arXiv:2601.02650v1 [math.NA] 6 Jan 2026

一种具有线性收敛速率的无导数鞍点搜索算法*

杜强†，史宝明‡，张磊§，和郑祥成¶

摘要。我们提出一种无导数鞍点搜索算法，仅通过函数求值来定位过渡态。该算法采用嵌套架构，由内层特征向量搜索和外层鞍点搜索组成。通过严格的数值分析，我们证明了在适当假设下，内层步骤几乎必然收敛。此外，我们证明了采用递减步长的外层搜索能够收敛，并在固定步长和有界性条件下证明了其线性收敛。我们还给出数值实验，以验证理论结果并展示该算法的实际适用性。

关键词。鞍点、无导数算法、几乎必然收敛

美国数学学会学科分类。65C20, 65K10,65L20,37N30

1. 引言。自然科学和工程学中的许多问题反复呈现这样一种情形：定义在系统物理变量上的能量（自由能或目标函数）具有复杂的景观，其中分布着大量局部极小值和鞍点。例子包括相变、[18, 28], 化学反应 [11], 蛋白质折叠 [37, 46], 以及深度神经网络的损失景观 [17, 26]。局部极小值对应稳定态，而鞍点则代表过渡态或激发态，在决定极小值之间的过渡路径方面发挥着关键作用 [52] ，并揭示能量景观的层级结构 [47]。这些特征使得鞍点计算在过去几十年中受到广泛关注。

与识别稳定极小值相比，鞍点的计算要困难得多，这是因为其稳定性显然较差。此外，通常事先并不知道鞍点的不稳定方向。人们已开发出大量用于计算鞍点的数值算法，这些算法通常分为两类：路径搜索方法 [9, 10]和曲面行走方法 [52]。我们重点关注后者，这类方法通常利用一阶或二阶导数，在能量曲面上迭代，逐步逼近鞍点。这类方法的例子包括最温和上升动力学[12, 38, 30] 和高指标鞍点动力学 [48, 44], ，它们需要计算目标函数的梯度（和海森矩阵）。二聚体方法和收缩二聚体动力学 [20, 50, 51]采用两点梯度差，在一维空间中近似投影海森矩阵（该空间由二聚体方向确定）。

*本工作部分得到美国国家科学基金会 DMS-2309245（资助 QD 和 BS）；国家重点研发计划 2024YFA0919500，以及国家自然科学基金委员会 12225102、T2321001、12288101 和 12426653（资助 LZ）；山东省自然科学基金（ZR2025QB01）、国家自然科学基金委员会（12301555）、国家重点研发计划（2023YFA1008903）和山东省泰山学者计划（tsqn202306083）（资助 XZ）的支持。作者姓名按字母顺序排列。

†哥伦比亚大学应用物理与应用数学系和数据科学研究院，美国纽约州纽约市 10027（qd2125@columbia.edu）。

‡通讯作者。哥伦比亚大学应用物理与应用数学系，美国纽约州纽约市 10027（bs3705@columbia.edu）。§北京大学北京国际数学研究中心、定量生物学中心和机器学习中心，中国北京 100871（zhangl@math.pku.edu.cn）。

¶数学学院、密码学与数字经济安全全国重点实验室，山东大学，济南，250100，中国（xzheng@sdu.edu.cn）。

本手稿仅供审稿使用。

虽然这种方法避免了显式使用海森矩阵，但仍需要能量的梯度。因此，由于依赖导数信息，这些方法无法直接用于可能无法获得或难以获取此类信息的问题，例如分布式学习 [27]，分子模拟中的自由能面 [33]，（它并非原子坐标的显式函数，而是对微观态进行统计平均的结果）、机器学习中的超参数调优 [43]，以及科学问题中的经验目标函数 [8, 19]。关于其他无法获得或难以获取一阶及高阶导数的实际问题，请参阅 [2, 39] 。因此，开发用于计算鞍点的无导数算法具有重要意义。

零阶算法仅依赖函数求值，适用于上述问题，也称为无导数方法或黑箱方法。例如，（随机）零阶算法已在优化问题中得到广泛发展和应用 [6, 25]。然而，将零阶方法扩展到鞍点计算的研究却少之又少。与零阶优化算法相比，零阶鞍点搜索算法在算法设计和收敛性分析方面都面临额外挑战。从算法设计的角度来看，零阶鞍点搜索算法还应考虑不稳定方向的更新（即海森矩阵中与负特征值对应的特征向量）；这些方向是与构型空间中的迭代点相关的未知量。这意味着需要根据零阶信息重构二阶信息，这可能导致维数灾难 [21, 35]。从收敛性分析的角度来看，我们需要处理鞍点搜索算法中构型和不稳定子空间的迭代，而包含鞍点的目标函数必然是非凸的，这可能会给收敛性证明带来挑战。本文的主要贡献是推导出一种零阶鞍点搜索算法并对其进行收敛性分析，该算法适用于许多实际问题。值得注意的是，我们的理论和数值结果均表明，所提出的采用固定步长的零阶算法，其迭代点以线性收敛速率趋近鞍点；这一速度与传统的（离散化）鞍点动力学相当，同时仅需进行函数求值。

本文结构安排如下。在第 2 节中，我们回顾鞍点的概念和鞍点动力学。在第 3 节中，我们首先介绍零阶特征向量搜索算法，随后介绍零阶鞍点搜索算法。我们在第 4 节中对它们进行收敛性分析。第 5 节通过若干基准测试和实际问题的数值实验，验证理论结果。最后，我们在第 6 节中给出结论并进行讨论。

2. 鞍点与鞍点动力学。考虑一个由光滑能量函数 $f(\mathbf{x}): \mathbb{R}^d \to \mathbb{R}$ 刻画的系统，其中 $d \geqslant 2$，我们假设 f 是C^6 连续的。临界点 $\mathbf{x}^*$ 定义为梯度为零的构型，即 $\nabla f(\mathbf{x}^*) = 0$。可根据海森矩阵 $\nabla^2 f(\mathbf{x}^*)$ 判断 $\mathbf{x}^*$ 的局部稳定性。若 $\nabla^2 f(\mathbf{x}^*)$ 正定，则 $\mathbf{x}^*$是一个局部极小值 [36]；若$\nabla^2 f(\mathbf{x}^*)$ 有 $k \leqslant d-1$ 个负特征值及其对应的 k 个特征向量，记作 $\{\mathbf{v}_1^*, \cdots, \mathbf{v}_k^*\}$，则 $\mathbf{x}^*$ 被归类为指标为 k 的鞍点，而 $\{\mathbf{v}_1^*, \cdots, \mathbf{v}_k^*\}$在本文中称为“不稳定特征向量”或“不稳定方向”。

这里，我们考虑一种用于搜索

指标-k 鞍点 $\mathbf{x}^*$ 的特定版本，其定义为

$$\dot{\mathbf{x}} = -\left(\mathbf{I} - 2\sum_{i=1}^{k}\mathbf{v}_i\mathbf{v}_i^\top\right)\nabla f(\mathbf{x}), \tag{2.1}$$

其中 $\{\mathbf{v}_i\}_{i=1}^k$，$1 \leqslant k \leqslant d-1$ 是对应于k 的最小特征值的归一化特征基 $\nabla^2 f(\mathbf{x})$，而 $\mathbf{I}$ 是恒等算子。与沿负梯度方向演化并使目标函数值下降的梯度流不同，鞍点动力学的右端沿不稳定方向对负梯度应用豪斯霍尔德变换。因此，能量沿不稳定方向增加，同时沿稳定方向减少，从而使动力学能够找到鞍点。式 $\{\mathbf{v}_i\}_{i=1}^k$ （2.1）中的特征基可由特征值求解器给出，该求解器作用于 $\nabla^2 f(\mathbf{x})$ 并生成不稳定方向。

对鞍点动力学（2.1）采用最简单的时间离散化，可得到以下离散动力学：给定$\mathbf{x(0)}$,

$$\begin{aligned}&\mathbf{x}(n+1) = \mathbf{x}(n) - \alpha(n)\left(\mathbf{I} - 2\sum_{i=1}^{k}\mathbf{v}_i(n)\mathbf{v}_i(n)^\top\right)\nabla f(\mathbf{x}(n)),\ n \geqslant 0,\\ &\text{where } \{\mathbf{v}_i(n)\}_{i=1}^k = \text{EigenSolver}(\nabla^2 f(\mathbf{x}(n))),\ \text{and } \alpha(n) \text{ is the step size.}\end{aligned} \tag{2.2}$$

3. 无导数/零阶鞍点搜索。 每一步迭代（2.2）都需要计算目标函数的梯度和海森矩阵，这在实际应用中可能不可行，或计算成本高昂 [2, 39]。我们考虑一种无导数鞍点搜索算法，它利用随机差分近似，只需进行零阶函数求值。

3.1. 导数的零阶近似。推导零阶鞍点搜索算法的一种自然方法，是在 $\nabla f(x)$ 和 $\nabla^2 f(\mathbf{x})$ 中，用零阶估计器 $F(\mathbf{x}, \mathbf{r}\ l)$ 和 $\mathbf{H}(\mathbf{x}, \mathbf{r}\ l)$ 替换（2.2）。这里，我们采用基于双点随机零阶梯度估计器的无导数形式，该估计器已广泛应用于零阶（无导数）优化方法中 [35,29, 3]。其定义如下：

$$F(\mathbf{x}, \mathbf{r}, l) = \frac{f(\mathbf{x}+l\mathbf{r}) - f(\mathbf{x}-l\mathbf{r})}{2l}\mathbf{r}, \mathbf{r} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3.1}$$

其中，$1 \gg l > 0$ 是差分长度，类似于二聚体长度； [20, 50], $\mathbf{r}$ 是从标准正态分布或单位球面上的均匀分布中抽取的独立同分布随机方向（在后一种情况下，我们需要将 F 乘以 d [40, 16]）。对于$l > 0$，尽管零阶梯度估计器会给真实梯度引入偏差，但它对于参数为 l的所谓高斯平滑函数的梯度仍然是无偏的，该函数定义为

$$f_l(\mathbf{x}) = \mathbb{E}_{\mathbf{r}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}(f(\mathbf{x}+l\mathbf{r})). \tag{3.2}$$

根据 Stein 引理，$\nabla f_l(\mathbf{x}) = \mathbb{E}_{\mathbf{r}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\nabla f(\mathbf{x}+l\mathbf{r})] = \frac{1}{l}\mathbb{E}_{\mathbf{r}\approx\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}+l\mathbf{r})\mathbf{r}]$,并且（3.1）实际上是双点采样，因此它是 $f_l(\mathbf{x})$ 的无偏梯度估计量。我们注意到，对于鞍点搜索，二聚体方法 [20, 50, 51]采用梯度差分来近似二聚体方向上的投影海森矩阵。然而，二聚体方向并非随机采样，因此这些方法仍然基于梯度。我们也可以利用（3.2）获得海森矩阵的无导数近似，

本手稿仅供审阅。

$\nabla^2 f_l(\mathbf{x}) = \frac{1}{l^2}\mathbb{E}_{\mathbf{r}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(\mathbf{r}\mathbf{r}^\top - \mathbf{I})f(\mathbf{x}+l\mathbf{r})] = \frac{1}{l^2}\mathbb{E}_{\mathbf{r}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(\mathbf{r}\mathbf{r}^\top - \mathbf{I})(f(\mathbf{x}+l\mathbf{r}) - f(\mathbf{x}))]$。因此，离散采样

$$\mathbf{H}(\mathbf{x},\mathbf{r},l) = \frac{f(\mathbf{x}+l\mathbf{r}) + f(\mathbf{x}-l\mathbf{r}) - 2f(\mathbf{x})}{2l^2}(\mathbf{r}\mathbf{r}^\top - \mathbf{I}), \mathbf{r}\sim\mathcal{N}(\mathbf{0},\mathbf{I}), \tag{3.3}$$

是 $f_l(\mathbf{x})$ 的无偏零阶海森矩阵估计量。

3.2. 零阶/无导数鞍点搜索算法。由(2.2)可知，鞍点搜索算法由两个部分组成：

- 内层迭代，即特征向量搜索算法，在每个迭代点识别不稳定方向，见 算法 3.1；以及
- 外层迭代，即鞍点搜索算法，沿内层步骤获得的不稳定方向反射零阶梯度估计器，以更新迭代点，见 算法 3.2。

我们首先考虑将梯度下降的无导数版本应用于特征向量搜索中的瑞利商，其中 $\nabla^2 f(\mathbf{x})$ 替换为零阶海森矩阵估计器 $\mathbf{H}$，其定义见(3.3)：对于 $n \geqslant 0$ 和 $k = 2, \cdots, k$,

$$\begin{aligned}\hat{\mathbf{v}}(n+1) =& \mathbf{v}(n) - \alpha(n)(\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)\mathbf{H}(\mathbf{x},\mathbf{r}(n),l(n))\mathbf{v}(n),\\ \mathbf{v}(n+1) =& \hat{\mathbf{v}}(n+1)/\|\hat{\mathbf{v}}(n+1)\|_2,\end{aligned} \tag{3.4}$$

$$\begin{aligned}\hat{\mathbf{v}}_{\bar{k}}(n+1) =& \mathbf{v}_{\bar{k}}(n) - \alpha(n)(\mathbf{I} - V_{n,k}V_{n,k}^\top)\mathbf{H}(\mathbf{x},\mathbf{r}(n),l(n))\mathbf{v}_{\bar{k}}(n),\\ \mathbf{v}_{\bar{k}}(n+1) =& \hat{\mathbf{v}}_{\bar{k}}(n+1)/\|\hat{\mathbf{v}}_{\bar{k}}(n+1)\|_2,\end{aligned} \tag{3.5}$$

其中 $V_{n,\,k} = [\mathbf{v}(n)、\mathbf{v}_1、\ldots、\mathbf{v}_{\bar{k}-1}]$，且步长 $\alpha(n) > 0$除非另有说明，均满足以下通常假设。

Assumption 3.1 (Diminishing step size). $\sum_{n=0}^{\infty}\alpha(n) = \infty$, *and* $\sum_{n=0}^{\infty}\alpha(n)^2 < \infty$.

