arXiv:2602.01627v1 [math.OC] 2 Feb 2026

为什么自适应零阶优化有效?

Haishan Ye * Luo Luo †

2026年2月3日

摘要

零阶（ZO）优化在实际应用中十分普遍，因为获取梯度信息的成本高昂或无法获取。近年来，通过函数值的经验标准差对梯度估计量进行归一化的自适应零阶方法展现出出色的实际性能，尤其是在大语言模型微调中。然而，这种策略的理论理解仍然有限。在本文中，我们证明，以高概率而言，经验标准差与（随机）梯度的范数近似成正比。基于这一洞见，我们在相对于矩阵范数的广义（L_0，L_1）-光滑性条件下分析自适应零阶方法。我们针对确定性和随机情形分别给出了明确的收敛率和查询复杂度界，证明了与采用固定步长的基础零阶优化方法相比，自适应零阶方法具有更快的收敛速度和更高的查询效率。

1 引言

零阶（ZO）优化已成为现代机器学习的基础支柱。在无法获取梯度信息或获取成本高得难以承受的各种应用中，零阶优化得到了广泛应用，包括黑盒对抗攻击（Ilyas 等人，2018；Chen 等人，2017）、强化学习（Chrabaszcz 等人，2018；Salimans 等人，2017；Mania 等人，2018）、内存高效的大语言模型（LLM）微调（Malladi 等人，2023；Dang 等人，2025）等。普通零阶算法使用一批函数值查询来估计梯度。具体而言，梯度估计量的构造如下：

$$\mathbf{g}(\boldsymbol{x}) = \sum_{i=1}^{\ell} \frac{f(\boldsymbol{x} + \alpha \boldsymbol{s}_i) - f(\boldsymbol{x})}{\alpha} \cdot \boldsymbol{s}_i,$$

其中，$f(\boldsymbol{x})$ 是目标函数，ℓ 是批量大小，$\alpha > 0$ 是平滑参数，而$\boldsymbol{s}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d/\ell)$ 中的 $\boldsymbol{I}_d$ 是 $d \times d$ 单位矩阵。

*西安交通大学；邮箱：hsye cs@outlook.com
†复旦大学；邮箱：luoluo@gmail.com

当最小化随机目标时，由于只能获取噪声函数评估 $f(\boldsymbol{x};\ \xi)$，相应的随机梯度估计器为

$$\hat{\mathbf{g}}(\boldsymbol{x}) = \sum_{i=1}^{\ell} \frac{f(\boldsymbol{x} + \alpha \boldsymbol{s}_i;\ \xi) - f(\boldsymbol{x};\ \xi)}{\alpha} \cdot \boldsymbol{s}_i.$$

使用任一种估计器时，普通零阶算法（Nesterov与Spokoiny，2017；Ghadimi与Lan，2013）按如下方式更新迭代点

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \mathbf{g}(\boldsymbol{x}_t) \qquad \text{or} \qquad \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \hat{\mathbf{g}}(\boldsymbol{x}_t),$$

其中 $0 < \eta$ 表示步长。这种普通零阶方案已成功应用于黑盒对抗攻击（Ilyas 等人，2018）以及内存高效的大语言模型调优（Malladi 等人，2023）。

为进一步提升基础零阶优化方法的性能，一种常见的改进是通过用采样函数值 σ_t的 ℓ 标准差对估计梯度进行归一化，来自适应地控制步长（见公式（2.5）和公式（2.7））。由此得到的更新规则形式如下

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \cdot \frac{\mathbf{g}(\boldsymbol{x}_t)}{\sigma_t} \qquad \text{or} \qquad \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \cdot \frac{\hat{\mathbf{g}}(\boldsymbol{x}_t)}{\sigma_t}. \tag{1.1}$$

这种归一化背后的直觉是，较小的标准差 σ_t 表明当前迭代点 $\boldsymbol{x}_t$ 处于景观中相对平坦的区域，因此较大的有效步长更为理想。

从实证结果来看，这种自适应零阶策略相比原始零阶方法取得了显著改进，尤其是在大规模语言模型应用中。例如，Dang 等人（2025）将公式（1.1）中的自适应零阶更新应用于大语言模型微调，结果表明其查询效率显著高于 MeZO（Malladi 等人，2023），后者采用固定步长的普通零阶方法。近期，自适应零阶方法在基于人类反馈的强化学习中也展现出良好的性能（Qiu等人，2025）。

