)$ $f(x^0+t_i^{j_i})$ 和 $f(x^0-t_i^{j_i})$ 都可以写成关于 x^0 的三阶泰勒展开，外加余项 $R_3(\cdot)$。由此可得

$$
\begin{aligned}
&\left|\frac{1}{2}\left(f(x^0+s^i+t_i^{j_i})+f(x^0-s^i-t_i^{j_i})-f(x^0+s^i)-f(x^0-s^i)-f(x^0+t_i^{j_i})-f(x^0-t_i^{j_i})+2f(x^0)\right)-(s^i)^\top H t_i^{j_i}\right| \\
&\le \frac{1}{2}\left(\|R_3(x^0+s^i+t_i^{j_i})\|+\|R_3(x^0-s^i-t_i^{j_i})\|+\|R_3(x^0+s^i)\|+\|R_3(x^0-s^i)\|+\|R_3(x^0+t_i^{j_i}\|+\|R_3(x^0-t_i^{j_i})\|\right) \\
&\le \frac{1}{48}\left(L_{\nabla^3 f}\|s^i+t_i^{j_i}\|^4+\frac{1}{24}L_{\nabla^3 f}\|-s^i-t_i^{j_i}\|^4+L_{\nabla^3 f}\|-s^i\|^4+L_{\nabla^3 f}\|s^i\|^4+L_{\nabla^3 f}\|t_i^{j_i}\|+L_{\nabla^3 f}\|-t_i^{j_i}\|\right) \\
&\le \frac{1}{8}L_{\nabla^3 f}(\Delta_S+\Delta_{T_i})^4. \qquad (20)
\end{aligned}
$$

将（20）中的界代入（19），得到

$$
\begin{aligned}
\|\nabla_c^2 f(x^0;S,T_{1:m})-\mathrm{Proj}_{S,T_{1:m}}H\| &\le \sum_{i=1}^m \frac{1}{\Delta_S\Delta_{T_i}}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\widehat{T}_i^\dagger\right\|\sqrt{k_i}\frac{1}{8}L_{\nabla^3 f}(\Delta_S+\Delta_{T_i})^4 \\
&\le \sum_{i=1}^m \frac{1}{\Delta_l^2}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\widehat{T}^\dagger\right\|\sqrt{k}\frac{1}{8}L_{\nabla^3 f}(2\Delta_u)^4 \\
&= 2m\sqrt{k}L_{\nabla^3 f}\left(\frac{\Delta_u}{\Delta_l}\right)^2\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\left(\widehat{T}\right)^\dagger\right\|\Delta_u^2.
\end{aligned}
$$

现在，假设 $T_1=T_2=\cdots=T_m=:\overline{T}$. 我们有

$$
\begin{aligned}
\|\nabla_c^2 f(x^0;S,\overline{T})-\mathrm{Proj}_{S,\overline{T}}H\| &= \left\|(S^\top)^\dagger\delta_c^2(x^0;S,\overline{T})-(S^\top)^\dagger S^\top H\overline{T}\overline{T}^\dagger\right\| \\
&\le \frac{1}{\Delta_S}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\delta_c^2(x^0;S,\overline{T})-S^\top H\overline{T}\overline{T}^\dagger\right\| \\
&\le \frac{1}{\Delta_S\Delta_T}\|(\widehat{S}^\top)^\dagger\|\widehat{T}^\dagger\|\|\|(D_c^2)^\top-S^\top H\overline{T}\|, \qquad (21)
\end{aligned}
$$

其中

$$
(D_c^2)^\top=\frac{1}{2}\begin{bmatrix}\left(\delta_s(x^0+s^1;\overline{T})+\delta_s(x^0-s^1;-\overline{T})-\delta_s(x^0;\overline{T})-\delta_s(x^0;-\overline{T})\right)^\top \\ \vdots \\ \left(\delta_s(x^0+s^m;\overline{T})+\delta_s(x^0-s^m;-\overline{T})-\delta_s(x^0;\overline{T})-\delta_s(x^0;-\overline{T})\right)^\top\end{bmatrix}\in\mathbb{R}^{m\times k}.
$$

每个元素 $\left[(D_c^2)^\top-S^\top H\overline{T}\right]_{i,j}$ 都由（20）限定。由于

$$
\|(D_c^2)^\top-S^\top H\overline{T}\|\le\|(D_f^c)^\top-S^\top H\overline{T}\|_F,
$$

将 (20) 代入 (21)，可得不等式

$$
\begin{aligned}
\|\nabla_c^2 f(x^0;S,\overline{T})-\mathrm{Proj}_{S,\overline{T}}H\| &\le \frac{\sqrt{mk}}{24}L_{\nabla^3 f}\frac{(\Delta_S+\Delta_T)^4}{\Delta_S\Delta_T}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\widehat{T}^\dagger\right\| \\
&\le \frac{2\sqrt{mk}}{3}L_{\nabla^3 f}\frac{\Delta_u}{\Delta_l}\left\|(\widehat{S}^\top)^\dagger\right\|\left\|\widehat{T}^\dagger\right\|\Delta_u^2.
\end{aligned}
$$

□

定理 4.3 中针对 GCSH 提出的误差界与定理 4.2 中针对 GSH 的误差界之间的主要差异是

- 定理 4.3 中的误差界假设 $f \in \mathcal{C}^4$，而定理 4.2 中的误差界假设 $f \in \mathcal{C}^3$,
- 定理 4.3中的误差界涉及利普希茨常数 $L_{\nabla^3 f}$，而定理 4.2 中的误差界涉及利普希茨常数 $L_{\nabla^2 f}$,
- 定理4.3中的误差界具有二阶精度，而定理 4.2 中的误差界具有一阶精度。

定理 4.2和 4.3的条款 (ii)比条款 (i) 限制性更强，但作为回报，它可以带来额外的好处，例如对称性。在下面的命题中，我们证明，f 在 x^0 上关于 S 和$\overline{T}$ 的GSH（GCSH）的转置，与 f 在 x^0 上关于 $\overline{T}$和 S 的 GSH（GCSH）相同。

命题 4.4. 令 f: dom $f \subseteq \mathbb{R}^n \to \mathbb{R}$，并令 $x^0 \in$dom f 为关注点。令$S = \begin{bmatrix} s^1 & s^2 & \cdots & s^m \end{bmatrix} \in \mathbb{R}^{n\times m}$和 $\overline{T} = \begin{bmatrix} t^1 & t^2 & \cdots & t^k \end{bmatrix} \in \mathbb{R}^{n\times k}$满足 $x^0 \oplus S \oplus T_i, x^0 \oplus (-S) \oplus (-T_i)$ $x^0 \oplus (\pm S)$，且 $x^0 \oplus (\pm T_i)$ 对所有 $i \in \{1, \ldots, m\}$. 均包含于 dom f 中。则

$$\left(\nabla_s^2 f(x^0; S, \overline{T})\right)^\top = \nabla_s^2 f(x^0; \overline{T}, S),$$

and

$$\left(\nabla_c^2 f(x^0; S, \overline{T})\right)^\top = \nabla_c^2 f(x^0; \overline{T}, S).$$