然而，在第4.2小节中，我们将证明$\mathbf{H}(\mathbf{x},\ \mathbf{r}\ l)$ 的变化幅度会受到维数灾难的影响，进而减慢收敛速度。在(3.4)和(3.5)中，采用近似海森矩阵-向量积 [50]即可。也就是说，

$$H_{\mathbf{v}}(\mathbf{x},\mathbf{r},l) = \frac{F(\mathbf{x}+l\mathbf{v},\mathbf{r},l) - F(\mathbf{x}-l\mathbf{v},\mathbf{r},l)}{2l} \tag{3.6}$$

可用于替代$\mathbf{H}(\mathbf{x},\ \mathbf{r}\ l)\mathbf{v}$，从而降低变化幅度。由于 $F(\mathbf{x},\ \mathbf{r}\ l)$、$\mathbf{H}(\mathbf{x},\ \mathbf{r}\ l)\mathbf{v}$ 和 $\mathbf{H}_{\mathbf{v}}(\mathbf{x},\ \mathbf{r}\ l)$ 都是 $\nabla f(\mathbf{x})$ 或 $\nabla^2 f(\mathbf{x})\mathbf{v}$ 的有偏估计量，因此有时需要使用以下递减差分长度 $l(n)$来控制偏差误差。

Assumption 3.2. 存在一个常数 $L > 0$，使得 $l(n) \leqslant L\sqrt{\alpha(n)}$。

基于离散动力学（3.4）和（3.5），以及零阶海森矩阵-向量乘积估计器 $H_{\mathbf{v}}(\mathbf{x},\mathbf{r}\ l)$，我们在 算法 3.1 中提出了零阶特征向量搜索算法。借助 算法 3.1，我们提出 算法 3.2 来实现无导数鞍点搜索的外层迭代，随后对其进行收敛性分析。在继续讨论算法的收敛性之前，我们将在本节余下部分给出一些假设和技术结果。

3.3. 关于能量的假设。为简洁起见，我们考虑一个满足以下正则性假设的目标函数 $f(\mathbf{x})$。

本手稿仅供审阅。

算法 3.1用于 k 不稳定方向的无导数特征向量搜索算法

初始条件: $n \leftarrow 0$, $\mathbf{v}_i, i = 1, \cdots, k$
对于 $n \leqslant n_{\mathbf{v}}^{max}$, **执行**
  $\mathbf{r}$ 取自 $\mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{v}_1 \leftarrow \mathbf{v}_1 - \alpha_{\mathbf{v}}(n)(\mathbf{I} - \mathbf{v}_1\mathbf{v}_1^\top)H_{\mathbf{v}_1}(\mathbf{x}, \mathbf{r}, l(n))$
  $\mathbf{v}_1 \leftarrow \mathbf{v}_1/\|\mathbf{v}_1\|_2$, $n \leftarrow n + 1$
结束循环
for j in $\{2, \cdots, k\}$ **do**
  $n \leftarrow 0$
  $\mathbf{v}_j \leftarrow \mathbf{v}_j - (\mathbf{I} - \sum_{i=1}^{j-1} \mathbf{v}_i\mathbf{v}_i^\top)\mathbf{v}_j$, $\mathbf{v}_j \leftarrow \mathbf{v}_j/\|\mathbf{v}_j\|_2$
  for $n \leqslant n_{\mathbf{v}}^{max}$ **do**
    $\mathbf{r}$ is drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{v}_j \leftarrow \mathbf{v}_j - \alpha_{\mathbf{v}}(n)(\mathbf{I} - \mathbf{v}_j\mathbf{v}_j^\top - \sum_{i=1}^{j-1} \mathbf{v}_i\mathbf{v}_i^\top)H_{\mathbf{v}_j}(\mathbf{x}, \mathbf{r}, l(n))$
    $\mathbf{v}_j \leftarrow \mathbf{v}_j/\|\mathbf{v}_j\|_2$, $n \leftarrow n + 1$
  end for
end for

算法 3.2用于寻找指标为 k 的鞍点的无导数鞍点搜索算法

初始条件: $n \leftarrow 0$, $\mathbf{x}, \mathbf{v}_i, i = 1, \cdots, k$
对于 $n \leqslant n_{\mathbf{x}}^{max}$, **执行**
  $\mathbf{r}$ 取自 $\mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{x} \leftarrow \mathbf{x} - \alpha_{\mathbf{x}}(n)(\mathbf{I} - \sum_{i=1}^{k} 2\mathbf{v}_i\mathbf{v}_i^T)F(\mathbf{x}, \mathbf{r}, l(n))$
  搜索 k 不稳定方向 $\{\mathbf{v}_i\}_{i=}^{k}$ $_1$ 的 $\nabla^2 f(\mathbf{x})$ 使用 算法 3.1
  $n \leftarrow n + 1$
结束循环

ASSUMPTION 3.3. 假设 $f \in C^6$, 对于所有 $\mathbf{x} \in \mathbb{R}^n$, 有 $\|\nabla^i f(\mathbf{x})\|_2 \leqslant M$ 其中正常数为 $M < \infty$, 且 $i = 2,3,4,5,6$。

备注 3.1. 为了分析鞍点搜索算法, 以最简单的二次函数为例, 我们可以像 [54, 53]中那样, 对高阶导数施加一致有界性。在许多应用中, 存在满足假设 3.3的各种目标函数, 例如 Minyaev-Quapp 能量函数 [34] 和 Eckhardt 能量函数 [13]。

ASSUMPTION 3.4. 存在 $\delta > 0$ 和 $\mu \in (0, M)$, 使得 $\forall \mathbf{x} \in \{\mathbf{x} \ \|\mathbf{x} - \mathbf{x}^*\|_2 \leqslant \delta\}$, $\nabla^2 f(\mathbf{x})$ 的特征值满足

$$-M \leqslant \lambda_1 \leqslant \cdots \leqslant \lambda_k < -\mu < 0 < \mu < \lambda_{k+1} \leqslant \cdots \leqslant \lambda_d \leqslant M. \tag{3.7}$$

注 3.2. 对每个严格鞍点 $\mathbf{x}^*$, 特征值假设均成立。根据 假设 3.3, L_2-海森矩阵的范数一致有界于 M。此外, 根据 假设 3.3, 对于任意满足 $\mathbf{v}$ 条件的 $\|\mathbf{v}\|_2 = 1$, 有

$$\|(\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y}))\mathbf{v}\|_2 = \left\| \int_0^1 \nabla^3 f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))[\mathbf{x} - \mathbf{y}, \mathbf{v}] \mathrm{d}t \right\|_2 \leqslant M\|\mathbf{x} - \mathbf{y}\|_2,$$

这给出了 Lipschitz 条件 $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leqslant M\|\mathbf{x} - \mathbf{y}\|_2$, $\forall \mathbf{x}, \mathbf{y} \in U$, 其中常数 M 由以下条件给出: 假设 3.3。

本手稿仅供审阅。

3.4. 近似梯度和海森矩阵的估计。以下引理表明，在任意有界域 $\{\mathbf{x}\ \|\mathbf{x}\|_2 \leqslant C\}$ 内，零阶梯度和海森矩阵的二阶矩有界。

LEMMA 3.3（近乎无偏估计量）。根据假设3.3，我们有

$$\mathbb{E}_{\mathbf{r}}[F(\mathbf{x},\mathbf{r},l)] = \nabla f_l(\mathbf{x}), \mathbb{E}_{\mathbf{r}}[\mathbf{H}(\mathbf{x},\mathbf{r},l)] = \nabla^2 f_l(\mathbf{x}),$$

$$\mathbb{E}_{\mathbf{r}}[\|F(\mathbf{x},\mathbf{r},l)\|_2^2] \leqslant (2d+4)\|\nabla f(\mathbf{x})\|_2^2 + \frac{M^2 l^4 d(d+2)(d+4)(d+6)}{18},$$

$$\mathbb{E}_{\mathbf{r}}[\|\mathbf{H}(\mathbf{x},\mathbf{r},l)\|_2^2] \leqslant \frac{M^2 d(d+2)(d^2+12d+33)}{2} + \frac{M^2 l^4 d(d+2)(d+4)(d+6)(d^2+20d+97)}{288},$$

$$\|\nabla f(\mathbf{x}) - \nabla f_l(\mathbf{x})\|_2 \leqslant \frac{Ml^2 d}{2}, \|\nabla^2 f(\mathbf{x}) - \nabla^2 f_l(\mathbf{x})\|_2 \leqslant \frac{Ml^2 d}{2}. \tag{3.8}$$

证明。该引理可由泰勒展开及正态分布的高阶矩直接推出，证明见附录。 □

LEMMA 3.4。根据假设3.3 和假设3.4， $\nabla^2 f_l(\mathbf{x})$ 是利普希茨连续的。此外，如果 $Ml^2 d/2 \ll 1$，其特征值满足：对于 $\mu_l = \mu - Ml^2 d/2 > 0$ 以及任意$\mathbf{x} \in \{\mathbf{x}\ \|\mathbf{x} - \mathbf{x}^*\|_2 \leqslant \delta\}$,

$$-M \leqslant \lambda_1 \leqslant \cdots \leqslant \lambda_k < -\mu_l < 0 < \mu_l < \lambda_{k+1} \leqslant \cdots \leqslant \lambda_d \leqslant M. \tag{3.9}$$

证明。为验证 $\nabla^2 f_l(\mathbf{x})$的利普希茨连续性，我们计算

$$\|\nabla^2 f_l(\mathbf{x}) - \nabla^2 f_l(\mathbf{y})\|_2 = \left\|\mathbb{E}_{\mathbf{r}}\left[\nabla^2 f(\mathbf{x}+l\mathbf{r}) - \nabla^2 f(\mathbf{y}+l\mathbf{r})\right]\right\|_2 \leqslant M\|\mathbf{x}-\mathbf{y}\|_2.$$

根据假设 3.3，我们有

$$\|\nabla^i f_l(\mathbf{x})\|_2 = \|\mathbb{E}_{\mathbf{r}}[\nabla^i f(\mathbf{x}+l\mathbf{r})]\|_2 \leqslant \mathbb{E}_{\mathbf{r}}[\|\nabla^i f(\mathbf{x}+l\mathbf{r})\|_2] \leqslant M, i = 2,3,4,5,6.$$

因此，（3.9）是（3.8）的直接推论，且满足假设 3.3和假设 3.4。 □

我们通常要求差分长度 l 足够小，具体由以下假设规定。

ASSUMPTION 3.5. 差分长度 l满足 $dl \ll 1$、 $Ml^2 d/2 \ll 1$。

3.5. 一些技术结果。 我们首先引用一个关于矩阵特征值问题扰动的经典结果，该结果将用于评估海森矩阵不稳定子空间的近似。

LEMMA 3.5（戴维斯–卡汉定理 [49]）。设 Σ、$\hat{\Sigma} \in \mathbb{R}^{d\times d}$ 为对称矩阵，其特征值分别为 $\lambda_d \geqslant \cdots \geqslant \lambda_1$和 $\lambda_d \geqslant \cdots \geqslant \lambda_1$。固定 $1 \leqslant r \leqslant s \leqslant d$，并假设 $\min(\lambda_r - \lambda_{r-1},\ \lambda_{s+1} - \lambda_s) > 0$，其中我们定义 $\lambda_0 = -\infty$ 和$\lambda_{d+1} = +\infty$。设 $p = s - r + 1$，并令 $V = [v_r、v_{r+1}、\ldots\ldots、v_s] \in \mathbb{R}^{d\times p}$ 以及 $\hat{V} = [\hat{v}_r、\hat{v}_{r+1}、\ldots\ldots、\hat{v}_s] \in \mathbb{R}^{d\times p}$ 为具有正交归一列的矩阵，满足 $\Sigma v_j = \lambda_j v_j$、$\hat{\Sigma}\hat{v}_j = \lambda_j \hat{v}_j$以及 $\|v_j\|_2 = \|\hat{v}_j\|_2 = 1$，其中 $j = r、r+1、\ldots\ldots、s$。则

$$\left\|VV^\top - \hat{V}\hat{V}^\top\right\|_F \leqslant \frac{2\sqrt{2}\ \min\left(p^{1/2}\|\hat{\Sigma} - \Sigma\|_2,\ \|\hat{\Sigma} - \Sigma\|_F\right)}{\min(\lambda_r - \lambda_{r-1},\ \lambda_{s+1} - \lambda_s)}.$$

本手稿仅供审阅使用。

为了帮助我们证明鞍点搜索算法的收敛性，我们引用 [42,引理 4。3] 关于含噪递归过程的收敛性。

LEMMA 3.6. 令 $A(x): \mathbb{R}^d \to \mathbb{R}$ 为一个下有界函数， $B(x): \mathbb{R}^d \to \mathbb{R}$为一个非负函数，即对于所有 $x \in \mathbb{R}^d$， $B(x) \geqslant 0$ ；令 $\mathcal{F}(n)$表示由随机变量 $\mathcal{F}(n) = \{x_0, x1, \ldots, x_n\}$. 生成的滤子。在假设 3.1 的步长假设下，假设存在常数 C_1、 $C_2 > 0$ ，使得

$$\mathbb{E}(A(\mathbf{x}(n+1)) - A(\mathbf{x}(n))|\mathcal{F}(n)) \leqslant -C_1\alpha(n)B(\mathbf{x}(n)) + C_2\alpha(n)^2, \tag{3.10}$$

那么 $A(\mathbf{x}(n))$ 几乎处处收敛，当 $n \to \infty$ 时。如果 $B(\mathbf{x}(n))$ 也几乎处处收敛，那么它几乎处处收敛到零。

4. 无导数鞍点搜索算法的收敛性。 我们首先在 第4.1小节 研究带海森矩阵估计器的特征向量搜索算法的收敛性。我们还在第4.2小节研究维数对收敛性的影响，并在算法 3.1 中研究内部特征向量搜索算法第4.3小节。然后，基于内层迭代输出的不稳定方向，我们在第4.4小节 分析外层迭代的收敛性，即无导数鞍点搜索算法（算法 3.2）。

4.1. 使用零阶海森矩阵估计量的内层无导数特征向量搜索算法的收敛性。对于鞍点搜索，关键在于识别不稳定方向。因此，内层迭代的收敛性分析侧重于投影算子在不稳定子空间上的收敛性。我们注意到，这项分析主要沿用了 [42] 中采用的方法。

LEMMA 4.1. 在假设 3.1 和假设 3.5 下，由 (3.4) 生成的 $v(n)$ 的极限点几乎处处属于 $\nabla^2 f_l(x)$ 的特征子空间，其中 $l(n) \equiv l$ 。由 (3.5) 生成的迭代点 $v(n)$ 的极限点 $v_\infty \notin span\{v_{l,1}, \cdots, v_{l,\bar{k}-1}\}$ 几乎处处属于 $\nabla^2 f_l(x)$ 的特征子空间，前提是 $v_{l,i}, i = 1, \cdots, \bar{k}-1$ 是 $\nabla^2 f_l(x)$ 的特征向量，且 $v(0)$ 与 $v_i, i = 1, \cdots, \bar{k}-1$ 正交。

证明。 这是由引理 3.3 和 [42, 命题 4。9]。 □

备注 4.2. 考虑 Rayleigh商问题

$$\min_{\mathbf{v}} \ \mathbf{v}^\top \nabla^2 f_l(\mathbf{x})\mathbf{v} \text{ s.t. } \|\mathbf{v}\|^2 = 1, \mathbf{v}_{l,j}^T\mathbf{v} = 0, \ j = 1, \cdots, \bar{k}-1.$$

上述问题的解给出了特征向量 $\mathbf{v}_{l,\bar{k}}$ ，它对应于 $\nabla^2 f_l(\mathbf{x})$ 的第 $\bar{k}$ 小特征值，而 $\mathbf{v}_{l,i}$, $i = 1, \cdots, \bar{k}-1$ 是对应于第 i 小特征值的特征向量。若 $\lambda_{\bar{k}}$ 是单重的，则 $\mathbf{v}_{l,\bar{k}}$ 是全局最小值点，而 $\mathbf{v}_{l,i}, i > k$ 是不稳定鞍点。即使在最坏情况下，算法收敛到鞍点或对应于正特征值的特征向量，我们也可以重复进行特征向量搜索，以识别不稳定方向。因此，以下假设引理 4.1 中的特征子空间实际上由 $\nabla^2 f_l(\mathbf{x})$ 的 k 个最小特征向量张成。

分别用 $V = [\mathbf{v}_1(\mathbf{x}) \ldots, \mathbf{v}_k(\mathbf{x})]$ 和 $V_l = [\mathbf{v}_{1,l}(\mathbf{x}) \ldots, \mathbf{v}_{k,l}(\mathbf{x})]$, 表示 $\nabla^2 f(\mathbf{x})$ 和 $\nabla^2 f_l(\mathbf{x})$ 对应于 k 个最小特征值的特征向量，¯并用 $\bar{V} = [\bar{\mathbf{v}}_1, \ldots, \bar{\mathbf{v}}_k]$ 表示 (3.4) 和 (3.5) 的输出。

LEMMA 4.3. 假设假设3.1 成立。给定任意容差 $\epsilon_{\mathbf{v}} > 0$，¯$V$ 达到该容差

$$\left\|\bar{V}\bar{V}^\top - V_l V_l^\top\right\|_2 \leqslant \epsilon_{\mathbf{v}}, \tag{4.1}$$