尽管取得了这些实证成功，但对于自适应零阶方法为何优于基础零阶优化方法，目前仍缺乏深入的严格理论理解。本文旨在弥合这一空白，为自适应零阶方法的优势提供有理有据的理论解释。我们的主要贡献概括如下。

- 首先，我们证明，标准差 σ_t 以高概率近似正比于（随机）梯度的范数。这一结果从形式上证实了这样的直觉：较小的 σ_t 意味着当前迭代点周围的景观较为平坦。

- 其次，我们分析了自适应零阶方法在广义（L_0，L_1）光滑函数上的收敛速率和查询复杂度，其中光滑性是相对于 $\boldsymbol{H}$-范数而言的，$\boldsymbol{H}$ 其中 $d \times d$是一个半正定矩阵，而非标准的欧几里得范数。原始（L_0，L_1）光滑性的概念最早由 Zhang 等人(2020b) 提出，用以展示梯度裁剪在现实应用中的优越性。在本文中，我们进一步推广了这一设定，以适用于 $\boldsymbol{H}$-范数光滑性。

- 最后，我们针对广义（L_0，L_1）光滑函数，分别在确定性和随机情形下，为自适应零阶方法建立了明确的收敛率和查询复杂度界。我们的结果表明，自适应零阶方法的收敛速度可以任意快于固定步长原始零阶方法。因此，自适应零阶方法的查询复杂度可以任意小于固定步长原始零阶方法的查询复杂度。关于这些改进的详细讨论见注释 3 和注释 6。

1.1 相关工作

自适应零阶方法。自适应零阶方法曾被多次独立地重新发现。Salimans 等人（2017）采用了与公式（1.1）类似的归一化方式，并将其称为神经网络策略的虚拟批归一化。Mania 等人（2018）在 Salimans 等人（2017）的工作基础上进行改进，并加入二阶信息以实现更快的收敛速率。Lyu 和 Tsang（2019）提出了一种名为 INGO 的算法，该算法与自适应零阶方法类似，但利用了二阶信息。此外，Lyu 和 Tsang（2019）还表明，即使目标函数是非凸的，INGO 也会收敛到一个驻点。最近，Dang 等人（2025）提出了 FZOO，这也是一种自适应零阶方法，但它使用拉德马赫随机向量而非高斯向量作为搜索方向。Dang 等人（2025）首先意识到，标准差 σ_t 与（随机）梯度范数有关，并证明$\mathbb{E}\left[\sigma_t^2\right]$ 几乎与（随机）梯度范数的平方成正比。然而， $\mathbb{E}\left[\sigma_t^2\right]$ 的性质无法直接推导出 σ_t 应与（随机）梯度范数成正比。此外，Dang 等人（2025）仅证明了 FZOO 会收敛到一个驻点，却没有提供充分的理论来说明 FZOO 会优于原始零阶方法。

松弛光滑性。训练神经网络中的实证观察表明，相较于标准光滑性，松弛光滑性能够更好地刻画目标函数（Zhang 等人，2020b；Crawshaw 等人，2022；Cooper，2022）。近年来，松弛光滑性条件下优化问题的（随机）一阶方法已得到广泛研究（Chen 等人，2023；Vankov 等人，2025；Li 等人，2024，2023；Gaash 等人，2025；Gorbunov 等人，2024；Koloskova 等人，2023；Chezhegov 等人，2025；Tovmasyan 等人，2025；Crawshaw 和 Liu，2025；Lobanov 等人，2024；Tyurin，2025；Reisizadeh 等人，2025；Zhang 等人，2020a；Yu 等人，2025；Khirirat 等人，2024b，a；Jiang 等人，2025；Borodich 和 Kovalev，2025）。值得注意的是，在随机梯度下降（SGD）中加入归一化或裁剪步骤，可以高效地找到松弛光滑函数的近似驻点；当初始化较差时，其速度可能任意快于采用固定步长的标准 SGD（Zhang 等人，2020b）。此外，Xie 等人（2024）引入了信赖域方法，并额外假设二阶光滑性条件，以进一步提高收敛速度。近期，Lobanov 和 Gasnikov（2025）采用均匀平滑，构造了裁剪归一化梯度下降的零阶变体；当批量大小足够大时，这些方法对凸目标函数实现了线性收敛率。