证明。我们有

$$\begin{aligned}
\left(\nabla_s^2 f(x^0; S, \overline{T})\right)^\top &= \left((S^\top)^\dagger \delta_s^2(x^0; S, \overline{T})\right)^\top \\
&= (\delta_s^2(x^0; S, \overline{T}))^\top S^\dagger \\
&= \left[\nabla_s f(x^0 + s^1; \overline{T}) - \nabla_s f(x^0; \overline{T}) \quad \cdots \quad \nabla_s f(x^0 + s^m; \overline{T}) - \nabla_s f(x^0; \overline{T})\right] S^\dagger \\
&= (\overline{T}^\top)^\dagger \left[\delta_s(x^0 + s^1; \overline{T}) - \delta_s(x^0; \overline{T}) \quad \cdots \quad \delta_s(x^0 + s^m; \overline{T}) - \delta_s(x^0; \overline{T})\right] S^\dagger.
\end{aligned}$$

令 $D_s^2 = \left[\delta_s(x^0 + s^1; \overline{T}) - \delta_s(x^0; \overline{T}) \cdots \quad \delta_s(x^0 + s^m; \overline{T}) - \delta_s(x^0; \overline{T})\right] \in \mathbb{R}^{k\times m}$.。则

$$[D_s^2]_{i,j} = f(x^0 + s^i + t^j) - f(x^0 + s^i) - f(x^0 + t^j) + f(x^0).$$

请注意

$$D_s^2 = \begin{bmatrix} (\delta_s(x^0 + t^1; S) - \delta_s(x^0; S))^\top \\ \vdots \\ \left(\delta_s(x^0 + t^k; S) - \delta_s(x^0; S)\right)^\top \end{bmatrix}.$$

因此，

$$\begin{aligned}\left(\nabla_s^2 f(x^0;S,\overline{T})\right)^\top &= (\overline{T}^\top)^\dagger \begin{bmatrix} \left(\delta_s(x^0+t^1;S)-\delta_s(x^0;S)\right)^\top \\ \vdots \\ \left(\delta_s(x^0+t^k;S)-\delta_s(x^0;S)\right)^\top \end{bmatrix} S^\dagger \\ &= (\overline{T}^\top)^\dagger \begin{bmatrix} \left((S^\dagger)^\top\delta_s(x^0+t^1;S)-(S^\dagger)^\top\delta_s(x^0;S)\right)^\top \\ \vdots \\ \left((S^\dagger)^\top\delta_s(x^0+t^k;S)-(S^\dagger)^\top\delta_s(x^0;S)\right)^\top \end{bmatrix} \\ &= (\overline{T}^\top)^\dagger \delta_s^2(x^0;\overline{T},S) = \nabla_s^2 f(x^0;\overline{T},S).\end{aligned}$$

第二个等式由 GCSH 的定义（定义 2.7）得出。 □

在下一节中，我们将探讨如何用极少的采样点获得完整海森矩阵的 1 阶精度近似。

5 极小确定集与二次插值

本节将探讨如何选择方向矩阵 S 和 T_i，以用极少的采样点获得完整海森矩阵的近似。减少不同点的数量极为重要，因为这会减少计算 GSH 所需的不同函数求值次数。在本节中，我们均设 $T_1 = \cdots = T_m =: \overline{T} \in \mathbb{R}^{n\times k}$。我们证明，若 S 和 $\overline{T}$ 具有特定结构，则计算 GSH 所需的不同函数求值次数为 $(n+1)(n+2)/2.$。随后，我们将探讨使用这种结构时得到的一些结果。首先，我们定义一个用于计算 GSH 的集合。

定义 5.1. 用于计算 $\nabla_s^2 f(x^0;\, S,\overline{T})$ 的所有互异点的集合，称为 GSH 计算集，记作 $\mathcal{S}(x^0;\, S,\overline{T})$。

注意，$\mathcal{S}(x^0;\, S,\overline{T})$最多包含 $(m+1)(k+1)$ 个不同的点，但如果其中一些点重合，则包含的点可能更少。接下来，我们介绍 GSH 的极小确定集定义。

定义 5.2（GSH 的极小确定集）。设 $x^0 \in \mathbb{R}^n$ 为关注点。设 $S = \begin{bmatrix} s^1 & s^2 & \cdots & s^n \end{bmatrix} \in \mathbb{R}^{n\times n}$ 和 $\overline{T} = \begin{bmatrix} t^1 & t^2 & \cdots & t^n \end{bmatrix} \in \mathbb{R}^{n\times n}$。当且仅当 S 和 $\overline{T}$ 满秩，且 $\mathcal{S}(x^0;\, S,\overline{T})$ 恰好包含 $(n+1)(n+2)/2$ 个不同的点时，我们称 $\mathcal{S}(x^0;\, S,\overline{T})$ 是 x^0 处 GSH 的极小确定集。

该定义要求 S and $\overline{T}$ 恰好具有 n 列，并且为满秩。这是 GSH 为 S-确定和 $\overline{T}$-确定的情形。这意味着 S and $\overline{T}$ 的大小是最小的，从而确保 GSH 是完整海森矩阵的具有一阶精度的近似。

注意，GCSH 的极小确定集可以用类似的方式定义。显然，这样的集合必须包含多于 $(n+1)(n+2)/2.$ 对这一主题的研究留待未来研究方向。接下来我们展示可以为 GSH 创建一个极小确定集。

命题 5.3。设 x^0 为关注点。设 $S = \begin{bmatrix} s^1 & s^2 & \cdots & s^n \end{bmatrix} \in \mathbb{R}^{n\times n}$。对每个指标 $\ell \in \{0, 1, \ldots, n\}$，定义集合 U_ℓ 为

$$U_0 = S$$

and

$$U_\ell = \begin{bmatrix} s^1 - s^\ell & s^2 - s^\ell & \cdots & s^{\ell-1} - s^\ell & -s^\ell & s^{\ell+1} - s^\ell & \cdots & s^n - s^\ell \end{bmatrix} \in \mathbb{R}^{n\times n}, \ell \neq 0.$$

然后，对于每个 ℓ、$|\mathcal{S}(x^0;\ S、\ U_\ell)| \leq (n+1)\ (n+2)/2$。此外，如果 S 具有满秩，则 $|\mathcal{S}(x^0;\ S、\ U_\ell)| = (n+1)\ (n+2)/2$。