本手稿仅供评审使用。

且几乎处处在有限次迭代内实现。

证明。 这是定理 [42, 定理 4。13]。 □

COROLLARY 4.4. 假设假设 3.1，假设 3.3和假设 3.4¯成立。若 V 满足(4.1)中的容差，那么它也是 V 的良好近似，如果差分长度 l 足够小。具体而言，

$$\left\|\bar{V}\bar{V}^{\top}-VV^{\top}\right\|_2 \leqslant \frac{\sqrt{2k}Ml^2d}{4\mu}+\epsilon_{\mathbf{v}}.$$

证明。 令 $\Sigma=\nabla^2 f(\mathbf{x})$, $\hat{}\ \Sigma=\nabla^2 f_l(\mathbf{x})$, $s=k$, $r=1$ 代入引理 3.5，可得 $\min(\lambda_{k+1}-\lambda_k, \lambda_1-\lambda_0) \geqslant 2\mu$，并且

$$\|V_lV_l^{\top}-VV^{\top}\|_2 \leqslant \left\|V_lV_l^{\top}-VV^{\top}\right\|_F \leqslant \frac{\sqrt{2k}}{2\mu}\|\nabla^2 f_l(\mathbf{x})-\nabla^2 f(\mathbf{x})\|_2 \leqslant \frac{\sqrt{2k}Ml^2d}{4\mu},$$

$$\|\bar{V}\bar{V}^{\top}-VV^{\top}\|_2 \leqslant \|\bar{V}\bar{V}^{\top}-V_lV_l^{\top}\|_2+\|V_lV_l^{\top}-VV^{\top}\|_2 \leqslant \frac{\sqrt{2k}Ml^2d}{4\mu}+\epsilon_{\mathbf{v}}.$$

4.2. 误差的维数依赖性。 需要注意的是，式（3.4）和（3.5）中的零阶海森矩阵估计量受到维数灾难的影响。忽略 $O(l^4d^6)$ 项（该项可通过足够小的 l 在引理 3.3中加以控制），可得

$$\mathbb{E}_{\mathbf{r}}[\|\mathbf{H}(\mathbf{x},\mathbf{r},l)\|_2^2] \lesssim \frac{d(d+2)(d^2+12d+33)M^2}{2}=O(d^4). \tag{4.2}$$

事实上，$\mathbf{H}(\mathbf{x},\mathbf{r},l)$ 仅包含一维曲率信息，用于近似 $d\times d$ 海森矩阵。因此，随着维度 d 增大，变化幅度迅速增加也就不足为奇了。接下来，我们证明这种较大的变化幅度会减缓收敛。为不失一般性，假设$\mathbb{E}_{\mathbf{r}}(\|\mathbf{H}(\mathbf{x},\mathbf{r},l)\|_2^2) \leqslant C_3d^4$。考虑式 (3.4) 中的迭代，并将瑞利商及其黎曼梯度的平方范数定义为 $A(\mathbf{v})=\mathbf{v}^T\nabla^2 f(\mathbf{x})\mathbf{v}$ 和$B(\mathbf{v})=\|(\mathbf{I}-\mathbf{v}\mathbf{v}^{\top})\nabla^2 f(\mathbf{x})\mathbf{v}\|_2^2$，则

$$\begin{aligned}
&A(\mathbf{v}(n+1))-A(\mathbf{v}(n))\\
&\leqslant -2\alpha(n)\langle(\mathbf{I}-\mathbf{v}(n)\mathbf{v}(n)^{\top})\mathbf{H}(\mathbf{x},\mathbf{r}(n),l(n))\mathbf{v}(n),(\mathbf{I}-\mathbf{v}(n)\mathbf{v}(n)^{\top})\nabla^2 f_l(\mathbf{x})\mathbf{v}(n)\rangle\\
&\quad+M_1\alpha(n)^2\|\mathbf{H}(\mathbf{x},\mathbf{r}(n),l(n))\|_2^2,
\end{aligned}$$

其中，M_1 是一个依赖于 M 的正常数。对两边取期望，

$$\begin{aligned}
&\mathbb{E}[A(\mathbf{v}(n+1))-A(\mathbf{v}(n))|\mathcal{F}(n)]\\
&\leqslant -2\alpha(n)B(\mathbf{v}(n))+M_1\alpha(n)^2\mathbb{E}[\|\mathbf{H}(\mathbf{x},\mathbf{r}(n),l(n))\|_2^2]\\
&\leqslant -2\alpha(n)B(\mathbf{v}(n))+M_1C_3\alpha(n)^2d^4.
\end{aligned}$$

对 $\mathcal{F}(n)$ 取期望，可得

$$\mathbb{E}[A(\mathbf{v}(n+1))]-\mathbb{E}[A(\mathbf{v}(n))] \leqslant -2\alpha(n)\mathbb{E}[B(\mathbf{v}(n))]+M_1C_3\alpha(n)^2d^4.$$

通过望远镜求和论证，我们得到

$$\begin{aligned}
\min_{1\leqslant i\leqslant n}\mathbb{E}[B(\mathbf{v}(i))] &\leqslant \frac{A(\mathbf{v}(0))-\mathbb{E}[A(\mathbf{v}(n+1))]+M_1C_3\sum_{i=0}^{\infty}\alpha(i)^2d^4}{2\sum_{i=0}^{n}\alpha(i)}\\
&\leqslant \frac{A(\mathbf{v}(0))+M+M_1C_3\sum_{i=0}^{\infty}\alpha(i)^2d^4}{2\sum_{i=0}^{n}\alpha(i)}.
\end{aligned} \tag{4.3}$$

本手稿仅供审阅。

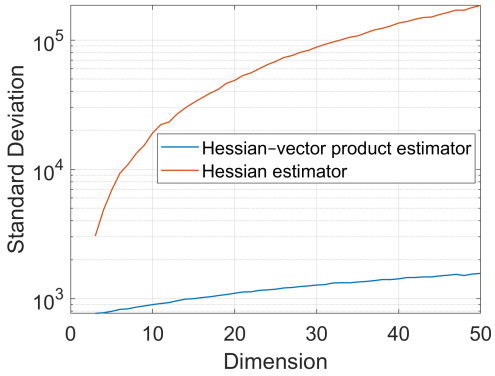

FIG. 1。 *The standard deviations are computed over 10000 samples of* **海森矩阵-向量估计器** *(Hessian estimator)and* $H_{\mathbf{v}}$ *(Hessian-vector estimator), where* $\mathbf{v}$ *is a fixed vector drawn from* $\mathcal{N}(\mathbf{0}, \mathbf{I})$。

因此，维度越大， d 收敛速度越慢。此外，在第5.3小节中，我们说明，较大的变化幅度还会导致潜在的数值不稳定性。幸运的是，我们不需要完整的海森矩阵信息。在（3.4）和（3.5）中，只需近似计算海森矩阵-向量积 $\nabla^2 f(\mathbf{x})\mathbf{v}$即可，这也是 [50] 中避免显式海森矩阵计算的一个关键思想。为此，可以使用由（ $H_{\mathbf{v}}(\mathbf{x}, \mathbf{r}\ l)$ 定义的（3.6）来近似 $\nabla^2 f(\mathbf{x})\mathbf{v}$。

LEMMA 4.5. 假设 假设 3.3 和 假设 3.5成立，则

$$\|\mathbb{E}_{\mathbf{r}}[H_{\mathbf{v}}(\mathbf{x}, \mathbf{r}, l)] - \nabla^2 f(\mathbf{x})\mathbf{v}\|_2 \leqslant Ml^2 d,$$

$$\mathbb{E}_{\mathbf{r}}[\|H_{\mathbf{v}}(\mathbf{x}, \mathbf{r}, l)\|_2^2] \leqslant 4(d+2)M^2 + \frac{M^2 d(d+2)l^4}{9} + \frac{M^2 d(d+2)(d+4)(d+6)l^4}{9} + O(l^8 d^8) \leqslant 10M^2 d.$$

证明。 证明见附录。 □

使用 $H_{\mathbf{v}}$ 显著降低了(4.2)中的变化幅度，从 $O(d^4)$ 降至 $O(d)$ （另见 图 1），并加快了特征向量搜索的收敛。

4.3. 使用零阶海森矩阵-向量估计器的内层无导数特征向量搜索算法的收敛性。 以下命题表明， 算法 3.1 配合海森矩阵-向量估计器 $H_{\mathbf{v}}$的输出可作为原函数的不稳定方向的近似$f(\mathbf{x})$ 以及高斯平滑函数 $f_{l_0}(\mathbf{x})$ 的不稳定方向的近似，且误差较小 l_0。

˜将算法 3.1 的输出记为 $V = [\tilde{\mathbf{v}}_1, \cdots, \tilde{\mathbf{v}}_k]$。

PROPOSITION 4.6. 假设假设3.1、假设3.2以及假设3.3 均成立。那么，对于任意 $\epsilon_{\mathbf{v}}$，几乎处处存在一个有限的 $n_{\mathbf{v}}^{max}$（取决于˜和 $\epsilon_{\mathbf{v}}$ 样本的随机实现$\mathbf{r}(n)$）使得输出 V 的算法3.1 满足 $\left\|\tilde{V}\tilde{V}^\top - VV^\top\right\|_2 \leqslant \epsilon_{\mathbf{v}}$ 且 $\left\|\tilde{V}\tilde{V}^\top - V_l V_l^\top\right\|_2 \leqslant \epsilon_{\mathbf{v}} + \frac{\sqrt{2k}Ml_0^2 d}{4\mu}$.

证明。 证明思路与 [42, 定理 4。13], 类似，而定理 4 本身依赖于 [42, 命题 4。8], 也就是说，式（3.4）中的单特征向量迭代几乎必然收敛到真实不稳定特征向量。通过选择足够大的 n_{max}，我们可以确保算法的输出接近真实特征向量。这里的一个关键区别在于， [42, 命题 4。8] 是在假设存在

本手稿仅供审阅。

无偏估计量的条件下证明的，而估计量 $H_{\mathbf{v}}$ 存在偏差。我们利用条件 $l(n) \leqslant L\sqrt{\alpha(n)}$ 控制偏差，使偏差引入的额外误差渐近可忽略，因此几乎必然收敛的结论也适用于我们的情形。具体而言，对于任意满足 $\|\mathbf{v}\|_2 = 1$ 的$\mathbf{v}$，我们将瑞利商定义为 $A(\mathbf{v}) = \mathbf{v}^T\nabla^2 f(\mathbf{x})\mathbf{v}$，于是

$$
\begin{aligned}
&A(\mathbf{v}(n+1)) - A(\mathbf{v}(n)) \\
=&\mathbf{v}(n+1)^\top \nabla^2 f(\mathbf{x})\mathbf{v}(n+1) - \mathbf{v}(n)^\top \nabla^2 f(\mathbf{x})\mathbf{v}(n) \\
&- \langle (\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top) H_{\mathbf{v}(n)}(\mathbf{x}, \mathbf{r}(n), l(n)), \mathbf{v}(n)\mathbf{v}(n)^\top \nabla^2 f(\mathbf{x})\mathbf{v}(n)\rangle \\
\leqslant& -2\alpha(n)\langle (\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top) H_{\mathbf{v}(n)}(\mathbf{x}, \mathbf{r}(n), l(n)), (\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)\nabla^2 f(\mathbf{x})\mathbf{v}(n)\rangle \\
&+ M_2\alpha(n)^2 \|H_{\mathbf{v}(n)}(\mathbf{x}, \mathbf{r}(n), l(n))\|_2^2,
\end{aligned}
$$

其中 M_2 是一个仅依赖于 M 的正常数。对两边取期望，并利用 引理 4.5，可得

$$
\begin{aligned}
&\mathbb{E}(A(\mathbf{v}(n+1)) - A(\mathbf{v}(n))|\mathcal{F}(n)) \\
&\leqslant -2\alpha(n)B(\mathbf{v}(n)) + M_2\alpha(n)^2 \mathbb{E}_{\mathbf{r}}[\|H_{\mathbf{v}(n)}\|_2^2] \\
&\quad + 2\alpha(n)\langle (\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)(\nabla^2 f(\mathbf{x})\mathbf{v}(n) - \mathbb{E}_{\mathbf{r}}[H_{\mathbf{v}(n)}]), (\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)\nabla^2 f(\mathbf{x})\mathbf{v}(n)\rangle \\
&\leqslant -2\alpha(n)B(\mathbf{v}(n)) + \left(10dM_2M^2 + 2L^2M^2d\right)\alpha(n)^2,
\end{aligned}
$$

其中 $B(\mathbf{v}) = \|(\mathbf{I} - \mathbf{v}\mathbf{v}^\top)\nabla^2 f(\mathbf{x})\mathbf{v}\|_2^2$。然后，与 [42, 命题 4.8] 的证明相同，我们可以证明 $A(\mathbf{v}(n))$ 和 $B(\mathbf{v}(n))$ 满足引理 3.6 中的条件，因此 $\mathbf{v}(n)$ 几乎处处收敛于真实不稳定特征向量。

根据命题 4.6，我们知道内层特征向量搜索的迭代 $\mathbf{v}(n)$ 几乎处处收敛于真实不稳定特征向量。然而，所需的迭代次数是一个取决于样本路径的随机量。在本小节的其余部分中，我们提出一种更加自动的停止准则；为简洁起见，将讨论限制在 $k = 1$ 的情形。根据 [42, 定理 4.13], ，$\mathbf{v}(n)$ 与真实不稳定特征向量之间的距离可以通过残差$\|(\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)\nabla^2 f(\mathbf{x})\mathbf{v}(n)\|_2$进行控制。这促使我们使用零阶残差$\|(\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)\bar{H}_{\mathbf{v}(n)}\| < \epsilon$ 作为潜在的停止准则，其中

$$
\bar{H}_{\mathbf{v}(n)} := \frac{1}{m}\sum_{i=1}^{m} H_{\mathbf{v}(n)}\big(\mathbf{x}, \mathbf{r}(i), l(n)\big).
$$

接下来，我们首先证明停止准则几乎处处能在有限次迭代内成立。我们还证明，使用递增的 m 时，该停止准则以正概率保证在每次内层特征向量搜索中，$\|(\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)\nabla^2 f(\mathbf{x})\mathbf{v}(n)\|_2$ 足够小，因此根据 [42, 定理 4.13]，返回的向量足够接近真实不稳定特征向量。

PROPOSITION 4.7. 在与命题 4.6, $\forall m \geqslant 1, \epsilon > 0$相同的条件下，迭代 $\mathbf{v}(n)$ 几乎处处在有限次迭代内满足停止准则 $\|(\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)\bar{H}_{\mathbf{v}(n)}\|_2 < \epsilon$。

证明。定义终止时间 $T := \inf\{n \geqslant 1, \|(\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)\bar{H}_{\mathbf{v}(n)}\|_2 < \epsilon\}$. 对于每个 n，我们将其分解为

$$
\begin{aligned}
(\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)\bar{H}_{\mathbf{v}(n)} = &(\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)\nabla^2 f(\mathbf{x})\mathbf{v}(n) \\
&+ (\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)(\mathbb{E}_{\mathbf{r}}[\bar{H}_{\mathbf{v}(n)}] - \nabla^2 f(\mathbf{x})\mathbf{v}(n)) \\
&+ (\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)(\bar{H}_{\mathbf{v}(n)} - \mathbb{E}_{\mathbf{r}}[\bar{H}_{\mathbf{v}(n)}])\,.
\end{aligned} \tag{4.4}
$$