2 预备知识

本节首先形式化介绍符号和问题设定，然后介绍零阶优化的背景。

2.1 符号说明

我们介绍本文中使用的符号。若对任意非零向量 $\boldsymbol{x} \in \mathbb{R}^d$，均有 $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$，则称对称矩阵 $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ 为半正定矩阵。我们令 $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ 为正定矩阵 $\boldsymbol{A}$ 的特征值分解，其中 $\boldsymbol{U}$ 是一个标准正交矩阵，包含 $\boldsymbol{A}$ 的特征向量；$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots\ldots, \lambda_d)$ 是一个对角矩阵，且 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$。本文使用 $\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^d A_{ii}$ 表示矩阵 $\boldsymbol{A}$ 的迹。给定两个向量 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$，我们用 $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{i=1}^d x_i y_i$ 表示它们的内积。给定半正定对称矩阵 $\boldsymbol{A}$，我们令 $\|\boldsymbol{A}\|_F \triangleq (\sum_{i,j} A_{ij}^2)^{1/2} = (\sum_i \lambda_i^2)^{1/2}$ 为其弗罗贝尼乌斯范数，并令 $\|\boldsymbol{A}\| \triangleq \lambda_1$ 为其谱范数。对于给定向量 $\boldsymbol{x} \in \mathbb{R}^d$，我们将其欧几里得范数和 $\boldsymbol{A}$-范数分别定义为 $\|\boldsymbol{x}\| \triangleq \sqrt{\sum_{i=1}^d x_i^2}$ 和 $\|\boldsymbol{x}\|_{\boldsymbol{A}} \triangleq \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$。

2.2 问题设定

首先，我们考虑以下确定性最小化问题

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}), \tag{2.1}$$

其中，给定 $\boldsymbol{x} \in \mathbb{R}^d$，零阶算法可以访问函数值 $f(\boldsymbol{x})$。我们还研究随机形式

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) = \mathbb{E}_\xi \left[f(\boldsymbol{x};\xi)\right], \tag{2.2}$$

其中，给定 $\boldsymbol{x} \in \mathbb{R}^d$，零阶算法可以抽取随机变量 ξ 以获取随机函数值。

接下来，我们对上述优化问题提出假设。

假设 1. 我们假设可微函数 $f: \mathbb{R}^d \to \mathbb{R}$ 关于半正定矩阵 $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ 是 (L_0, L_1)-光滑的，也就是说，若对于所有满足 $\|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{H}} \leq 1/(L_1\sqrt{\|\boldsymbol{H}\|})$ 的 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$，均有

$$\left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle\right| \leq \frac{L_0 + L_1\|\nabla f(\boldsymbol{x})\|}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{H}}^2. \tag{2.3}$$

假设 2。 我们假设目标函数 $f: \mathbb{R}^d \to \mathbb{R}$ 下有界，即满足 $f^* = \inf_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) > -\infty$。

假设3 (Zhang 等人 (2020b))。 我们假设公式（$\nabla f(\boldsymbol{x};\ \xi)$ （2.2）中的随机梯度具有有界偏差，即满足 $\|\nabla f(\boldsymbol{x};\ \xi) - \nabla f(\boldsymbol{x})\| \leq \nu$ 对所有 $\boldsymbol{x} \in \mathbb{R}^d$ 均成立，其中 $\nu > 0$ 是一个常数。

注意，假设1 是对（L_0, L_1）-光滑性的推广，由Zhang 等人(2020a）提出，因为我们使用 $\boldsymbol{H}$-范数来描述该条件下的距离。显然，假设1在令L_0, L_1-光滑性时，会退化为标准的（ $\boldsymbol{H}$ 为单位矩阵。事实上，假设1对自适应零阶方法的分析至关重要，也有助于说明自适应零阶方法能够实现比原始零阶方法快得多的收敛速率。

2.3 无关随机草图矩阵

在本小节中，我们介绍无关随机草图矩阵的概念。值得注意的是，近期的研究（Ye等人，2025）表明，从随机草图矩阵中抽取的零阶（ZO）方法搜索方向，可以帮助零阶方法实现弱维度依赖的查询复杂度（参见备注 2和备注 5）。