证明。不失一般性，令 $x^0 = \mathbf{0}$。首先，假设 $\ell \in \{1, 2, \ldots, n\}$. 对于任意函数$f$，考虑矩阵 $\delta_s^2(x^0; S, T)$，其中 $\overline{T} = U_\ell$. 对 $\nabla_s f(x^0; U_\ell)$ 的计算在这些点处计算 f

$$\begin{aligned} &\{\mathbf{0}, (s^1 - s^\ell), \ldots, (s^{\ell-1} - s^\ell), -s^\ell, (s^{\ell+1} - s^\ell), \ldots, (s^n - s^\ell)\} \\ &= \{\mathbf{0}\} \cup \{-s^\ell\} \cup \bigcup_{\substack{i=1 \\ i\neq\ell}}^{n} \left\{s^i - s^\ell\right\}. \end{aligned}$$

对于 $i \neq \ell$，$\nabla_s f(x^0 + s^i; U_k)$ 的计算在该点对 f 求值 s

$$\begin{aligned} &\{s^i, s^i + (s^1 - s^\ell), \ldots, s^i + (s^{\ell-1} - s^\ell), s^i - s^\ell, s^i + (s^{\ell+1} - s^\ell), \ldots, s^i + (s^n - s^\ell)\} \\ &= \{s^i\} \cup \{s^i - s^\ell\} \cup \bigcup_{\substack{j=1 \\ j\neq\ell}}^{n} \{s^i + s^j - s^\ell\}. \end{aligned}$$

$x^0 + s^\ell$；U_ℓ 各点处的 f 计算结果为 $\nabla_s f$

$$\begin{aligned} &\{s^\ell, s^\ell + (s^1 - s^\ell), \ldots, s^\ell + (s^{\ell-1} - s^\ell), s^\ell - s^\ell, s^\ell + (s^{\ell+1} - s^\ell), \ldots, s^\ell + (s^n - s^\ell)\} \\ &= \{s^\ell\} \cup \{\mathbf{0}\} \cup \bigcup_{\substack{i=1 \\ i\neq\ell}}^{n} \{s^i\} = \{\mathbf{0}\} \cup S. \end{aligned}$$

因此，f 在这些点处求值。

$$\begin{aligned} &\left(\{\mathbf{0}\} \cup \{-s^\ell\} \cup \bigcup_{\substack{i=1 \\ i\neq\ell}}^{n} \left\{s^i - s^\ell\right\}\right) \cup \left(\bigcup_{\substack{i=1 \\ i\neq\ell}}^{n} \{s^i\} \cup \bigcup_{\substack{i=1 \\ i\neq\ell}}^{n} \{s^i - s^\ell\} \cup \bigcup_{\substack{i=1 \\ i\neq\ell}}^{n} \bigcup_{\substack{j\geq i \\ j\neq\ell}} \{s^i + s^j - s^\ell\}\right) \cup \left(\{\mathbf{0}\} \cup S\right) \\ &= \{\mathbf{0}\} \cup S \cup \{-s^\ell\} \cup \bigcup_{\substack{i=1 \\ i\neq\ell}}^{n} \left\{s^i - s^\ell\right\} \cup \bigcup_{\substack{i=1 \\ i\neq\ell}}^{n} \bigcup_{\substack{j\geq i \\ j\neq\ell}} \{s^i + s^j - s^\ell\}. \end{aligned} \tag{22}$$

这最多为 $1 + n + 1 + (n-1)\ + (n-1)(n-2)/2 = (n+1)(n+2)/2$ 个点。

现在，假设 $\ell = 0$. 按照与上述类似的过程，我们发现 f 在这些点处求值。

$$\left(\{\mathbf{0}\} \cup \bigcup_{i=1}^{n} \{s^i\}\right) \ \cup \ \left(\bigcup_{i=1}^{n} \{2s^i\} \cup \bigcup_{i=1}^{n} \bigcup_{j>i} \{s^i + s^j\}\right). \tag{23}$$

这至多为 $(n+1)(n+2)/2$ 个点。

最后，如果 S 是满秩的，那么 (22) 中的四个集合与 (23)中的四个集合互不相交，因此我们恰好有 $(n+1)(n+2)/2$ 函数求值。 □

在命题 5.3 中使用 $S = \mathrm{Id}_n$，我们可以为 GSH 构造 $n+1$ 个典范极小确定集。

定义 5.4（GSH 的第 ℓ 个典范极小确定集）。设 $x^0 \in \mathbb{R}^n$ 为关注点。设 $S = \mathrm{Id}_n$。固定 $\ell \in \{0, 1, \ldots, n\}$。令

$$E_0 = \mathrm{Id}_n$$

and

$$E_\ell = \begin{bmatrix} e^1 - e^\ell & e^2 - e^\ell & \cdots & e^{\ell-1} - e^\ell & -e^\ell & e^{\ell+1} - e^\ell & \cdots & e^n - e^\ell \end{bmatrix}, \ell \neq 0.$$

则 $\mathcal{S}(x^0; \mathrm{Id}_n, E_\ell)$ 称为 GSH 在 x^0 处的第 ℓ 个典范极小确定集。

根据命题 5.3，以及定义 5.4 中任意矩阵 E_k 都是满秩这一事实，GSH 的第 ℓ 个典范极小确定集确实是 GSH 的一个极小确定集。此后，我们用记号 $\mathcal{M}(x^0; S, U_k)$ 表示 GSH 在 x^0 处的一个极小确定集，其形式为命题 5.3 中构造的形式。

注意，S 和 $\overline{T}$ 中方向的顺序是任意的。因此，显然，若 $\mathcal{S}(x^0; S, \overline{T})$ 是 GSH 的一个极小确定集，且 $P_1, P_2 \in \mathbb{R}^{n\times n}$ 是置换矩阵，则 $\mathcal{S}(x^0; SP_1, \overline{T}P_2)$ 也是 GSH 在 x^0 处的一个极小确定集。下一个命题将进一步阐述这一思路，并展示如何构造 GSH 的极小确定集。

命题 5.5. 设 $x^0 \in \mathbb{R}^n$ 为关注点。设 $S, \overline{T} \in \mathbb{R}^{n\times n}$。设 $N \in \mathbb{R}^{n\times n}$ 为可逆矩阵，$P_1, P_2 \in \mathbb{R}^{n\times n}$ 为置换矩阵。当且仅当 $\mathcal{S}(x^0; NSP_1, N\overline{T}P_2)$ 是 x^0处 GSH 的极小确定集时，$\mathcal{S}(x^0; S, \overline{T})$ 才是 x^0处 GSH 的极小确定集。

证明。该证明可直接由矩阵的性质得出。 □

由此可得，$\mathcal{S}(x^0; S, \overline{T})$ 是 x^0处 GSH 的极小确定集，当且仅当集合$\mathcal{S}(x^0; \beta SP_1, \beta\overline{T}P_2)$ 是 x^0处 GSH 的极小确定集，其中 β是非零标量，P_1, P_2是置换矩阵。