本手稿仅供审稿使用。

根据命题 4.6，对于几乎每个随机事件 ω，存在 $N(\omega)$，使得对于所有 $n > N_1(\omega)$，都有 $\|(\mathbf{I}-\mathbf{v}(n)\mathbf{v}(n)^\top)\nabla^2 f(\mathbf{x})\mathbf{v}(n)\|_2 < \epsilon/4$。注意到 $l(n) \leqslant L\sqrt{\alpha(n)} \to 0$，由引理 4.5，存在一个确定性的 N_2，使得对于所有 $n > N_2$（使得 $l(n)$ 足够小），都有 $\|(\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)(\mathbb{E}_{\mathbf{r}}[\bar{H}_{\mathbf{v}(n)}] - \nabla^2 f(\mathbf{x})\mathbf{v}(n))\|_2 < \epsilon/4$。接下来，根据 f 在 $\mathbf{x}$ 处的泰勒展开，对于每个固定的 $\mathbf{v}$ 和 l，我们有：

$$H_{\mathbf{v}} = \frac{F(\mathbf{x}+l\mathbf{v},\mathbf{r},l) - F(\mathbf{x}-l\mathbf{v},\mathbf{r},l)}{2l} = \mathbf{r}\mathbf{r}^\top\nabla^2 f(\mathbf{x})\mathbf{v} + O(l^2),$$

其中，$O(l^2)$ 项在对 f 的正则性假设下关于 $\mathbf{r}$ 一致。由于每个 $\mathbf{r}(i)$ 都取自标准正态分布，当 $l(n)$ 较小时，存在 $N_3 \in \mathbb{N}$ 和一个常数 $\delta_p > 0$，使得对于所有 $n > N_3$，都有 $\mathbb{P}(\|(\mathbf{I} - \mathbf{v}(n)\mathbf{v}(n)^\top)(\bar{H}_{\mathbf{v}(n)} - \mathbb{E}_{\mathbf{r}}[\bar{H}_{\mathbf{v}(n)}])\|_2 < \epsilon/2) \geqslant \delta_p$。因此，$\forall n > N_4 = \max(N_1, N_2, N_3)$ 时，有 $\mathbb{P}(T > n) \leqslant (1-\delta_p)^{n-N_4}$，这意味着 $\mathbb{P}(T = \infty) = 0$。 □

PROPOSITION 4.8. 令 T 为终止时间，即停止准则 $\|(\mathbf{I}-\mathbf{v}(T)\mathbf{v}(T)^\top)\bar{H}_{\mathbf{v}(T)}\| < \epsilon$ 成立的时刻，并且 m 足够大，使得 $10M^2 d < m\epsilon^2$，那么

$$\mathbb{P}\left(\|(\mathbf{I}-\mathbf{v}(T)\mathbf{v}(T)^\top)\nabla^2 f(\mathbf{x})\mathbf{v}(T)\|_2 < 2\epsilon + Ml(T)^2 d\right) \geqslant 1 - \frac{10M^2 d}{m\epsilon^2}.$$

证明。由（4.4），我们有

$$\begin{aligned}
&\|(\mathbf{I}-\mathbf{v}(T)\mathbf{v}(T)^\top)\nabla^2 f(\mathbf{x})\mathbf{v}(T)\|_2 \\
&\leqslant \|(\mathbf{I}-\mathbf{v}(T)\mathbf{v}(T)^\top)\bar{H}_{\mathbf{v}(T)}\|_2 + \|(\mathbf{I}-\mathbf{v}(T)\mathbf{v}(T)^\top)(\mathbb{E}_{\mathbf{r}}[\bar{H}_{\mathbf{v}(T)}] - \nabla^2 f(\mathbf{x})\mathbf{v}(T))\|_2 \\
&\quad + \|(\mathbf{I}-\mathbf{v}(T)\mathbf{v}(T)^\top)(\bar{H}_{\mathbf{v}(T)} - \mathbb{E}_{\mathbf{r}}[\bar{H}_{\mathbf{v}(T)}])\|_2 \\
&< \epsilon + Ml(T)^2 d + \|(\mathbf{I}-\mathbf{v}(T)\mathbf{v}(T)^\top)(\bar{H}_{\mathbf{v}(T)} - \mathbb{E}_{\mathbf{r}}[\bar{H}_{\mathbf{v}(T)}])\|_2.
\end{aligned}$$

注意

$$\begin{aligned}
&\mathbb{E}_{\mathbf{r}}[\|(\mathbf{I}-\mathbf{v}(T)\mathbf{v}(T)^\top)(\bar{H}_{\mathbf{v}(T)} - \mathbb{E}_{\mathbf{r}}[\bar{H}_{\mathbf{v}(T)}])\|_2^2] \leqslant \mathbb{E}_{\mathbf{r}}[\|\bar{H}_{\mathbf{v}(T)} - \mathbb{E}_{\mathbf{r}}[\bar{H}_{\mathbf{v}(T)}]\|_2^2] \\
&\leqslant \frac{\mathbb{E}_{\mathbf{r}}[\|H_{\mathbf{v}(T)} - \mathbb{E}_{\mathbf{r}}[H_{\mathbf{v}(T)}]\|_2^2]}{m} \leqslant \frac{\mathbb{E}_{\mathbf{r}}[\|H_{\mathbf{v}(T)}\|_2^2]}{m} \leqslant \frac{10M^2 d}{m},
\end{aligned}$$

其中，第一个不等式由以下界得到：$\|\mathbf{I} - \mathbf{v}(T)\mathbf{v}(T)^\top\|_2 \leqslant 1$；第二个不等式成立是因为 $H_{\mathbf{v}}$ 是 m 个 $H_{\mathbf{v}}$ 的独立副本的平均值，而最后一个不等式由引理 4.5 得出。由切比雪夫不等式，我们得到

$$\mathbb{P}(\|(\mathbf{I}-\mathbf{v}(T)\mathbf{v}(T)^\top)(\bar{H}_{\mathbf{v}(T)} - \mathbb{E}_{\mathbf{r}}[\bar{H}_{\mathbf{v}(T)}])\|_2 \geqslant \epsilon) \leqslant \frac{10M^2 d}{m\epsilon^2}.$$

COROLLARY 4.9. 假设命题 4.6 中的条件成立。在外层鞍点算法的第 n 次迭代中，在内层特征向量搜索的停止准则中，选择样本大小 $m_n = O(n^p)$，其中 $p > 1$。也就是说，在内循环中，我们构造 $\bar{H}_{\mathbf{v}} := \frac{1}{m_n}\sum_{i=1}^{m_n} H_{\mathbf{v}}(x_n, \mathbf{r}(i), l)$，$\mathbf{r}(i) \sim \mathcal{N}(0, I)$，并在满足 $\|(I - vv^\top)\bar{H}_{\mathbf{v}}\| < \epsilon$ 时立即停止。于是，存在一个正概率，使得内层算法返回的向量 $v_n(T)$ 满足

$$\|(\mathbf{I}-\mathbf{v}_n(T)\mathbf{v}_n(T)^\top)\nabla^2 f(\mathbf{x}_n)\mathbf{v}_n(T)\|_2 < 2\epsilon + Ml(0)^2 d, \forall n \in \mathbb{N}.$$

证明。注意到 $l(0) \geqslant l$（T）由 假设 3.2 可得。根据 命题 4.8，

$$\mathbb{P}(\|(\mathbf{I}-\mathbf{v}_n(T)\mathbf{v}_n(T)^\top)\nabla^2 f(\mathbf{x}_n)\mathbf{v}_n(T)\|_2 < 2\epsilon + Mdl(0)^2) \geqslant 1 - \frac{10M^2 d}{m_n\epsilon^2}.$$

本手稿仅供审阅。

此外，选择带有 $p > 1$ 的 $m_n = O(n^p)$ 后，

$$\|(\mathbf{I} - \mathbf{v}_n(T)\mathbf{v}_n(T)^\top)\nabla^2 f(\mathbf{x}_n)\mathbf{v}_n(T)\|_2 < 2\epsilon + Mdl(0)^2 \text{ for every } n \in \mathbb{N}$$

成立，且其下界为 $\prod_{n=1}^{\infty}\left(1 - O\left(\frac{1}{n^p}\right)\right) > 0$。 □

4.4. 外层无导数鞍点搜索算法的收敛性。根据命题 4.6，算法 3.1 的输出 $\{\tilde{\mathbf{v}}_i\}_1^k$ 几乎处处能够在有限次迭代内，以任意小的误差近似 $f(\mathbf{x})$ 的不稳定方向（即 $\{\mathbf{v}_i\}_1^k$）和 $f_l(\mathbf{x})$ 的不稳定方向（即 $\{\mathbf{v}_{l,i}\}_1^k$），前提是 l 足够小。值得注意的是，$\mathbf{x}(n)$ 通常接近 $\mathbf{x}(n+1)$，因此 $\{\mathbf{v}_i(n)\}_1^k$ 可作为 $\{\mathbf{v}_i(n+1)\}_1^k$ 的良好初始条件；所以在实际应用中，通常只需较小的 $n_{\mathbf{v}}^{max}$（例如 $n_{\mathbf{v}}^{max} = 10$），便能以令人满意的精度得到 $\mathbf{v}(n+1)$。因此，我们建议在实践中直接设定最大迭代次数 $n_{\mathbf{v}}^{max}$，以降低计算成本，而不采用推论 4.9 中要求进行大量函数求值的停止准则。尽管如此，推论 4.9 中的停止准则能够保证每次内层特征向量搜索都以正概率返回良好的结果。本小节假定这一事件发生，即假定内循环返回的向量足够接近不稳定特征向量，具体如下文所述。

ASSUMPTION 4.1. 存在一个小参数 θ，使得在每隔 n 次外层迭代时，

$$\theta > \frac{\sqrt{2k}Ml(n)^2 d}{4\mu} > 0, \text{ and } (1-\sqrt{\theta})\mu_{l(n)} - M\theta - 5M\sqrt{\theta} > 0 \tag{4.5}$$

成立，其中 $\mu_{l(n)} = \mu - Ml(n)^2 d/2$，且以下条件几乎处处成立，

$$\min(\|\tilde{V}(n)\tilde{V}(n)^\top - V(n)V(n)^\top\|_2^2, \|\tilde{V}(n)\tilde{V}(n)^\top - V_{l(n)}V_{l(n)}^\top\|_2^2) \leqslant \theta, \tag{4.6}$$

其中，$\tilde{V}(n)\tilde{V}(n)^\top = \sum_{i=1}^k \tilde{\mathbf{v}}_i(\mathbf{x}(n))\tilde{\mathbf{v}}_i(\mathbf{x}(n))^\top$ 分别是相对于算法 3.1 在 $x = \mathbf{x}(n)$ 时的输出的投影算子。同样，算子$V(n)V(n)^\top = \sum_{i=1}^k \mathbf{v}_i(\mathbf{x}(n))\mathbf{v}_i(\mathbf{x}(n))^\top$, $V_{l(n)}V_l^\top(n) = \sum_{i=1}^k \mathbf{v}_{l(n),i}(\mathbf{x}(n))\mathbf{v}_{l(n),i}(\mathbf{x}(n))^\top$ 是相对于 k 对应于 $\nabla^2 f(\mathbf{x}(n))$ 和 $\nabla^2 f_{l(n)}(\mathbf{x}(n))$ 的 k 个最小特征值的特征向量的投影。

THEOREM 4.10（衰减 $l(n)$，采用衰减步长 $\alpha(n)$）。假设假设 3.1、假设 3.2、假设 3.3、假设 3.4 和假设 4.1 均成立。若初始点 $\mathbf{x}(0)$ 足够接近 $\mathbf{x}^*$，且 $\sum_{n=0}^{\infty}\alpha(n)^2$ 足够小（例如，$\alpha(n) = \frac{\gamma}{(n+m)^p}$，其中 $p \in (1/2, 1]$，且 m 足够大），则以下有界性事件以高概率发生，

$$E^\infty = \{\mathbf{x}(n) \in U, \forall n\},\ U = \left\{\mathbf{x}, \|\mathbf{x} - \mathbf{x}^*\|_2 \leqslant \min\left(\frac{(1-\sqrt{\theta})\mu}{M} - \theta - 5\sqrt{\theta}, \delta\right)\right\}. \tag{4.7}$$

在 E^∞的条件下，对于任意给定的容差 $\epsilon_{\mathbf{x}} > 0$，算法 3.2 几乎必然经过有限次迭代后到达 $\mathbf{x}^*$ 的 $\epsilon_{\mathbf{x}}$-邻域。

证明。证明见附录。由于 $F(\mathbf{x}, \mathbf{r}, l)$ 是一个有偏梯度估计器，其偏差项的阶为 $O(l^2)$（引理 3.3），我们可以利用步长 $\alpha(n)$（假设 3.2）对应的衰减 $l(n)$ 的界来控制这一偏差误差。 □

本手稿仅供审阅。

然而，如果 $l(n)$（它在 $F(\mathbf{x}(n), \mathbf{r}(n)\ l(n))$ 中作为分母）过小，那么 $F(\mathbf{x}(n), \mathbf{r}(n)\ l(n))$ 可能会受到舍入误差的主导，从而无法准确近似梯度。因此，在本节余下部分，我们考虑 $l(n) \equiv l > 0$ 足够小但保持为常数的情况。

LEMMA 4.11. 设假设 3.1、假设 3.3、假设 3.4、假-设 3.5以及假设 4.1 均成立。如果初始点 $\mathbf{x}(0)$ 充分接近于 $\mathbf{x}^*$ 且 $\sum_{n=0}^{\infty} \alpha(n)^2$ 充分小，则有界性事件

$$E_l^{\infty} = \{\mathbf{x}(n) \in U, \forall n\},\ U_l = \left\{\mathbf{x}, \|\mathbf{x} - \mathbf{x}^*\|_2 \leqslant \min\left(\frac{(1-\sqrt{\theta})\mu_l}{M} - \theta - 5\sqrt{\theta}, \delta\right)\right\}$$

以高概率发生。在给定 E_l^{∞}的条件下，对于任意容差 $\epsilon_{\mathbf{x}} > 0$，算法 3.2 几乎处处都会在有限次迭代内到达 $\epsilon_{\mathbf{x}}$-邻域$\mathbf{x}_l^*$。

证明。 注意到 $F(\mathbf{x}(n), \mathbf{r}(n)\ l)$ 是 $f_l(\mathbf{x}(n)$ 的无偏梯度估计量。因此，证明过程与定理 4.10中的步骤相同，只需假设偏差误差为零。 □

引理 4.11 表明，迭代点以较大的概率收敛到 $\mathbf{x}_l^*$。因此，自然要估计 $\mathbf{x}_l^*$ 与目标鞍点 $\mathbf{x}^*$之间的距离。根据（3.8），$\nabla f(\mathbf{x})$ 与 $\nabla f_l(\mathbf{x})$ 的梯度和海森矩阵之差为 $O(l^2)$。因此，我们可以预期它们的零点，即$\mathbf{x}^*$ 和 $\mathbf{x}_l^*$，相距 $O(l^2$，这一点在引理 4.12中得到了精确说明。

LEMMA 4.12. 假设 假设 3.3 和 假设 3.4 成立。若 $l > 0$ 是一个充分小的参数，使得 $\mu_l > 0$，$\frac{Ml^2 d}{\mu_l} = \delta_l \leqslant \frac{\delta}{2}$，$\frac{M\delta_l}{\mu_l} \leqslant \frac{1}{2}$. 则存在~~一个~~ $\mathbf{x}_l^* \in \{\mathbf{x}\ \|\mathbf{x}-\mathbf{x}$~~f v41'>~~，它是 k 的一个指标为 f_l 的鞍点，使得

$$\|\mathbf{x}_l^* - \mathbf{x}^*\|_2^2 \leqslant \delta_l^2 = \left(\frac{Ml^2 d}{\mu_l}\right)^2 = O(l^4). \tag{4.8}$$

证明。 我们借鉴 [24, 第 5 节4] 中所述的技巧，来估计两个函数的零点之间的距离，这两个函数的梯度和海森矩阵彼此接近。定义牛顿迭代映射为 $T(\mathbf{x}) = \mathbf{x} - (\nabla^2 f_l(\mathbf{x}^*))^{-1}\nabla f_l(\mathbf{x})$。对于 $\mathbf{x}$，$\mathbf{y}$ $\in\{\mathbf{x}\ \|\mathbf{x} - \mathbf{x}^*\|_2 \leqslant \delta_l\}$，我们有

$$\begin{aligned}
\|T(\mathbf{x}) - T(\mathbf{y})\|_2 &= \|(\mathbf{x}-\mathbf{y}) - (\nabla^2 f_l(\mathbf{x}^*))^{-1}(\nabla f_l(\mathbf{x}) - \nabla f_l(\mathbf{y}))\|_2 \\
&= \left\|\left[(\nabla^2 f_l(\mathbf{x}^*))^{-1}\int_0^1 (\nabla^2 f_l(\mathbf{x}^*) - \nabla^2 f_l(\mathbf{y} + t(\mathbf{x}-\mathbf{y})))\mathrm{d}t\right](\mathbf{x}-\mathbf{y})\right\|_2 \\
&\leqslant \frac{M\delta_l}{\mu_l}\|\mathbf{x}-\mathbf{y}\|_2 \leqslant \frac{\|\mathbf{x}-\mathbf{y}\|_2}{2}.
\end{aligned}$$

由式（3.8），可得 $\|\nabla f_l(\mathbf{x}^*)\|_2 = \|\nabla f_l(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)\|_2 \leqslant \frac{Ml^2 d}{2}$；由式（3.9），可知 $\|(\nabla^2 f_l(\mathbf{x}^*))^{-1}\|_2 \leqslant \frac{1}{\mu_l}$。因此，如果 $\mathbf{x} \in \{\mathbf{x}\ \|\mathbf{x} - \mathbf{x}^*\|_2 \leqslant \delta_l\}$，

$$\begin{aligned}
&\|T(\mathbf{x}) - \mathbf{x}^*\|_2 = \|\mathbf{x} - (\nabla^2 f_l(\mathbf{x}^*))^{-1}\nabla f_l(\mathbf{x}) - \mathbf{x}^*\|_2 \\
&\leqslant \underbrace{\|\mathbf{x} - (\nabla^2 f_l(\mathbf{x}^*))^{-1}(\nabla f_l(\mathbf{x}) - \nabla f_l(\mathbf{x}^*)) - \mathbf{x}^*\|_2}_{=\|T(\mathbf{x}) - T(\mathbf{x}^*)\|_2} + \underbrace{\|(\nabla^2 f_l(\mathbf{x}^*))^{-1}\nabla f_l(\mathbf{x}^*)\|_2}_{\leqslant\|(\nabla^2 f_l(\mathbf{x}^*))^{-1}\|_2\|\nabla f_l(\mathbf{x}^*)\|_2} \\
&\leqslant \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2 + \frac{Ml^2 d}{2\mu_l} \leqslant \frac{\delta_l}{2} + \frac{\delta_l}{2} = \delta_l,
\end{aligned}$$

本手稿仅供评审使用。

这意味着 $T(\mathbf{x})$ 在 $\{\mathbf{x} \ \|\mathbf{x}-\mathbf{x}^*\|_2 \leqslant \delta_l\}$ 中是一个压缩映射。根据巴拿赫不动点定理，存在一个 $\mathbf{x}_l^* \in \{\mathbf{x} \ \|\mathbf{x}-\mathbf{x}^*\|_2 \leqslant \delta_l\}$，使得$T(\mathbf{x}_l^*) = \mathbf{x}_l^*$，这意味着 $\mathbf{x}_l^*$ 是 $f_l(\mathbf{x})$ 的临界点。由于 $\mathbf{x}_l^* \in \{\mathbf{x} \ \|\mathbf{x}-\mathbf{x}^*\|_2 \leqslant \delta_l\} \subset \{\mathbf{x} \ \|\mathbf{x}-\mathbf{x}^*\|_2 \leqslant \delta\}$，可得（4.8）。由式（3.9）可知，$\mathbf{x}_l^*$ 是 $f_l(\mathbf{x})$ 的指标为k的鞍点。 □

根据 引理 4.11 和 引理 4.12，可直接得到以下定理。

THEOREM 4.13（常数 $l(n)$，采用衰减步长 $\alpha(n)$）。在引理 *4.11* 和引理 *4.12* 给出的相同条件和假设下，算法 *3.2* 的输出满足 $\|x^{end}-x^*\|_2 \leqslant \|x^{end}-x_l^*\|_2 + \|x^*-x_l^*\|_2 < \epsilon_{\mathbf{x}} + \frac{Ml^2d}{\mu_l}$，这意味着，当迭代次数趋于无穷大、容差 $\epsilon_{\mathbf{x}}$趋近于零时，输出 x^{end} 与目标鞍点 x^*之间的距离，即 $\|x^{end}-x^*\|_2^2$，几乎处处不大于一个小常数 $\left(\frac{Ml^2d}{\mu_l}\right)^2 = O(l^4)$。

尽管上述讨论从理论上表明，算法 3.2的迭代几乎处处收敛到$\mathbf{x}^*$（采用衰减的 $l(n)$）或收敛到$\mathbf{x}_l^*$（采用固定的$l(n)$），且这一结论以有界性事件为条件，但假设 3.1 中定义的衰减步长会导致次线性收敛速率。根据引理 3.3，随着迭代点接近鞍点，$\|\nabla f(\mathbf{x})\|_2$会趋于零，而 $F(\mathbf{x}, \mathbf{r}\ l)$ 的方差也几乎趋于零。这种现象称为方差缩减 [23, 7], ，随机梯度方法通常利用它来加速收敛。在这种情况下，通常无需使用衰减步长来控制随机噪声，因此可以采用固定步长以实现线性收敛速率。

THEOREM 4.14（固定 $l(n) \equiv l$，采用固定步长 $\alpha(n) \equiv \alpha$）。假设假设 *3.3*、假设 *3.4* 和假设 *4.1* 均成立。选取足够小的步长 $\alpha(n) \equiv \alpha > 0$，使得 $0 < \alpha < \frac{(1-\sqrt{\theta})\mu - M\theta - 5M\sqrt{\theta}}{(2d+4)M^2+1}$。那么，只要有界性事件 E^∞ 以正概率发生，算法 *3.2* 生成的迭代就会以由常数 $0 < \sigma < 1$确定的线性衰减速率趋近于 x^* 的一个邻域，该邻域的半径为 $O(l^2/\sqrt{\alpha})$，即

$$\mathbb{E}[\|\mathbf{x}(n)-\mathbf{x}^*\|_2^2 | E^\infty] \leqslant O(\sigma^n + l^4/\alpha). \tag{4.9}$$

证明。定义事件 $E^n = \{\mathbf{x}(i) \in U, \ i = 1, 2, \cdots, n\}$ 及其对应的指示函数 $\mathbf{1}_{E^n}$。根据 附录中的引理 6.1 和引理 3.3，可得

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}_{E^{n+1}}\|\mathbf{x}(n+1)-\mathbf{x}^*\|_2^2] \leqslant \mathbb{E}[\mathbf{1}_{E^n}\|\mathbf{x}(n+1)-\mathbf{x}^*\|_2^2] \\
&\leqslant \mathbb{E}[\mathbb{E}[\mathbf{1}_{E^n}\|\mathbf{x}(n+1)-\mathbf{x}^*\|_2^2|\mathcal{F}(n)]] \\
&\leqslant (1-\alpha((1-\sqrt{\theta})\mu - M\theta - 5M\sqrt{\theta}))\mathbb{E}[\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2^2] \\
&\quad + \alpha M d l^2 \mathbb{E}[\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2] \\
&\quad + \alpha^2\left((2d+4)\mathbb{E}[\mathbf{1}_{E^n}\|\nabla f(\mathbf{x})\|_2^2] + M^2 l^4 d(d+2)(d+4)(d+6)/18\right).
\end{aligned}$$

由杨氏不等式和詹森不等式可得

$$\begin{aligned}
\alpha M d l^2 \mathbb{E}[\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2] &\leqslant \alpha^2(\mathbb{E}[\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2])^2 + M^2 d^2 l^4/4 \\
&\leqslant \alpha^2 \mathbb{E}[\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2^2] + M^2 d^2 l^4/4.
\end{aligned}$$

注意，$\mathbf{1}_{E^n}\|\nabla f\mathbf{x}(n))\|_2^2 \leqslant \mathbf{1}_{E^n} M^2 \|\mathbf{x}(n) - \mathbf{x}^*\|_2^2$。于是，我们得到

$$\mathbb{E}[\mathbf{1}_{E^{n+1}}\|\mathbf{x}(n+1)-\mathbf{x}^*\|_2^2] \leqslant \sigma \mathbb{E}[\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2^2] + C_4,$$

本手稿仅供审阅。

其中，$\sigma = \sigma(\alpha, \theta, \mu, M, d) = 1 - \alpha((1-\sqrt{\theta})\mu - M\theta - 5M\sqrt{\theta}) + \alpha^2(2d+4)M^2 + \alpha^2$以及 $C_4 = \frac{M^2 d^2 l^4}{4} + \frac{\alpha^2 M^2 l^4 d}{(d+2)(d+4)(d+6)} \frac{(d+2)(d+4)(d+6)}{18} = O(l^4)$。通过望远镜求和，我们得到

$$\mathbb{E}[\mathbf{1}_{E^n}\|\mathbf{x}(n) - \mathbf{x}^*\|_2^2] \leqslant \sigma^n \|\mathbf{x}(0) - \mathbf{x}^*\|_2^2 + C_4 \sum_{i=0}^{n-1} \sigma^i \leqslant \sigma^n \|\mathbf{x}(0) - \mathbf{x}^*\|_2^2 + \frac{C_4}{1-\sigma}.$$

由于 $C_4 = O(l^4)$ 且 $1 - \sigma = O(\alpha)$，证明至此完成。 □

由（4.9），我们可以预期 算法 3.2 最初呈现线性收敛速率 σ^n，随后趋于一个较小的误差 $O(l^4/\alpha)$，这就产生了优化中通常所称的平台误差。

备注 4.15。第 5 节 中的数值测试 表明，第 定理 4.14 中的有界性假设确实成立。如果已知目标函数的适当全局信息，则可以去除有界性假设。例如，如果 $\nabla^2 f(\mathbf{x})$ 的特征值满足（3.7）且

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leqslant C_5 < (1-\sqrt{\theta})\mu - M\theta - 5M\sqrt{\theta}, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d,$$

可以证明，鞍点搜索算法的吸引域是整个空间 $\mathbb{R}^d$。假设 α 和 θ 均足够小，则 $\sigma \approx 1 - \alpha\mu$。因此，较大的 α 不仅能加快误差衰减，还能将平台误差降低至 $O(l^4/\alpha)$ 的量级；关于这一点的进一步说明，请参见 第 5.1 小节 中的数值实验。因此，定理 4.14 建议在保持数值稳定性的同时，将步长 α选取得尽可能大。

5. 数值实验。本节通过一系列数值示例展示该算法的实际适用性和收敛速度，并说明平台误差如何依赖于 l 和 α。

注意，当$\mathbf{x}(n)$接近严格鞍点$\mathbf{x}^*$时，我们有

$$\mu^2\|\mathbf{x}(n) - \mathbf{x}^*\|_2^2 \leqslant \|\nabla f(\mathbf{x}(n))\|_2^2 = \|\nabla f(\mathbf{x}(n)) - \nabla f(\mathbf{x}^*)\|_2^2 \leqslant M^2\|\mathbf{x}(n) - \mathbf{x}^*\|_2^2.$$

因此，$\|\mathbf{x}(n) - \mathbf{x}^*\|_2^2$ 和 $\|\nabla f(\mathbf{x}(n))\|_2^2$ 均可用于衡量误差。

5.1. M¨uller-Brown 势能。这个数值示例展示了所提出的无导数鞍点搜索算法的实用性，并验证了理论分析中推导的平台误差阶数。M¨uller-Brown（MB）势能是理论化学中的一个二维函数，用于描述一个具有多个局部极小值和鞍点的系统，是测试鞍点搜索算法的标准基准。MB势能定义为 [4]

$$E(x, y) = \sum_{i=1}^{4} A_i e^{a_i(x-\bar{x}_i)^2 + b_i(x-\bar{x}_i)(y-\bar{y}_i) + c_i(y-\bar{y}_i)^2},$$

其中 $A = [-200, -100, -170, 15]$，$a = [-1, -1, -6.5, 0.7]$，$b = [0, 0, 11, 0.6]$，$c = [-10, -10, -6.5, 0.7]$，$\bar{x} = [1, 0, -0.5, -1]$，以及 $\bar{y} = [0, 0.5, 1.5, 1]$. 我们展示了 MB 势能的等高线图，见图 2 (a)，图中有两个局部极小点（反应物和产物）以及一个连接这两个极小点、指标为 1 的鞍点（过渡态）。

我们将初始条件设为 $\mathbf{x}(0) = (0, 1)^\top$，并计算 $l = 0.001$(图 2(b)) 和 $l = 0.0001$ (图 2(c)) 中误差随迭代次数的变化。结果表明，在迭代初期，误差以

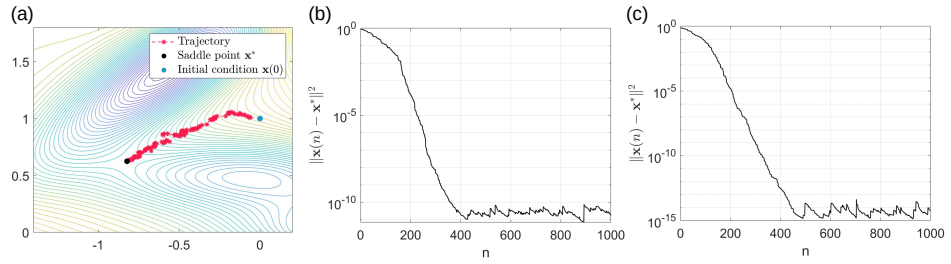

FIG. 2 *Contour plot of MB potential and iteration points (a) of one-time run of Algorithm 3 2. 。。, , , , Error plots of the iteration points with (b)* $l = 0.001$ $\alpha_{\mathbf{x}} \equiv 0.0001$ $n_{\mathbf{x}}^{max} = 1000$ $n_{\mathbf{v}}^{max} = 100$ $\alpha_{\mathbf{v}} \equiv 0.0002$ *and (c)* $l = 0.0001$ $\alpha_{\mathbf{x}} \equiv 0.0001$ $n_{\mathbf{x}}^{max} = 1000$ $n_{\mathbf{v}}^{max} = 100$ $\alpha_{\mathbf{v}} \equiv 0.0002/d$. , , , ,

指数速率下降，最终趋于一个较小的平台误差，这与定理 4.14 一致。此外，随着 l 从 0.001 降至 0.0001，平台误差从 10^{-11} 降至 10^{-15}，这与定理 4.13 和定理 4.14 预测的 $O(l^4)$ 误差相符。为进一步验证定理 4.14 中的 $O(l^4/\alpha)$ 平台误差，我们在表 1中考察了不同差分长度 l 和步长 α 下的平台误差。平台误差定义为 100 次运行中误差极小值的平均值，计算公式为 $\frac{1}{100}\sum_{i=1}^{100}\min_n \|\mathbf{x}^i(n) - \mathbf{x}^*\|_2^2$。对于 l，平台误差的衰减速率约为 $O(l^4)$。对于 α，当 $\overline{\alpha}$ 从 0.0001 增至 0.0002 时，平台误差大致减半。表1中的数值测试与 $O(l^4/\alpha)$ 的标度关系一致。