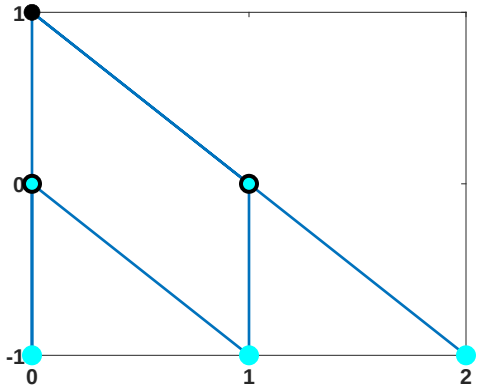
例 5.6. 设 $x^0 = (0, 0)$。GSH 在 $\mathbb{R}^2$ 中的第二规范极小确定集包含点 $(0, -1)$、$(0, 0)$、$(0, 1)$、$(1, -1)$、$(1, 0)$ 和 $(2, -1)$。在此情况下，$S = \{e^1, e^2\}$ 和 $\overline{T} = \{(e^1 - e^2) \ -e^2\}$。图 1 展示了这个集合。

$\{$中的点 $(0, 0)\} \cup \{(0,0) \oplus S\} = \{0, e^1, e^2\}$ 以实心黑色边框表示。这些是将计算单纯形梯度的基点。线表示对应于从 $\{(0,0)\} \cup \{(0,0) \oplus S\}$发出的 $\overline{T}$向量。$(\{(0, 0)\} \cup \{(0,0) \oplus S\}) + \overline{T}$中的点以青色核心表示。这些是用于构造单纯形梯度的点。注意，点$(0, 0)$ 和 $(1, 0)$ 同时具有黑色边框和青色核心。这些是使函数求值次数减少到 $(n+1)(n+2)/2 = 6$的公共点。

接下来，我们证明，对于形如 $\mathcal{M}(x^0; S, U_\ell)$的GSH 的每个极小确定集，它都适用于二次插值。然后我们证明逆命题并不成立；可以构造出一个适用于二次插值的集合，但它不具有 $\mathcal{M}(x^0; S, U_\ell)$ 的形式。

命题 5.7。令 $S \in \mathbb{R}^{n\times n}$ 满秩。选取 $\ell \in \{0, 1, \ldots, n\}$，并按照命题5.3 定义 $\overline{T} := U_\ell$。那么 $\mathcal{M}(x^0; S, \overline{T})$ 对二次插值是适定的。

图1：GSH 在 x^0 处、 $\mathbb{R}^2$ 中的第二典范极小确定集。



证明。注意到 $\mathcal{M}(x^0;\, S, \overline{T})$ 是二次插值适定的，当且仅当集合 $\mathcal{M}(x^0;\, S, \overline{T})\oplus -x^0$ 是二次插值适定的，因此不失一般性地假设 $x^0 = \mathbf{0}$。
情况I： $\overline{T} = U_\ell$ 对于 $\ell \in \{1, 2, \ldots, n\}$。不失一般性，根据命题5.5，假设$\ell = n$。包含在 $\mathcal{M}(x^0;\ S, U_n)$ 中的点是

$$\{\mathbf{0}\} \cup S \cup \{-s^n\} \cup \bigcup_{i=1}^{n-1} \{s^i - s^n\} \cup \bigcup_{i=1}^{n-1} \bigcup_{\substack{j \geq i \\ j \neq n}} \{s^i + s^j - s^n\}$$

$$= \{\mathbf{0}\} \cup S \cup \{-s^n\} \cup \bigcup_{i=1}^{n-1} \{s^i - s^n\} \cup \{2s^i - s^n\}_{i=1}^{n-1} \cup \bigcup_{i=1}^{n-1} \bigcup_{\substack{j > i \\ j \neq n}} \{s^i + s^j - s^n\}.$$

我们证明，利用这个集合，(1) 的唯一解是平凡解。考虑点 0，我们得到

$$\alpha_0 = 0.$$

考虑点 s^i 对于 $i \in \{1, 2, \ldots, n\}$，并注意到对于所有 $i \in \{1, 2, \ldots, n\}$、 $s^i = Se^i$，我们得到

$$\overline{\alpha}^\top e^i + \frac{1}{2}(e^i)^\top \widehat{\mathbf{H}} e^i = 0, \tag{24}$$

其中 $\overline{\alpha}^\top = \alpha^\top S$ 和 $\widehat{\mathbf{H}} = S^\top \mathbf{H} S$. 注意， $\widehat{\mathbf{H}}$ 是对称的。考虑点 $s^i - s^n$，其中$i \in \{1, 2, \ldots, n-1\}$，我们得到

$$\overline{\alpha}^\top e^i - \overline{\alpha}^\top e^n - (e^i)^\top \widehat{\mathbf{H}} e^n + \frac{1}{2}(e^i)^\top \widehat{\mathbf{H}} e^i + \frac{1}{2}(e^n)^\top \widehat{\mathbf{H}} e^n = 0.$$

利用 (24)，可简化为

$$-\overline{\alpha}^\top e^n - (e^i)^\top \widehat{\mathbf{H}} e^n + \frac{1}{2}(e^n)^\top \widehat{\mathbf{H}} e^n = 0. \tag{25}$$

考虑点 $-s^n$，我们发现

$$-\overline{\alpha}^\top e^n + \frac{1}{2}(e^n)^\top \widehat{\mathbf{H}} e^n = 0, \tag{26}$$

从而将 (25) 化为

$$-(e^i)^\top \widehat{\mathbf{H}} e^n = 0 \quad \text{for } \ i \in \{1, 2, \ldots, n-1\}.$$

因此，对于所有 $i \in \{1, 2, \ldots, n-1\}.$，$(e^i)^\top \widehat{\mathbf{H}} e^n = \widehat{\mathbf{H}}_{i,\ n} = \widehat{\mathbf{H}}_{n,\ i} = 0$。将 (24) 在 $i = n$ (26)
处结合并乘以 -1，可得

$$\overline{\alpha}^\top e^n + \frac{1}{2}(e^n)^\top \widehat{\mathbf{H}} e^n = 0 = \overline{\alpha}^\top e^n - \frac{1}{2}(e^n)^\top \widehat{\mathbf{H}} e^n,$$

这意味着 $(e^n)^\top \widehat{\mathbf{H}} e^n = \widehat{\mathbf{H}}_{n,\ n} = 0$。考虑点 $2s^i - s^n$（其中 $i \in \{1, 2, \ldots, n-1\}$），可得

$$2\overline{\alpha}^\top e^i - \overline{\alpha}^\top e^n + 2(e^i)^\top \widehat{\mathbf{H}} e^i + \frac{1}{2}(e^n)^\top \widehat{\mathbf{H}} e^n = 0.$$