TABLE 1 *Plateau error is computed as the mean of the minimum errors across* $k = 100$ *runs defined by,* $\frac{1}{k}\sum_{i=1}^{k} \quad {}_n\| \quad {}^i \; n \; - \; {}^*\|_2^2$ *with* $n_{\mathbf{x}}^{max} = 1000$ $n_{\mathbf{v}}^{max} = 100$ $\alpha_{\mathbf{v}} \equiv 0.0002$ $l \; n \; \equiv 0.001$. $\underline{\min}\mathbf{x}(\,)\,\mathbf{x}, \; , \; , \; , \; 0$

	$\alpha = 0.0001$		$\alpha = 0.0002$	
l	平台误差	消失阶	平台误差	消失阶
2^{-8}	2.71E-09		1.28E-09	
2^{-9}	1.58E-10	4.28	7.73E-11	4.15
2^{-10}	1.02E-11	3.89	4.84E-12	3.99
2^{-11}	6.40E-13	3.97	2.96E-13	4.09
2^{-12}	3.87E-14	4.14	2.02E-14	3.66

5.2. 隐式定义的目标函数。 这个数值示例表明，我们的方法适用于隐式定义的目标函数，这类函数的梯度和海森矩阵难以显式求得。例如，超参数优化 [14] 以及材料科学中的自洽场理论 [15]，都涉及此类问题，其中目标函数通常通过一个优化问题来定义。为简洁起见并便于可视化，我们考虑以下二维隐式函数：

$$f(x,y) = \min_{\mathbf{z}} g(x,y,\mathbf{z}) = \min_{\mathbf{z}} (x-z_1)^2 + (y-z_2)^2 + \sin(z_1 z_2), \tag{5.1}$$

其中，每次函数求值都被定义为求解一个相关的优化问题。在 图 3(a) 中，我们给出了能量景观的等高线图，图中表明 $\mathbf{x}^* = (0, \; 0)$ 是 $f(x, \; y)$ 的指标为 1 的鞍点。

实际上， f 的海森矩阵由隐函数定理给出

$$\nabla^2 f(x,y) = \begin{pmatrix} g_{xx} & g_{xy} \\ g_{yx} & g_{yy} \end{pmatrix} - \begin{pmatrix} g_{xz_1} & g_{xz_2} \\ g_{yz_1} & g_{yz_2} \end{pmatrix} \begin{pmatrix} g_{z_1z_1} & g_{z_1z_2} \\ g_{z_2z_1} & g_{z_2z_2} \end{pmatrix}^{-1} \begin{pmatrix} g_{z_1x} & g_{z_1y} \\ g_{z_2x} & g_{z_2y} \end{pmatrix}.$$

本手稿仅供评审使用。

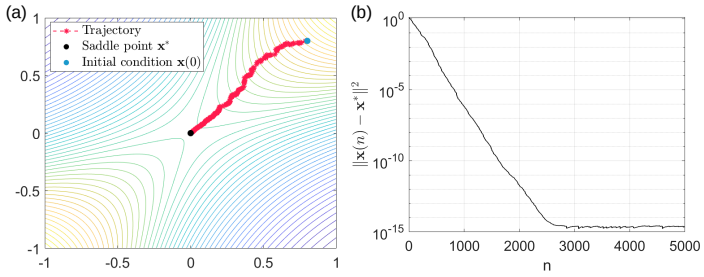

FIG. 3 *(a) Contour plot of* 5 1 *and iteration points of one-time run of Algorithm* 3 2. *(b)* 。（。）。，，，，*Error plot with* $l = 0.1$ $\alpha_{\mathbf{x}} \equiv 0.01$ $n_{\mathbf{x}}^{max} = 5000$ $n_{\mathbf{v}}^{max} = 100$ $\alpha_{\mathbf{v}} \equiv 0.0002/d$.

所有导数均在 $(x, y, \mathbf{z}) = (0, 0, 0)$ 处取值时，f 在 $\mathbf{x}^*$ 处的海森矩阵为

$$\nabla^2 f(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 8/3 & -4/3 \\ -4/3 & 8/3 \end{bmatrix} = \begin{bmatrix} -2/3 & 4/3 \\ 4/3 & -2/3 \end{bmatrix},$$

其特征值为 $\lambda_1 = 2/3, \lambda_2 = -2$, 这表明原点是 f（x, y）的指标为 1 的鞍点。见图 3（b）中绘制的误差随迭代次数的变化图，可以看出，迭代点在线性收敛速率下趋近鞍点，随后误差趋于一个较小的平台误差。

5.3. 修正的罗森布罗克函数。 以下数值实验表明，该算法可用于定位高维模型中的高指标鞍点；与海森矩阵近似相比，海森矩阵-向量积近似效率更高；此外，零阶算法在病态问题上也会遇到收敛缓慢的情况，这与典型的基于梯度的算法类似。Rosenbrock 函数 [41], 也称为香蕉函数，其特点是存在一条狭窄且弯曲的谷地，极小值附近的能量景观较为平坦。由于其几何形态颇具挑战性，难以收敛至全局极小值，因此它是优化算法的经典基准测试。为将 Rosenbrock 函数用于鞍点计算，我们通过添加额外的二次反正切项 [48],对其进行修正，得到

$$f(\mathbf{x}) = \sum_{i=1}^{d-1} 100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2 + \sum_{i=1}^{d} s_i \arctan^2(x_i - 1),$$

其中，$\{s_i\}_1^d$ 是用于调节临界点 $\mathbf{x}^* = (1, \cdots, 1)^\top$指标的参数。例如，当 $d = 2$, $s_1 = -50, s_2 = 1$时，$\mathbf{x}^*$是指标为 1 的鞍点，且条件数 $Cond(\nabla^2 f(\mathbf{x}^*)) \approx 47$。图 4(a) 展示了修正的罗森布罗克函数的能量景观，表明 $\mathbf{x}^*$ 附近的能量景观沿不稳定方向相对平坦。收敛速度在很大程度上取决于 $\nabla^2 f(\mathbf{x}^*)$ 的条件数。根据定理 4.14，误差大约按 $\sigma^n \approx (1 - \alpha\mu)^n$ 衰减。为保证数值稳定性，步长 α 应小于 $2/M$，由此得到的误差衰减率约为 $(1 - 2/Cond(\nabla^2 f(\mathbf{x}^*)))^n$。从图 4(a) 的迭代轨迹还可以看出，迭代点沿陡峭方向收敛得更快，而沿着对应于最小特征值绝对值的特征向量移动较慢。图 4(b) 显示，误差在迭代初期（沿陡峭方向）快速减小，但在后期（沿平坦方向）减慢。

本手稿仅供审阅。

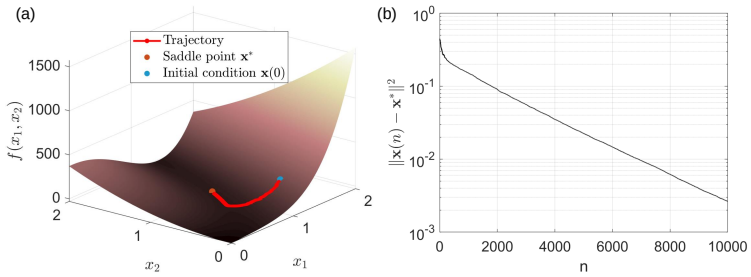

FIG. 4。 *(a) Energy landscape of the 2D Rosenbrock function and iteration points of one-time run of Algorithm 3 2. (b) Error plot of the iteration points of Algorithm 3 2 with* $l = 0.0001$ 。。, $\alpha_{\mathbf{x}} \equiv 0.00001$ $n_{\mathbf{x}}^{max} = 10000$ $n_{\mathbf{v}}^{max} = 100$ $\alpha_{\mathbf{v}} \equiv 0.0002/d$. ,,,

接下来，我们考察在特征向量搜索中使用海森矩阵-向量近似 $H_{\mathbf{v}}$相较于完整海森矩阵近似 H 的方差缩减优势。我们搜索 $\nabla^2 f(\mathbf{x}^* + 0.5\mathbf{r})$ 的最小特征值及其对应的特征向量，其中 r 取自标准正态分布，并在图 5(a) 中绘制所得误差。当维数较低，即 d =3 时，海森矩阵近似和海森矩阵-向量近似的表现都很好。然而，当 $d = 100$ 时，海森矩阵近似 H 的方差达到 10^8，量级，而海森矩阵 - 向量近似的方差仅约为 10^2。因此，使用海森矩阵近似进行特征向量搜索会变得不稳定，无法收敛到真实特征向量 $\mathbf{v}^*$；而海森矩阵-向量近似仍然表现良好。尽管图 5(a) 中需要数千次迭代才能得到精确结果，但在实际应用中，以 $\nabla^2 f(\mathbf{x}(n))$ 的近似不稳定方向作为初始条件，我们通常能在算法 3.1 中经过数十次迭代，找到 $\nabla^2 f(\mathbf{x}(n+1))$ 的近似不稳定方向。这是因为 $\mathbf{x}(n)$ 接近 $\mathbf{x}(n+1)$，所以 $\nabla^2 f(\mathbf{x}(n))$ 的不稳定方向（我们已得到）可作为求取 $\nabla^2 f(\mathbf{x}(n+1))$ 不稳定方向的良好初始条件。

现在，我们考虑高维情形下的一个高指标鞍点。我们选择 $d = 1000$ 并令 $s_1 = s_2 = s_3 = -1000$, $s_i = 1$ 其中 $i = 4$, ..., 1000。于是，$\mathbf{x}^* = (1,$ ....., $1)^\top$ 是一个条件数较大的指标为 3 的鞍点，$Cond(\nabla^2 f(\mathbf{x}^*)) \approx 722$。在 图 5(b) 中，我们绘制了零阶算法和确定性算法的误差随迭代次数变化的曲线。误差最初沿最陡方向快速衰减，随后速度放缓，这与图 4中观察到的行为相似，但由于海森矩阵的条件数较大，收敛速度更慢。因此，一个有趣且重要的未来研究方向是引入动量来缓解病态性，从而加快收敛。值得注意的是，零阶算法无需计算梯度和海森矩阵，却能达到与确定性算法相当的收敛速度。

5.4. 神经网络的损失景观。 在神经网络中，模型参数有时是隐藏的，无法获取梯度信息，例如零阶攻击 [29]。在这种情况下，通常必须采用无导数方法。我们在线性神经网络的损失景观上实现无导数鞍点搜索算法，以展示它同样适用于退化高指标鞍点。由于神经网络的过参数化，损失函数的大多数临界点都高度退化，即海森矩阵包含许多零特征值。我们考虑一个全连接线性神经

本手稿仅供审阅。

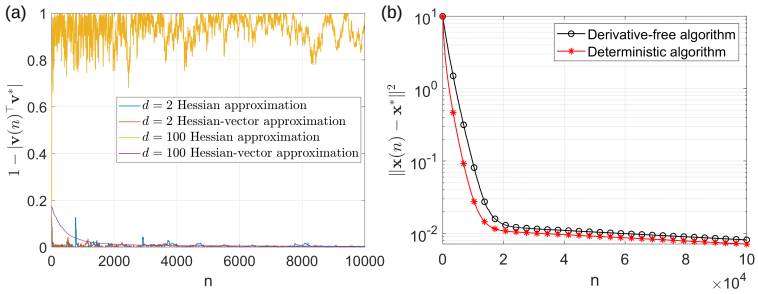

FIG. 5。 *(a)The error plots of the inner eigenvector search algorithm (Algorithm 3 1) with $d=2$。, and $d=100$ using the Hessian approximation ( ) and the Hessian–vector approximation ($H_{\mathbf{v}}$) $\mathbf{Hv}$, respectively. The step size is choosen as α n $=$ $\frac{1}{d\ 10n+10000}$. (b) Error plots of Algorithm 3 2 and ( ) ( )。 the deterministic saddle-search algorithm in 2 2 (。) applied to the modified Rosenbrock function with $d=1000$ $l=0.0001$ $\alpha_{\mathbf{x}}\equiv 0.000001$ $n_{\mathbf{x}}^{max}=100000$ $n_{\mathbf{v}}^{max}=100$ $\alpha_{\mathbf{v}}\equiv 0.0002/d$. , , , , ,*

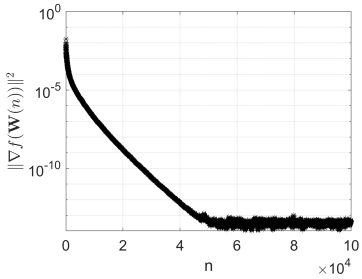

FIG. 6 *Plots of the squared gradient norm $\|\nabla f$ n $\|_2^2$ where n is the iteration points。 ($\mathbf{W}$ ( )) , $\mathbf{W}$ ( ) 。, , of Algorithm 3 2 with $\alpha_{\mathbf{x}}\equiv 0.01$ $l=0.0001$ for the neural network loss function.*

网络 $Net(\mathbf{W},\mathbf{x})$ $=\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_2\mathbf{W}_1\mathbf{x}$，其深度为 H，权重参数为$\mathbf{W}$ $_{i+1}\in\mathbb{R}^{d_{i+1}\times d_i}$，对于 $d_0=d_x, d_H=d_y$。相应的经验损失 f 定义为

$$f(\mathbf{W})=\sum_{i=1}^{N}\|Net(\mathbf{W},\mathbf{x}_i)-\mathbf{y}_i\|_2^2=\|\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_2\mathbf{W}_1\mathbf{X}-\mathbf{Y}\|_F^2,$$

其中 $\mathbf{X}=[\mathbf{x}_1,\cdots,\mathbf{x}_N]$，以及 $\mathbf{Y}=[\mathbf{y}_1,\cdots,\mathbf{y}_N]$，其中 $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^N$为训练数据。若 $d_i=d_0, 1\leqslant i\leqslant H-1$，则 $\mathbf{W}^*=[\mathbf{W}_1^*,\cdots,\mathbf{W}_H^*]$是一个鞍点，其中

$$\mathbf{W}_1^*=\begin{bmatrix} U_{\mathcal{S}}^\top\Sigma_{YX}\Sigma_{XX}^{-1} \\ \mathbf{0}\end{bmatrix}, \mathbf{W}_h^*=\mathbf{I}_{d_0} \text{ for } 2\leqslant h\leqslant H-1, \text{ and } \mathbf{W}_H^*=\begin{bmatrix}\mathbf{U}_{\mathcal{S}},\mathbf{0}\end{bmatrix},$$

且 $\mathcal{S}$是 $\{1,2,\ldots,r_{\max}\}$中针对 $r_{\max}=\min\{d_0,\ldots,d_H\}$, $\Sigma_{XX}=XX^\top$, $\Sigma_{YX}=YX^\top$, $\Sigma=\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{YX}^\top$的指标子集，并且 U满足 $\Sigma=U\Lambda U^\top$，其中 $\Lambda=\text{diag}(\lambda_1,\ldots,\lambda_{d_y})$ [1, 31]。

我们设深度 $H=5$、输入维数 $d_x=10$、输出维数$d_y=4$， $d_i=10$对于 $1\leqslant i\leqslant 4$，以及数据点数量 $N=100$。数据点（$\mathbf{x}_i$，$\mathbf{y}_i$）从标准正态分布中采样。在当前设置下，$\mathbf{W}^*$ 是一个退化鞍点，具有 16 个负特征值和若干个零特征值。在图6中，我们测试了无导数鞍点搜索算法，该算法成功找到了目标鞍点，且具有线性收敛速率。误差仍会趋于一个较小的平台值，约为 10^{-14}；通过减小 l，还可以进一步降低该平台误差。