利用 (26)，可将其化简为

$$2\overline{\alpha}^\top e^i + 2(e^i)^\top \widehat{\mathbf{H}} e^i = 0.$$

将 (24) 乘以 2 并代入上式，可得对于所有 $i \in \{1, 2, \ldots, n-1\}.$，$(e^i)^\top \widehat{\mathbf{H}} e^i = \widehat{\mathbf{H}}_{i,\ i} = 0$。这意味着对于所有 $i \in \{1, 2, \ldots, n\}$，均有 $\overline{\alpha}_i = \overline{\alpha}^\top e^i = 0$，即 $\overline{\alpha} = \mathbf{0}$。最后，考虑点 $s^i + s^j - s^n$（其中 $i \neq j$、i、$j \in \{1, 2, \ldots, n-1\}.$）。由于对于 $i \in \{1, 2, \ldots, n\}$，$\widehat{\mathbf{H}}_{i,\ i} = 0$，而对于 $i \in \{1, 2, \ldots, n-1\}$ 和 $\overline{\alpha} = \mathbf{0}$，$\widehat{\mathbf{H}}_{i,\ n} = \mathbf{H}_{n,\ i} = 0$，我们得到

$$(e^i)^\top \widehat{\mathbf{H}} e^j = 0.$$

因此 $\widehat{\mathbf{H}} = \mathbf{0}_{n \times n}$。所以，(1) 的唯一解是平凡解。

情形 II: $\overline{T} = U_0$. 该情形的证明与情形 I 的证明类似 . □

由前一个命题可知，当使用GSH的极小确定集时，GSH 等于 [17, 第9 节所述的单纯形海森矩阵。5]（另见 [21, 第 3] 节）。因此，在 [21, 17] 中获得的与单纯形海森矩阵相关的结果适用于GSH。在这种情况下，GSH（以及单纯形海森矩阵）等于通过采样点 $\mathcal{M}$（x^0； S, U_ℓ）的二次插值模型的海森矩阵。因此，在 [17, 16, 21] 中提出的与单纯形海森矩阵和二次插值模型的海森矩阵相关的结果对于 GSH 也是有效的。与计算二次插值模型的海森矩阵相比，计算GSH的一个优点已在备注 2.8中讨论。与 [17, 21] 中所述的单纯形海森矩阵相比，GSH 的另一个优点是，只要所采用的方向矩阵非空，它就能提供一个定义良好的简单显式公式。因此，对于定义2.9中定义的所有可能情况，GSH 的公式都提供了完整海森矩阵或部分海森矩阵的近似。

接下来，我们给出一个例子，以说明：实数集 $\mathbb{R}^n$ 中的一组 $(n+1)(n+2)/2$ 个互异点，即使能够进行二次插值，也不一定是 GSH 的极小确定集。

例 5.8。设 $x^0 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top$ 为关注点。考虑 $\mathcal{X} = \{x^0 = \mathbf{0},\, e^1,\, e^2,\, -e^1,\, -e^2, -e^1 - e^2\}$. 那么， $\mathcal{X}$ 能够进行二次插值，但不能表示为 x^0 处 GSH 的极小确定集。

证明。采用与命题 5.7 类似的方法，可以验证 $\mathcal{X}$ 是二次插值的确定集。现在我们通过穷举法证明， $\mathcal{X}$ 不是 GSH 在 x^0 处的极小确定集。我们需要构造 $S = \{s^1, s^2\}$，使得对应于 S 的矩阵满秩，且 $x^0 \oplus S \subseteq \mathcal{X}$。因此， S 的可能取值为

$$S \in \big\{ \{e^1, e^2\}, \{e^1, -e^2\}, \{e^1, -e^1 - e^2\}, \{-e^1, e^2\}, \\ \{-e^1 - e^2, e^2\}, \{-e^1, -e^2\}, \{-e^1, -e^1 - e^2\}, \{-e^1 - e^2, -e^2\} \big\}.$$

情形 I: $S = \{e^1, e^2\}$. 在这种情形下，我们需要构造 $\overline{T} = \{t^1, t^2\}$，使得对应于 $\overline{T}$ 的矩阵满秩，且

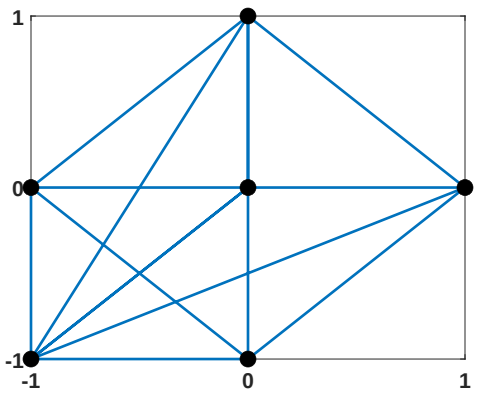
$$\{t^1, e^1 + t^1, e^2 + t^1, t^2, e^1 + t^2, e^2 + t^2\} = \mathcal{X}.$$

我们看到，使得 $\{t^1, e^1 + t^1, e^2 + t^1\} \subseteq \mathcal{X}$ 成立的 t^1 唯一可能的取值是 $t^1 = -e^1 - e^2$ （注意 $t^1 \neq \mathbf{0}$，因为我们要求满秩）。然而，使得 $\{t^2, e^1 + t^2, e^2 + t^2\} \subseteq \mathcal{X}$ 成立的 t^2 唯一可能的取值是 $t^2 = -e^1 - e^2$。由于满秩意味着 t^1 不能等于 t^2，我们看到 $S = \{e^1, e^2\}$ 无法满足所需性质。

情况 II 至 VIII： S 的其他选项也可以用类似的方法排除。
因此， $\mathcal{X}$ 不能表示为 x^0 处 GSH 的极小确定集。 □

图2展示了连接集合 $\mathcal{X}$ 中两点的所有可能方向，该集合来自示例 5.8。如果 $\mathcal{X}$ 是在 $x^0 = \mathbf{0}$ 处针对 GSH 的极小确定集，那么就可以选择两条从 x^0 发出的方向（蓝色线条），将其连接到 $\mathcal{Y}$ 中的其他点，并且同样的两条方向也会从另外两个点发出。

图 2：一个适用于二次插值的集合，但不是 x^0处 GSH 的极小确定集。



本节最后给出公式，以获得 f 在形如$\mathcal{M}(x^0;\ S, U_\ell)$ 的 GSH 的极小确定集上的二次插值函数中所涉及的所有系数值，我们将其记为 $Q_f(x^0;\ S, U_\ell)(x)$。

命题 5.9。设 f: dom $f \subseteq \mathbb{R}^n \to \mathbb{R}$。设 $\mathcal{M}(x^0;\ S, U_\ell) \subset \mathrm{dom}\, f$ 是按照命题 5.3 构造的、针对 x^0 处 GSH 的极小确定集。若 $\ell \in \{1, 2, \ldots, n\}$，则二次插值函数 $Q_f(x^0;\ S, U_\ell)(x)$ 的海森矩阵 $\mathbf{H}$ 由 $\mathbf{H} = S^{-\top}\widehat{\mathbf{H}}S^{-1}$ 给出，其中对称矩阵 $\mathbf{H} \in \mathbb{R}^{n\times n}$ 的元素$\widehat{}$为

$$
\begin{aligned}
\widehat{\mathbf{H}}_{i,\ell} &= -f(x^0 + s^i - s^\ell) + f(x^0 + s^i) + f(x^0 - s^\ell) - f(x^0), \quad i \in \{1, 2, \ldots, n\} \setminus \{\ell\},\\
\widehat{\mathbf{H}}_{i,i} &= f(x^0 + 2s^i - s^\ell) - 2f(x^0 + s^i - s^\ell) + f(x^0 - s^\ell), \quad i \in \{1, 2, \ldots, n\} \setminus \{\ell\},\\
\widehat{\mathbf{H}}_{\ell,\ell} &= f(x^0 + s^\ell) + f(x^0 - s^\ell) - 2f(x^0),\\
\widehat{\mathbf{H}}_{i,j} &= f(x^0 + s^i + s^j - s^\ell) - f(x^0 + s^i - s^\ell) - f(x^0 + s^j - s^\ell) + f(x^0 - s^\ell),
\end{aligned}
$$

对所有 $i, j \in \{1, 2, \ldots, n\} \setminus \{\ell\}, i \neq j.$，若 $\ell = 0$，则

$$
\begin{aligned}
\widehat{\mathbf{H}}_{i,i} &= f(x^0 + 2s^i) - 2f(x^0 + s^i) + f(x^0), \quad i \in \{1, 2, \ldots, n\},\\
\widehat{\mathbf{H}}_{i,j} &= f(x^0 + s^i + s^j) - f(x^0 + s^i) - f(x^0 + s^j) + f(x^0), \quad i, j \in \{1, 2, \ldots, n\}, i \neq j.
\end{aligned}
$$

对所有 $\ell \in \{0, 1, \ldots, n\}$，与 $Q_f(x^0;\ S, U_\ell)(x)$ 相关联的向量 α 由 $\alpha = S^{-\top}\overline{\alpha}$ 给出，其中 $\overline{\alpha} \in \mathbb{R}^n$ 的各项为

$$
\overline{\alpha}_i = f(x^0 + s^i) - f(x^0) - \frac{1}{2}\widehat{\mathbf{H}}_{i,i} - (x^0)^\top \mathbf{H} s^i, \quad i \in \{1, 2, \ldots, n\}.
$$

$Q_f(x^0;\ S, U_\ell)(x)$ 的标量 α_0 为

$$
\alpha_0 = f(x^0) - \alpha^\top x^0 - \frac{1}{2}(x^0)^\top \mathbf{H} x^0.
$$

证明。利用定义 2.2 可得该结果。令

$$
Q_f(x^0; S, U_\ell)(x) = \alpha_0 + \alpha^\top x + \frac{1}{2} x^\top \mathbf{H} x,
$$

其中 $a_0 \in \mathbb{R}$、$\alpha \in \mathbb{R}^n$ 和 $\mathbf{H} = \mathbf{H}^\top \in \mathbb{R}^{n\times n}$.。假设 $\ell \in \{1, 2, \ldots, n\}$.。在 x^0 处计算 $Q_f(x^0;\ S, U_\ell)(x)$，得到

$$
\alpha_0 + \alpha^\top x^0 + \frac{1}{2}(x^0)^\top \mathbf{H} x^0 = f(x^0). \tag{27}
$$

在 $x^0 + s^i$ 处计算 $Q_f(x^0;\ S, U_\ell)(x)$，并利用 (27)，得到

$$
f(x^0) + \overline{\alpha}^\top e^i + (x^0)^\top \overline{\mathbf{H}} e^i + \frac{1}{2}\widehat{\mathbf{H}}_{i,i} = f(x^0 + s^i), \quad i \in \{1, \ldots, n\}, \tag{28}
$$

其中 $\overline{\alpha}^\top = \alpha^\top S$，$\overline{\mathbf{H}} = \mathbf{H}S$ 且 $\widehat{\mathbf{H}} = S^\top \mathbf{H} S.$ 在 $x^0 + s^i - s^\ell$ 处计算 $Q_f(x^0;\ S, U_\ell)(x)$，并利用 (27) 和 (28)，得到

$$
f(x^0 + s^i) - \overline{\alpha}^\top s^\ell - (x^0)^\top \overline{\mathbf{H}} e^\ell - \widehat{\mathbf{H}}_{i,\ell} + \frac{1}{2}\widehat{\mathbf{H}}_{\ell,\ell} = f(x^0 + s^i - s^\ell), i \in \{1, \ldots, n\} \setminus \{\ell\}. \tag{29}
$$

在 $x^0 - s^\ell$ 处计算 $Q_f(x^0;\ S, U_\ell)(x)$，并利用 (27)，可得

$$-\overline{\alpha}^\top e^\ell - (x^0)^\top \overline{\mathbf{H}} e^\ell + \frac{1}{2}\widehat{\mathbf{H}}_{\ell,\ell} = f(x^0 - s^\ell) - f(x^0). \tag{30}$$

将 (30) 代入 (29)，得到

$$\widehat{\mathrm{H}}_{i,\ \ell} = \widehat{\mathrm{H}}_{\ell,i} = -f(x^0 + s^i - s^\ell) + f(x^0 + s^i) + f(x^0 - s^\ell) - f(x^0), i \in \{1, \ldots, n\} \setminus \{\ell\}. \tag{31}$$

在 $x^0 + 2s^i - s^\ell$ 处计算 $Q_f(x^0;\ S, U_\ell)(x)$，并利用 (27)、(28) 和 (30)，我们得到

$$2f(x^0 + s^i) - 2f(x^0) + f(x^0 - s^\ell) + \widehat{\mathbf{H}}_{i,i} - 2\widehat{\mathbf{H}}_{i,\ell} = f(x^0 + 2s^i - s^\ell), i \in \{1, \ldots, n\} \setminus \{\ell\}. \tag{32}$$

将（31）代入（32），可得

$$\widehat{\mathbf{H}}_{i,i} = f(x^0 + 2s^i - s^\ell) + f(x^0 - s^\ell) - 2f(x^0 + s^i - s^\ell), \quad i \in \{1, 2, \ldots, n\} \setminus \{\ell\}.$$

在 $x^0 + s^i + s^j - s^\ell$ 处计算 $Q_f(x^0;\ S, U_\ell)(x)$，并利用 (27)、(28) 和 (30)，我们得到

$$\widehat{\mathbf{H}}_{i,j} = \widehat{\mathbf{H}}_{j,i} = f(x^0 + s^i + s^j - s^\ell) - f(x^0 + s^i - s^\ell) - f(x^0 + s^j - s^\ell) + f(x^0 - s^\ell),$$

对于 $i, j \in \{1, 2, \ldots, n\} \setminus \{\ell\}, i \neq j.$，重新整理 (30)，得到

$$\frac{1}{2}\widehat{\mathbf{H}}_{\ell,\ell} - f(x^0 - s^\ell) + f(x^0) = \overline{\alpha}^\top e^\ell + (x^0)^\top \overline{\mathbf{H}} e^\ell. \tag{33}$$