本手稿仅供审阅。

6. 结论与讨论。 本文聚焦于一种用于鞍点搜索的无导数算法的开发与数值分析，该算法仅需进行函数求值。所提出的无导数鞍点搜索算法由两个主要部分组成：内层特征向量搜索步骤和外层鞍点搜索步骤。对于内层特征向量搜索步骤，我们采用零阶海森矩阵-向量近似来消除维数灾难，并证明该方法几乎处处能在有限次迭代内识别 $\nabla^2 f(\mathbf{x})$ 和 $\nabla^2 f_l(\mathbf{x})$ 的近似不稳定方向。利用这些近似不稳定方向，我们进一步证明，在衰减的差分长度 $l(n)$ 和步长 $\alpha(n)$ 下，外层无导数鞍点搜索步骤能够识别鞍点 $\mathbf{x}^*$ 。当差分长度固定、步长衰减时，外层迭代几乎处处收敛到$\mathbf{x}_l^*$，它是高斯平滑函数的一个鞍点，且与$\mathbf{x}^*$的距离为 $O(l^2)$。此外，在有界性假设下，我们证明，当差分长度和步长均固定时，迭代点以线性收敛速率趋近$\mathbf{x}^*$ 的一个邻域，且该邻域的半径为 $O(l^2/\sqrt{\alpha})$。这些结果表明，当差分长度 l 足够小时，算法输出能够很好地逼近目标鞍点，这一点也在我们的数值实验中得到了进一步验证。

本研究为进一步探索开辟了若干潜在方向。通过引入无导数优化方法中的某些技术，可以改进无导数特征向量搜索算法和无导数鞍点搜索算法的性能。例如，高斯平滑函数的梯度难以估计，因为它涉及一个 d维积分；方向高斯平滑可能为设计梯度估计器提供更高效的方法 [45]。此外，通过引入误差校正项，可以降低估计器的方差 [29, 22], ，从而加快收敛，如（4.3）中所述。此外，将动量引入算法（如 [5],所示）可能会加快收敛，尤其是当鞍点附近的能量景观 $\mathbf{x}^*$ 较为平坦时（或 $\nabla^2 f$（$\mathbf{x}^*$）病态时），例如在 第5.3小节中的修正的罗森布罗克函数情形。一个可能的理论挑战在于，如何确保迭代点始终处于目标鞍点的邻域内，至少以正概率如此。总体而言，由于无导数鞍点搜索算法可以被视为一种梯度估计方法，因此有理由认为，为无导数优化方法开发的许多技术同样可以应用于该算法。然而，鞍点固有的不稳定性以及目标函数的非凸性带来了额外困难。

A. 附录。

A.1. 证明引理 3.3。

Proof. Denote $\mathbf{r}^{\otimes i} = [\underbrace{\mathbf{r}, \cdots, \mathbf{r}}_{i \text{ times}}] := \underbrace{\mathbf{r} \otimes \mathbf{r} \otimes \cdots \otimes \mathbf{r}}_{i \text{ times}}$, for $i \geqslant 2$. We have

$$\nabla f_l(\mathbf{x}) = \mathbb{E}_{\mathbf{r}}[\nabla f(\mathbf{x} + l\mathbf{r})] = \nabla f(\mathbf{x}) + l^2 \mathbb{E}_{\mathbf{r}}[\nabla^3 f(\zeta_{1,l,\mathbf{r}})[\mathbf{r}, \mathbf{r}]]/2,$$
$$\nabla^2 f_l(\mathbf{x}) = \mathbb{E}_{\mathbf{r}}[\nabla^2 f(\mathbf{x} + l\mathbf{r})] = \nabla^2 f(\mathbf{x}) + l^2 \mathbb{E}_{\mathbf{r}}[\nabla^4 f(\zeta_{2,l,\mathbf{r}})[\mathbf{r}, \mathbf{r}]]/2,$$

由泰勒展开可得。因此

$$\|\nabla f_l(\mathbf{x}) - \nabla f(\mathbf{x})\|_2 \leqslant Ml^2 d/2, \ \|\nabla^2 f_l(\mathbf{x}) - \nabla^2 f(\mathbf{x})\|_2 \leqslant Ml^2 d/2.$$

为计算 $F(\mathbf{x}, \mathbf{r}\ l)$ 的二阶矩，我们考虑

$$f(\mathbf{x} \pm l\mathbf{r}) = f(\mathbf{x}) \pm l\nabla f(\mathbf{x})^\top \mathbf{r} + l^2 \mathbf{r}^\top \nabla^2 f(\mathbf{x}) \mathbf{r} 2 \pm l^3 \nabla^3 f(\xi_\pm)[\mathbf{r}, \mathbf{r}, \mathbf{r}]/6,$$

本手稿仅供审阅。

那么， $F(\mathbf{x},\mathbf{r}\ l) = (\nabla f(\mathbf{x})^\top \mathbf{r})\mathbf{r} + M_2\mathbf{r}$，其中 $|M_2| \leqslant \frac{l^2 M \|\mathbf{r}\|_2^3}{6}$. 根据伊瑟利斯定理，

$$\mathbb{E}_{\mathbf{r}}(\|(\nabla f(\mathbf{x})^\top \mathbf{r})\mathbf{r}\|^2) = \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} \nabla f(\mathbf{x})_j \nabla f(\mathbf{x})_k \mathbb{E}(r_i r_i r_j r_k) = (d+2)\|\nabla f(\mathbf{x})\|^2,$$

$$\begin{aligned}\mathbb{E}_{\mathbf{r}}(\|F(\mathbf{x},\mathbf{r},l)\|^2) &\leqslant 2\mathbb{E}_{\mathbf{r}}(\|(\nabla f(\mathbf{x})^\top \mathbf{r})\mathbf{r}\|_2^2) + M^2 l^4 \mathbb{E}_{\mathbf{r}}(\|\mathbf{r}\|^8)/18 \\ &\leqslant (2d+4)\|\nabla f(\mathbf{x})\|_2^2 + M^2 l^4 d(d+2)(d+4)(d+6)/18.\end{aligned}$$

对于海森矩阵近似，我们考虑以下泰勒展开，

$$f(\mathbf{x} \pm l\mathbf{r}) = f(\mathbf{x}) \pm l\nabla f(\mathbf{x})^\top \mathbf{r} + \frac{l^2}{2}\mathbf{r}^\top \nabla^2 f(\mathbf{x})\mathbf{r} \pm \frac{l^3}{6}\nabla^3 f(\mathbf{x})[\mathbf{r}^{\otimes 3}] + \frac{l^4}{24}\nabla^4 f(\xi_\pm)[\mathbf{r}^{\otimes 4}].$$

然后 $\mathbf{H}(\mathbf{x},\mathbf{r}\ l) = \frac{(\mathbf{r}^\top \nabla^2 f(\mathbf{x})\mathbf{r})(\mathbf{r}\mathbf{r}^\top - \mathbf{I})}{2} + M_3(\mathbf{r}\mathbf{r}^\top - \mathbf{I})$，其中 $|M_3| \leqslant \frac{l^2 M \|\mathbf{r}\|_2^4}{24}$. From

$$\begin{aligned}\mathbb{E}_{\mathbf{r}}[\|\mathbf{r}^\top \nabla^2 f(\mathbf{x})\mathbf{r}(\mathbf{r}\mathbf{r}^\top - \mathbf{I})\|_2^2/4] &\leqslant \|\nabla^2 f(\mathbf{x})\|_2^2 (\mathbb{E}_{\mathbf{r}}[\|\mathbf{r}\|^8 + 2\|\mathbf{r}\|^6 + \|\mathbf{r}\|^4])/4 \\ &\leqslant d(d+2)(d^2+12d+33)M^2/4,\end{aligned}$$

我们有

$$\begin{aligned}\mathbb{E}_{\mathbf{r}}[\|\mathbf{H}(\mathbf{x},\mathbf{r},l)\|_2^2] &\leqslant 2\mathbb{E}_{\mathbf{r}}[\|\mathbf{r}^\top \nabla^2 f(\mathbf{x})\mathbf{r}(\mathbf{r}\mathbf{r}^\top - \mathbf{I})\|_2^2/4] + 2\mathbb{E}_{\mathbf{r}}[M_3^2 \|\mathbf{r}\mathbf{r}^\top - \mathbf{I}\|_2^2] \\ &\leqslant d(d+2)(d^2+12d+33)M^2/2 \\ &\quad + M^2 l^4 d(d+2)(d+4)(d+6)(d^2+20d+97)/288.\end{aligned}$$

A.2. 引理的证明 引理 4.5。

证明。

$$\begin{aligned}\mathbb{E}_{\mathbf{r}}[H_{\mathbf{v}}] &= \frac{\mathbb{E}_{\mathbf{r}}[F(\mathbf{x}+l\mathbf{v},\mathbf{r},l)] - \mathbb{E}_{\mathbf{r}}[F(\mathbf{x}-l\mathbf{v},\mathbf{r},l)]}{2l} = \frac{\nabla f_l(\mathbf{x}+l\mathbf{v}) - \nabla f_l(\mathbf{x}-l\mathbf{v})}{2l} \\ &= \nabla^2 f_l(\mathbf{x})\mathbf{v} + \frac{l^2}{12}(\nabla^4 f_l(\xi_+) - \nabla^4 f_l(\xi_-))[\mathbf{v},\mathbf{v},\mathbf{v}],\end{aligned}$$

$$\|\mathbb{E}_{\mathbf{r}}[H_{\mathbf{v}}] - \nabla^2 f(\mathbf{x})\mathbf{v}\|_2 \leqslant \|\nabla^2 f(\mathbf{x}) - \nabla^2 f_l(\mathbf{x})\|_2 + Ml^2/6 \leqslant Ml^2(d/2+6) \leqslant Ml^2 d.$$

根据泰勒展开，我们得到

$$\begin{aligned}H_{\mathbf{v}} &= \frac{F(\mathbf{x}+l\mathbf{v},\mathbf{r},l) - F(\mathbf{x}-l\mathbf{v},\mathbf{r},l)}{2l} \\ &= \mathbf{r}\mathbf{r}^\top \nabla^2 f(\mathbf{x})\mathbf{v} + \frac{l^2}{6}(\nabla^4 f(\mathbf{x})[\mathbf{v},\mathbf{v},\mathbf{v},\mathbf{r}] + \nabla^4 f(\mathbf{x})[\mathbf{v},\mathbf{r},\mathbf{r},\mathbf{r}])\mathbf{r} + O(l^4).\end{aligned}$$

因此，由柯西–施瓦茨不等式可得

$$\mathbb{E}_{\mathbf{r}}(\|H_{\mathbf{v}}\|_2^2) \leqslant M^2(4(d+2) + l^4 d(d+2)/9 + l^4 d(d+2)(d+4)(d+6)/9) + O(l^8 d^8).$$

A.3. 定理 4.10 的证明。

LEMMA 6.1（ [42]命题 4.15）。假设假设3.3、假设3.4 和假设4.1 均成立。设邻域 U $\mathbf{x}^*$ 如(4.7)中所定义。若$\mathbf{x}(n) \in U$，则我们几乎处处有

$$\begin{aligned}\|\mathbf{x}(n+1) - \mathbf{x}^*\|_2^2/2 \leqslant &(1 - \alpha(n)((1-\sqrt{\theta})\mu - M\theta - 5M\sqrt{\theta}))\|\mathbf{x}(n) - \mathbf{x}^*\|_2^2/2 \\ &+ \alpha(n)\psi(n) + \frac{1}{2}\alpha(n)^2 \|F(\mathbf{x}(n),\mathbf{r}(n),l(n))\|_2^2,\end{aligned} \tag{6.1}$$

其中 $\psi(n) = -\langle(\mathbf{I} - 2\tilde{V}(n)\,\tilde{V}(n)^\top)(F(\mathbf{x}(n),\ \mathbf{r}(n)\ l(n)) - \nabla f(\mathbf{x}(n))),\ \mathbf{x}(n) - \mathbf{x}^*\rangle$。

本手稿仅供审阅。

LEMMA 6.2. 假设假设3.1、假设3.2、假设3.3、假设3.4以及假设4.1 均成立。若假设外层鞍点搜索算法采用满足以下条件的步长： $\sum_{n=0}^{\infty}\alpha(n)^2$ 足够小（例如， $\alpha(n) = \gamma_{(n+m)^p}$，其中 m 足够大且 $p\in(1/2,1]$)。令 ϵ 和 U_0 为

(6.2)
$$4\epsilon+2\sqrt{\epsilon}=\min\left(\frac{(1-\sqrt{\theta})\mu}{M}-\theta-5\sqrt{\theta},\delta\right)^2, U_0=\left\{\mathbf{x},\|\mathbf{x}-\mathbf{x}^*\|_2\leqslant\sqrt{2\epsilon}\right\}\subset U,$$

其中 U 在(4.7)中定义。然后，对于由鞍点搜索算法在初始条件$\mathbf{x}(0)\in U_0$下生成的 $\mathbf{x}(n)$，式(4.7) 中定义的事件 E^∞ 以正概率发生。特别地，如果鞍点搜索算法采用形式为 $\alpha(n)=\gamma_{(n+m)^p}$、$p\in(1/2,1]$ 且 m足够大的步长计划运行，那么$\mathbb{P}(E^\infty|\mathbf{x}(0)\in U_0)>1-\epsilon_{E^\infty}$，且当 $m\to\infty$ 时，$1-\epsilon_{E^\infty}\to 1$。

证明。由于引理 6.1 与 [32, 命题 D.1], 类似，而该命题是证明事件 E^∞以正概率发生的关键步骤，因此证明沿用与 [32, 定理 4]相同的思路。故略去细节，仅给出主要步骤。与原证明不同之处在于：(i) 梯度项表现为 $(\mathbf{I}-2\tilde{V}(n)\,\tilde{V}(n)^\top)\nabla f$，而非 ∇f，这不会影响其 L^2 范数；(ii) $\|\nabla f(\mathbf{x})\|_2$, $G=\max_{\mathbf{x}\in U}\|\nabla f(\mathbf{x})\|_2$ 的上界并非全局成立，而仅对 $\mathbf{x}\in U$成立，这不会改变证明步骤；(iii) $\psi(n)=-\langle(\mathbf{I}-2\tilde{V}(n)\,\tilde{V}(n)^\top)(F(\mathbf{x}(n),\mathbf{r}(n)\ l(n))-\nabla f(\mathbf{x}(n))),\mathbf{x}(n)-\mathbf{x}^*\rangle$ 并非零期望的鞅序列，因此我们需要估计有偏零阶梯度估计器引入的误差。

根据引理 6.1，迭代点与鞍点之间的距离几乎处处至多按 $\alpha(n)\psi(n)+\frac{1}{2}\alpha(n)^2\|F(\mathbf{x}(n),\ \mathbf{r}(n)\ l(n))\|_2^2$增长。定义累积均方误差

$$R^n=\left(\sum_{i=0}^{n}\alpha(n)\psi(n)\right)^2+\sum_{i=0}^{n}\frac{1}{2}\alpha(n)^2\|F(\mathbf{x}(n),\mathbf{r}(n),l(n))\|_2^2.$$

设 ϵ 和 $U_0\subset U$ 如（6.2）中所述。则若$\mathbf{x}(0)\in U_0$，且累积均方误差 $R^n<\epsilon$，几乎处处有

$$\begin{aligned}\|\mathbf{x}(n+1)-\mathbf{x}^*\|_2^2&\leqslant\|\mathbf{x}(0)-\mathbf{x}^*\|_2^2+2\sum_{i=0}^{n}\alpha(n)\psi(n)+\sum_{i=0}^{n}\alpha(n)^2\|F(\mathbf{x}(n),\mathbf{r}(n),l(n))\|_2^2\\&\leqslant 2\epsilon+2\sqrt{\epsilon}+2\epsilon=4\epsilon+2\sqrt{\epsilon},\end{aligned}$$