将 (33) 代入 (28) 中的 $i = \ell$，得到

$$\widehat{\mathbf{H}}_{\ell,\ell} = f(x^0 + s^\ell) + f(x^0 - s^\ell) - 2f(x^0).$$

通过在（28）中将 $\overline{\alpha}^\top e^i$ 单独列出，可以求得向量 $\overline{\alpha}$ 的各个分量。我们得到

$$\overline{\alpha}_i = f(x^0 + s^i) - f(x^0) - \frac{1}{2}\widehat{\mathbf{H}}_{i,i} - (x^0)^\top \mathbf{H} s^i, \quad i \in \{1, 2, \ldots, n\}$$

其中 $\alpha_i = S^{-\top}\overline{\alpha}_i$. 最后，标量 α_0 由（27）得到。我们发现

$$\alpha_0 = f(x^0) - \alpha^\top x^0 - \frac{1}{2}(x^0)^\top \mathbf{H} x^0.$$

若 $\ell = 0$，可采用类似过程来得到 $\mathbf{H}$ α 和 α_0. □

使用 $\overline{T} = U_\ell$，值得强调的是，只要已经计算出 $\nabla_s^2 f(x^0;\ S, \overline{T})$，就可以免费获得 $Q_f(x^0;\ S, \overline{T})(x)$，而无需额外进行函数求值。事实上，计算 $Q_f(x^0;\ S, \overline{T})$ 的所有系数时，使用的函数求值与求得 $\nabla_s^2 f(x^0;\ S, \overline{T})$ 时使用的相同。由于 $\mathcal{M}(x^0;\ S, \overline{T})$ 适用于二次插值，因此存在唯一一个海森矩阵 H。由此可知，二次插值函数 $Q_f(x^0;\ S, \overline{T})(x)$ 的海森矩阵 $\mathbf{H}$ 必须等于 $\nabla_s^2 f(x^0;\ S, \overline{T})$。

6 结论

我们介绍了广义单纯形海森矩阵（GSH），这是一种用于近似函数 f 的紧凑而简单的公式。只要矩阵 S andT_i 非空，GSH就有明确定义。特别地，矩阵 S 和 T_i 不需要是满行秩矩阵，也不需要是方阵，即可计算GSH。我们还介绍了GSH 的一个中心化版本，称为广义中心单纯形 Hessian（GCSH），并展示了其与 GSH 的关系（第 3 节）。

在定理4.2 和4.3中，我们为 GSH和 GCSH建立了误差界。在一些假设下，GSH 是部分海森矩阵 Proj$_{S,T_{1:m}}$ $\nabla^2 f(x^0)$ 的一阶精确近似，而 GCSH 是部分海森矩阵 Proj$_{S,T_{1:m}}$ $\nabla^2 f(x^0)$的二阶精确近似。事实上，当集合 $T_{1:m}$ 中的所有方向矩阵都相等时，我们得到的误差界涵盖了定义 2.9中定义的全部 16 种可能情况。当所有矩阵 T_i 不相等时，则涵盖了 16种情况中的12种。未来研究方向包括研究定理 4.2和 4.3 未涵盖的情况：四种 S-不定 \超定和 $T_{1:m}$-不定\欠定。

我们注意到，当矩阵不定或超定时，可以删除矩阵的列（或行），使其达到满秩。这可能有助于推导不定情形的误差界。

由于本文的结果是以阶-N 精度表示的，因此可以直接应用 [13] 中发展的无导数优化微积分理论。在 [13], 中，基于近似的乘积、商和链式法则得到了发展。对于每条法则，文中均表明，当目标函数可以分解为更简单函数的乘积、商或复合时，可以构造这些简单函数的海森矩阵近似，再据此构造目标函数的海森矩阵近似。将 [13] 与本文的研究相结合，便得到了一种采用基于微积分的近似技术 [26]的基于模型的信赖域算法。

在第 5 节中，我们证明，如果对 GSH 使用极小确定集，计算 GSH 所需的不同函数求值次数为$(n+1)(n+2)/2.$。我们研究了 GSH 的极小确定集与二次插值的确定性之间的关系，并证明，形式为 $\mathcal{M}(x^0;\ S, U_\ell)$的 GSH 极小确定集是二次插值的良态集，但反过来未必成立。在这种情况下，我们知道， $\nabla_s^2 f(x^0;\ S, \overline{T})$在$\overline{T} = U_\ell$时等于二次插值函数 $Q_f(x^0;\ S, U_\ell)$的海森矩阵。我们还推导出了显式公式，用于求出f在形式为 $\mathcal{M}(x^0;\ S, U_\ell)$的 GSH 极小确定集上的二次插值函数参数。

未来的研究应探讨命题 5.3 中定义的集合 $U_0, U_1, \ldots, U_n$是否是满足以下条件的 $\overline{T} \in \mathbb{R}^{n\times n}$ 的唯一可能选择： $\mathcal{S}(x^0;\ S, \overline{T})$ 中的 $S \in \mathbb{R}^{n\times n}$ 满秩，且该集合是 GSH 在 x^0 处的极小确定集。通过在 $\mathbb{R}$ 和 $\mathbb{R}^2$ 中使用穷举法，可以证明情况确实如此，但目前仍不清楚如何将这一结论推广到任意维度 n。

显然，GSH 的极小确定集可以用来构造 GCSH 的一个“较小”集合（利用命题 3.3），但目前尚不清楚这样的集合是否也是极小的。未来的研究将探讨 GCSH 的极小确定集的性质，以及可用于减少采样点数量的方向矩阵。

众所周知，在计算中心单纯形梯度时，可以通过函数求值免费获得海森矩阵的对角线元素。因此，未来研究方向可以是验证：在 GCSH 是

已计算。

最后需要指出的是，如有需要，可索取 GSH 和 GCSH 的 MATLAB 实现。

致谢

作者衷心感谢匿名审稿人所作的全面评审，这些评审提升了本文的质量。

Hare 的研究部分由加拿大自然科学与工程研究理事会（NSERC）探索计划资助，资助编号为 #2018-03865。Jarry-Bolduc 的研究也部分由加拿大自然科学与工程研究理事会（NSERC）探索计划资助，资助编号为 #2018-03865。Jarry-Bolduc 谨此感谢 UBC 通过大学研究生奖学金提供的资助。