即，$\mathbf{x}(n+1)\in U$. 定义小误差事件 $J^n=\{R_k\leqslant\epsilon,k=1,\cdots,n\}$以及有界性事件 $E^n=\{\mathbf{x}(i)\in U,i=1,2,\cdots,n\}$. 根据 [32, 引理 D.2。], 我们有 $E^{n+1}\subseteq E^n, J^{n+1}\subseteq J^n$, $J^{n-1}\subseteq E^n$，这意味着小误差确保有界性事件发生。不同于 [32, 引理 D.2。] 中使用无偏随机梯度，使得 $\psi(n)$是具有零期望的鞅，$\psi(n)=-\langle(\mathbf{I}-2\tilde{V}(n)\,\tilde{V}(n)^\top)(F(\mathbf{x}(n),\mathbf{r}(n)\ l(n))-\nabla f(\mathbf{x}(n))),\mathbf{x}(n)-\mathbf{x}^*\rangle$ 并不是具有零期望的鞅序列。不过，我们可以利用假设 3.2 中的 $l(n)\leqslant L\alpha(n)$ 来控制这一项。具体而言，根据引理 3.3，如果 $\mathbf{x}(n)\in U$,

$$\begin{aligned}\|\mathbb{E}_{\mathbf{r}}[F(\mathbf{x}(n),\mathbf{r},l(n))]-\nabla f(\mathbf{x}(n))\|_2&\leqslant Ml(n)^2d/2,\\\mathbb{E}_{\mathbf{r}}[\|F(\mathbf{x}(n),\mathbf{r},l(n))\|_2^2]&\leqslant(2d+4)G^2+M^2l(n)^4d(d+2)(d+4)(d+6)/18.\end{aligned}$$

本手稿仅供审阅。

通过将 [32] 中的方程 D.24a 和 D.25 替换为

$$2\alpha(n+1)\mathbb{E}\left[\left(\sum_{i=0}^{n}\alpha(i)\psi(i)\right)\psi(n+1)\mathbf{1}_{J^n}\right]\leqslant\sqrt{\epsilon}\sqrt{4\epsilon+2\sqrt{\epsilon}}ML^2\alpha(n+1)^2d,$$

并按照 [32, 引理 D.2], 中的相同步骤，我们得到

$$\epsilon\mathbb{P}(J^n\setminus J^{n+1})\leqslant\mathbb{E}(R^n\mathbf{1}_{J^{n-1}})-\mathbb{E}(R^{n+1}\mathbf{1}_{J^n})+R^*\alpha(n+1)^2, \tag{6.3}$$

其中 $R^0\mathbf{1}_{J^{-1}}=0$, $l=\max_n l(n)$,

$$R^*=\sqrt{\epsilon}\sqrt{4\epsilon+2\sqrt{\epsilon}}ML^2d+\left(\frac{1}{2}+8\epsilon+4\sqrt{\epsilon}\right)\left[(2d+4)G^2+\frac{M^2l^4d_4}{18}+G^2\right],$$

并且 $d_4=d(d+2)(d+4)(d+6)$。对（6.3）求和后，我们得到

$$\epsilon\mathbb{P}\left(\bigcup_{i=0}^{n}(J^i\setminus J^{i+1})\right)=\epsilon\sum_{i=0}^{n}\mathbb{P}(J^i\setminus J^{i+1})\leqslant\sum_{n=0}^{\infty}R^*\alpha(n)^2.$$

那么，对于 E^∞ 由（4.7）定义的量，我们有

$$\begin{aligned}\mathbb{P}(E^\infty)&=\inf_n\mathbb{P}(E^{n+1})\geqslant\inf_n\mathbb{P}(J^n)=\inf_n\mathbb{P}\left(\bigcap_{i=0}^{n}(J^i\setminus J^{i+1})^c\right)\\&=\inf_n\mathbb{P}\left(\Omega\setminus\bigcup_{i=0}^{n}(J^i\setminus J^{i+1})\right)\geqslant 1-R^*\sum_{n=0}^{\infty}\frac{\alpha(n)^2}{\epsilon}.\end{aligned}$$

因此，如果 $\sum_{n=0}^{\infty}\alpha(n)^2$ 足够小，使得 $1-R^*\sum_{n=0}^{\infty}\alpha(n)^2>0$ 成立，且初始距离不太大（即 $\mathbf{x}(0)\in U_0$），那么我们可以得出结论：E^∞ 以正概率发生。□

定理 4.10 的证明。由引理 6.2 可知，有界性事件以正概率发生。定义事件 $E^n=\{\mathbf{x}(i)\in U, i=1,2,\cdots,n\}$，其中 U 在 (4.7) 中定义；相应的指示函数为 $\mathbf{1}_{E^n}$，并令 $G=\max_{\mathbf{x}\in U}\|\nabla f(\mathbf{x})\|_2$。注意到 $l(n)\leqslant L\sqrt{\alpha(n)}$（假设 3.2），由引理 6.1 可得

$$\begin{aligned}&\mathbb{E}\left(\mathbf{1}_{E^{n+1}}\|\mathbf{x}(n+1)-\mathbf{x}^*\|_2^2-\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2^2|\mathcal{F}(n)\right)\\&\leqslant\mathbb{E}\left(\mathbf{1}_{E^n}\|\mathbf{x}(n+1)-\mathbf{x}^*\|_2^2-\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2^2|\mathcal{F}(n)\right)\\&\leqslant-\alpha(n)((1-\sqrt{\theta})\mu-M\theta-5M\sqrt{\theta}))\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2^2\\&\quad+2\alpha(n)\frac{Ml(n)^2d}{2}\frac{(1-\sqrt{\theta})\mu-M\theta-5M\sqrt{\theta}}{M}\\&\quad+\alpha(n)^2\left((2d+4)G^2+M^2l(n)^4d(d+2)(d+4)(d+6)/18\right)\\&\leqslant-\alpha(n)((1-\sqrt{\theta})\mu-M\theta-5M\sqrt{\theta}))\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2^2+M_4\alpha(n)^2,\end{aligned}$$

其中，$M_4>0$ 是依赖于 d、μ、M、G、θ、γ、m、p、L的常数。应用引理 3.6，我们得出结论：$\mathbf{1}_{E^\infty}\|\mathbf{x}(n)-\mathbf{x}^*\|_2^2\leqslant\mathbf{1}_{E^n}\|\mathbf{x}(n)-\mathbf{x}^*\|_2^2$ 几乎处处收敛到零。因此，在给定 E^∞的条件下，算法 3.2 几乎处处经过有限次迭代即可到达 $\mathbf{x}^*$ 的$\epsilon_{\mathbf{x}}$-邻域。□

本手稿仅供审阅。

[1] E. M. Achour, F. Malgouyres, and S. Gerchinovitz, *The loss landscape of deep linear neural networks: a second-order analysis*, 《机器学习研究杂志》, 25（2024）, 第1–76页。
[2] C. Audet and W. Hare, *Introduction: tools and challenges in derivative-free and blackbox optimization* 2017, 3 14 , 载于《无导数与黑盒优化》, 施普林格, 第–页。[3] K. Balasubramanian and S. Ghadimi, *Zeroth-order (non)-convex stochastic optimizationwia conditional gradient and gradient updates*, 神经信息处理系统进展, 31（2018）。[4] S. Bonfanti and W. Kob, *Methods to locate saddle points in complex landscapes*, 《化学物理学杂志》, 147（2017）。[5] X. Chen, S. Liu, K. Xu, X. Li, X. Lin, M. Hong, and D. Cox, *Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization*, 神经信息处理系统进展, 32（2019）。[6] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to derivative-free optimization*, 工业与应用数学学会, 2009年。[7] A. Cutkosky and F. Orabona, *Momentum-based variance reduction in non-convex SGD*, 神经信息处理系统进展, 32（2019）。[8] S. N. Deming, L. R. Parker Jr, and M. Bonner Denton, *A review of simplex optimization in analytical chemistry* 7 1978 187 , 《分析化学评论》, （）, 第–202页。[9] W. E, W. Ren, and E. Vanden-Eijnden, *String method for the study of rare events*, 《物理评论B》, 66（2002）, 第052301页。[10] W. E, W. Ren, and E. Vanden-Eijnden, *Simplified and improved string method for computingthe minimum energy paths in barrier-crossing events*, 《化学物理学杂志》, 126（2007）, 第164103页。[11] W. E and E. Vanden-Eijnden, *Transition-path theory and path-finding algorithms for the study of rare events* 61 2010 391 420 , 《物理化学年度评论》, （）, 第–页。[12] W. E and X. Zhou, *The gentlest ascent dynamics*, 《非线性》, 24（2011）, 第1831页。[13] B. Eckhardt, *Irregular scattering*, 《物理学D辑：非线性现象》, 33（1988）, 第89–98页。[14] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, *Bilevel programming for hyperparameter optimization and meta-learning*, 载于国际机器学习会议, 机器学习研究论文集, 2018年, 第1568–1577页。[15] G. Fredrickson, *The equilibrium theory of inhomogeneous polymers*, 国际物理学专著丛书第134号, 牛津大学出版社, 2006年。[16] X. Gao, B. Jiang, and S. Zhang, *On the information-adaptive variants of the admm: an iteration complexity perspective* 76 2018 327 363 , 《科学计算杂志》, （）, 第–页。[17] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, 第1卷, 麻省理工学院出版社（剑桥）, 2016年。[18] E. F. Gramsbergen, L. Longa, and W. H. de Jeu, *Landau theory of the nematic-isotropic phase transition* 135 1986 195 257 , 《物理学报告》, （）, 第–页。[19] G. A. Gray, T. G. Kolda, K. Sale, and M. M. Young, *Optimizing an empirical scor-ing function for transmembrane protein structure determination*, 《INFORMS计算杂志》, 16（2004）, 第406–418页。[20] G. Henkelman and H. Jónsson, *A dimer method for finding saddle points on high dimensionalpotential surfaces using only first derivatives*, 《化学物理学杂志》, 111（1999）, 第7010–7022页。[21] K. G. Jamieson, R. Nowak, and B. Recht, *Query complexity of derivative-free optimization*, 神经信息处理系统进展, 25（2012）。[22] K. Ji, Z. Wang, Y. Zhou, and Y. Liang, *Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization*, 载于国际机器学习会议, 机器学习研究论文集, 2019年, 第3100–3109页。[23] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction* 26 2013 , 神经信息处理系统进展, （）。[24] C. T. Kelley, *Iterative methods for linear and nonlinear equations*, 工业与应用数学学会, 1995年。[25] J. Larson, M. Menickelly, and S. M. Wild, *Derivative-free optimization methods*, 《数值学报》, 28（2019）, 第287–404页。[26] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, 《自然》, 521（2015）, 第436–444页。[27] Y. Li, Y. Tang, R. Zhang, and N. Li, *Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach*, 《IEEE自动控制汇刊》, 67（2021）, 第6429–6444页。[28] R. Lifshitz and H. Diamant, *Soft quasicrystals–why are they stable?*, 《哲学杂志》, 87（2007）, 第3021–3030页。

本手稿仅供审阅使用。

[29] S. Liu, B. Kailkhura, P. Y. Chen, P. Ting, S. Chang, and L. Amini, *Zeroth-order stochastic variance reduction for nonconvex optimization*, 神经信息处理系统进展, 31（2018年）。[30] W. Liu, Z. Xie, and Y. Yuan, *A constrained gentlest ascent dynamics and its applications to finding excited states of Bose–Einstein condensates*, Journal of Computational Physics, 473（2023年）, 第111719页。[31] Y. Luo, L. Zhang, and X. Zheng, *Accelerated high-index saddle dynamics method for searching high-index saddle points*, 102 2025 31 , 《科学计算杂志》, （）, 第 页。[32] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher, *On the almost sure convergenceof stochastic gradient descent in non-convex problems*, 神经信息处理系统进展, 33（2020年）, 第1117–1128页。[33] M. Mezei and D. L. Beveridge, *Free energy simulations*, 纽约科学院年鉴, 482（1986年）, 第1–23页。[34] R. M. Minyaev, W. Quapp, G. Subramanian, P. V. R. Schleyer, and Y. Mo, *Internalconrotation and disrotation in H2BCH2BH2 and diborylmethane 1, 3 H exchange*, Journal of Computational Chemistry, 18（1997年）, 第1792–1803页。[35] Y. Nesterov and V. Spokoiny, *Random gradient-free minimization of convex functions*, 计算数学基础, 17（2017年）, 第527–566页。[36] J. Nocedal and S. J. Wright, *Numerical optimization*, 施普林格, 1999年。[37] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, *Theory of protein folding: the energy landscape perspective*, 48 1997 545 , 《物理化学年度评论》, （）, 第 – 600 页。[38] W. Quapp and J. M. Bofill, *Locating saddle points of any index on potential energy surfaces by the generalized gentlest ascent dynamics*, 理论化学学报, 133（2014年）, 第1–14页。[39] L. M. Rios and N. V. Sahinidis, *Derivative-free optimization: a review of algorithms and comparison of software implementations*, 56 2013 1247 , 全球优化杂志, （）, 第 – 1293 页。[40] O. Shamir, *An optimal algorithm for bandit and zero-order convex optimization with two-point feedback*, 18 2017 1 11 , 机器学习研究杂志, （）, 第 – 页。[41] Y. W. Shang and Y. H. Qiu, *A note on the extended Rosenbrock function*, 进化计算, 14（2006年）, 第119–126页。[42] B. Su, L. Zhang, and Q. Du, *A stochastic algorithm for searching saddle points with convergence guarantee*, 2510 14144, 2025 , arXiv 预印本 arXiv: .（）。[43] J. Snoek, H. Larochelle, and R. P. Adams, *Practical bayesian optimization of machine learning algorithms*, 25 2012 , 神经信息处理系统进展, （）。[44] H. Su, H. Wang, L. Zhang, J. Zhao, and X. Zheng, *Improved high-index saddle dynamics for finding saddle points and solution landscape*, SIAM 数值分析杂志, 63（2025年）, 第1757–1775页。[45] H. Tran, Q. Du, and G. Zhang, *Convergence analysis for a nonlocal gradient descent method via directional gaussian smoothing*, 计算优化与应用, 90（2025年）, 第481–513页。[46] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, *Navigating the folding routes*, 科学, 267（1995年）, 第1619–1620页。[47] J. Yin, Y. Wang, J. Z. Chen, P. Zhang, and L. Zhang, *Construction of a pathway map on a complicated energy landscape*, 124 2020 090601 , 物理评论快报, （）, 第 页。[48] J. Yin, L. Zhang, and P. Zhang, *High-index optimization-based shrinking dimer methodfor finding high-index saddle points*, SIAM 科学计算杂志, 41（2019年）, 第A3576–A3595页。[49] Y. Yu, T. Wang and R. J. Samworth *A useful variant of the davis–kahan theorem for , , ,* 生物统计学, *statisticians* 102 2015 315 323 （）, 第 – 页。[50] J. Zhang and Q. Du, *Shrinking dimer dynamics and its applications to saddle point search*, SIAM 数值分析杂志, 50（2012年）, 第1899–1921页。[51] L. Zhang, Q. Du, and Z. Zheng, *Optimization-based shrinking dimer method for finding transition states*, 38 2016 528 544 , SIAM 科学计算杂志, （）, 第A –A 页。[52] L. Zhang, W. Ren, A. Samanta, and Q. Du, *Recent developments in computational modelling of nucleation in phase transformations*, 2 2016 1 9 , NPJ 计算材料, （）, 第 – 页。[53] L. Zhang, P. Zhang, and X. Zheng, *Error estimates for euler discretization of high-index saddle dynamics*, 60 2022 2925 2944 , SIAM 数值分析杂志, （）, 第 – 页。[54] L. Zhang, P. Zhang, and X. Zheng, *Discretization and index-robust error analysis for constrained high-index saddle dynamics on the high-dimensional sphere*, 中国科学: 数学, 66（2023年）, 第2347–2360页。

本手稿仅供审阅。