参考文献

[1] N. Amaioua、C. Audet、A. Conn 和 S. Le Digabel。在网格自适应直接搜索算法中高效求解带二次约束的二次子问题。European Journal of Operational Research，268(1)：13–24，2018。[2] C. Audet、V. B´echard 和 S. Le Digabel。通过网格自适应直接搜索和变邻域搜索进行非光滑优化。Journal of Global Optimization，41(2)：299–318，2008。[3] C. Audet 和 W. Hare。无导数与黑箱优化。Springer，2017。[4] C. Audet 和 W. Hare。从方向导数算法构造次微分。集值与变分分析，26(3)：431–447，2018。[5] C. Audet、A. Ianni、S. Le Digabel 和 C. Tribes。减少网格自适应直接搜索算法中的函数求值次数。SIAM 优化杂志，24(2)：621–642，2014。[6] C. Audet 和 C. Tribes。用于不等式约束优化的基于网格的 Nelder–Mead 算法。计算优化与应用，71(2)：331–352，2018。[7] A. Bagirov、B. Karas¨ozen 和 M. Sezer。离散梯度方法：用于非光滑优化的无导数方法。Journal of Optimization Theory and Applications，137(2)：317–334，2008。[8] A. Berahas、R. Byrd 和 J. Nocedal。通过拟牛顿方法对含噪函数进行无导数优化。SIAM 优化杂志，2019。即将出版。[9]F. Berghen 和 H. Bersini。CONDOR：Powell 的 UOBYQA 算法的一种新的并行约束扩展：实验结果及其与 DFO 算法的比较。计算与应用数学杂志，181(1)：157–175，2005。

[10] D. Bortz 和 C. Kelley。单纯形梯度与含噪优化问题。载于《最优设计与控制的计算方法》，第 77–90 页。Springer，1998。[11] P. Braun、W. Hare 和 G. Jarry-Bolduc。当固定区间 r 中的点数趋于无穷时，导数近似技术的极限行为。计算与应用数学杂志，386：113218，2021。[12] R. Burden、J. Faires 和 A. Burden。数值分析第 10 版。Brooks/Cole Cengage Learning，2016。[13] Y. Chen、W. Hare 和 G. Jarry-Bolduc。通过对子模型进行运算构造的代理模型的误差分析。运筹学数学，2022。[14]G. Cocchi、G. Liuzzi、A. Papini 和 M. Sciandrone。带盒约束的无导数多目标优化的隐式滤波算法。计算优化与应用，69(2)：267–296，2018。[15] A. Conn、K. Scheinberg 和 L. Vicente。无导数优化中插值集的几何性质。数学规划，111(1-2)：141–172，2008。[16] A. Conn、K. Scheinberg 和 L. Vicente。无导数优化中样本集的几何性质：多项式回归与欠定插值。IMA 数值分析杂志，28(4)：721–748，2008。[17] A. Conn、K. Scheinberg 和 L. Vicente。无导数优化导论。SIAM，2009。[18] I. Coope 和 R. Tappenden。正则单纯形梯度的高效计算。计算优化与应用，72(3)：561–588，2019。[19] I. Coope 和 R. Tappenden。利用二次插值模型和对齐的正则基近似梯度与对角海森矩阵。数值算法，88：767–791，2021。[20] A. Cust´odio、J. Dennis 和 L. Vicente。在直接搜索方法中使用非光滑函数的单纯形梯度。IMA 数值分析杂志，28(4)：770–784，2008。[21] A. Cust´odio 和 L. Vicente。在模式搜索方法中使用采样和单纯形导数。SIAM 优化杂志，18(2)：537–555，2007。[22] S. Gratton、C. Royer 和 L. Vicente。用于非凸非线性无约束优化的解耦的一阶/二阶步长技术及改进的复杂度界。数学规划，179：195–222，2020。[23] S. Gratton、C. Royer、L. Vicente 和 Z. Zhang。基于概率可行下降、用于有界约束与线性约束问题的直接搜索。计算优化与应用，72(3)：525–559，2019。[24] W. Hare。关于无导数优化中变分分析的讨论。集值与变分分析，第 1–17 页，2020。

[25] W. Hare 和 G. Jarry-Bolduc。广义单纯形梯度的微积分恒等式：规则与应用。SIAM 优化杂志，30(1):853–884，2020。[26] W. Hare 和 G. Jarry-Bolduc。关于基于微积分的方法在无导数信赖域算法中构建模型函数的性能。Algorithms，16(2)，2023。[27] W. Hare、G. Jarry-Bolduc 和 C. Planiden。固定形状上点数趋于无穷时广义单纯形梯度的极限行为。集值与变分分析，31(1):1–33，2023。[28] W. Hare、G. Jarry–Bolduc 和 C. Planiden。超定与欠定广义中心单纯形梯度的误差界。IMA 数值分析杂志，42(1):744–770，2020年12月。[29] W. Hare 和 J. Nutini。有限极小极大问题的无导数近似梯度采样算法。计算优化与应用，56(1):1–38，2013。[30] W. Hare、C. Planiden 和 C. Sagastiz´abal。凸有限极大问题的无导数 VU 算法。Optimization Methods and Software，35(3):521–559，2020。[31] W. Hare、C. Sagastiz´abal 和 M. Solodov。基于近似信息的非光滑非凸函数近端束方法。计算优化与应用，63(1):1–28，2016。[32] G. Jarry-Bolduc。海森矩阵对角线的近似：应使用哪组采样点。数值算法，91(3):1349–1361，2022。[33] J. Larson、M. Menickelly 和 S. Wild。非凸优化的流形采样。SIAM 优化杂志，26(4):2540–2563，2016。[34] G. Liuzzi、S. Lucidi、F. Rinaldi 和 L. Vicente。非光滑黑箱函数无导数优化的信赖域方法。SIAM 优化杂志，29(4):3012–3035，2019。[35] A. Maggiar、A. W¨achter、I. Dolinskaya 和 J. Staum。采用自适应多重重要性采样，通过高斯卷积平滑函数的无导数信赖域优化算法。SIAM 优化杂志，28(2):1478–1507，2018。[36] M. Menickelly 和 S. Wild。通过外逼近进行无导数鲁棒优化。数学规划，2019。即将发表。[37] J. Nocedal 和 S. Wright。数值优化。Springer Science & Business Media，2006。[38] M. Powell。关于由插值定义的二次模型的拉格朗日函数。Optimization Methods and Software，16(1-4):289–309，2001。[39] M. Powell。关于无导数无约束最小化的信赖域方法。数学规划，97(3):605–623，2003。

[40] M. Powell. 用于无导数有界约束优化的 BOBYQA 算法。剑桥 NA 报告 NA2009/06，剑桥大学，剑桥，第 26–46 页，2009 年。[41] R. Regis. 单纯形梯度的微积分。Optimization Letters，9(5):845–865，2015 年。[42] R. Rockafellar 和 R. Wets. 变分分析。数学科学基础丛书 [数学科学基本原理]。Springer-Verlag，柏林，1998 年。[43] S. Shashaani、F. Hashemi 和 R. Pasupathy. ASTRO-DF：一类用于无导数随机优化的自适应采样信赖域算法。SIAM 优化杂志，28(4):3145–3176，2018 年。[44] A. Verd´erio、E. Karas、L. Pedroso 和 K. Scheinberg. 无导数信赖域算法中二次模型的构造。EURO Journal on Computational Optimization，5(4):501–527，2017 年。[45] S. Wild 和 C. Shoemaker. 无导数优化径向基函数信赖域算法的全局收敛性。SIAM REVIEW，55(2):349–371，2013 年。