定义 1 （矩阵乘积草图）。给定一个矩阵 $\boldsymbol{A} \in \mathbb{R}^{d\times d}$、参数 $\gamma \in (0,1], k \geq 1$ 和 $\delta > 0$，若矩阵 $\boldsymbol{S} \in \mathbb{R}^{d\times \ell}$ 与 $\boldsymbol{A}$ 独立，且以下条件以至少 $1-\delta$ 的概率成立，则称 $\boldsymbol{S} \in \mathbb{R}^{d\times \ell}$ 是用于矩阵乘积的(γ, k, δ)-随机草图矩阵：

$$\left\|\boldsymbol{A}^\top \boldsymbol{S}\boldsymbol{S}^\top \boldsymbol{A} - \boldsymbol{A}^\top \boldsymbol{A}\right\| \leq \gamma \left(\|\boldsymbol{A}\|^2 + \frac{\|\boldsymbol{A}\|_F^2}{k}\right). \tag{2.4}$$

- **高斯草图矩阵：**最经典的草图矩阵是高斯草图矩阵 $\boldsymbol{S} \in \mathbb{R}^{\ell\times d}$，其元素独立同分布于均值为0、方差为 $1/\ell$ 的单变量高斯分布。由于其具有众所周知的集中性质（Woodruff 等人，2014），高斯随机矩阵极具吸引力。只需 $\ell = \Omega\left((k+\log(1/\delta))\gamma^{-2}\right)$，即可满足定义 1 中定义的（$\gamma$， k，δ）-近似矩阵乘积性质（Cohen 等人，2016）。

- **拉德马赫草图化：** 作为一种离散替代方案，拉德马赫矩阵使用独立同分布的条目，其取值来自$\{\pm 1/\sqrt{\ell}\}$。它与高斯矩阵具有几乎相同的样本复杂度界，但随机比特消耗更少、效率更高。近期研究还表明，它能够有效加速大型语言模型微调中的零阶优化（Dang 等人，2025）。

- **SRHT：** 通过 SRHT 实现的结构化草图化，形式化表示为 $\boldsymbol{S} = \boldsymbol{P}\boldsymbol{H}_d\boldsymbol{D}$，旨在弥合统计效率与计算速度之间的差距（Woodruff 等人，2014）。矩阵$\boldsymbol{H}_d \in \mathbb{R}^{d\times d}$ 是沃尔什–哈达玛矩阵，其中 $+1$ 和 -1 为条目。 $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ 是一个对角矩阵，其对角元素从 $\{+1,-1\}$中均匀采样；而 $bP \in \mathbb{R}^{s\times d}$ 是均匀采样矩阵。尽管所需的 ℓ比高斯矩阵略高（相差对数因子），但快速沃尔什–哈达玛变换使 $\boldsymbol{S}\boldsymbol{A}$ 能够在 $\mathcal{O}(d^2\log\ell)$ 内计算（对于 $d\times d$ 矩阵），从而显著加快稠密数据的处理速度。

- **稀疏嵌入：**这些矩阵通过每列仅保留 $s \ll \ell$ 个非零元素来优化稀疏性。根据Cohen (2016)，草图大小为 $\ell = \Omega(k\log(k/\delta)\gamma^{-2})$，稀疏性为 $s = \mathcal{O}(\log(k/\delta)\gamma^{-1})$ 时，足以满足 (γ, k, δ)-性质。这种构造对于高维稀疏数据尤其有效，因为在此类数据中，保持稀疏性至关重要。

算法 1 确定性优化的自适应零阶方法

1: **输入：** 初始向量 x_0、样本量 ℓ 和步长 η。2: **对于** $t = 0, 1, 2, \ldots, T-1$ **执行**3: 生成一个 (γ, k, δ)-无偏草图矩阵 $\boldsymbol{S}_t \in \mathbb{R}^{d\times\ell}$，用于与 $\boldsymbol{s}_i$ 进行矩阵乘积，其中 $\boldsymbol{s}_i$ 为 $\boldsymbol{S}_t$ 的第 i 列。4: 访问 $f(\cdot)$ 的值并构造近似梯度

$$\mathbf{g}(\boldsymbol{x}_t) = \sum_{i=1}^{\ell} \frac{f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_i) - f(\boldsymbol{x}_t)}{\alpha} \boldsymbol{s}_i.$$

5: 计算标准差:

$$\sigma_t = \sqrt{\frac{1}{\ell-1}\sum_{j=1}^{\ell}\left(f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_j) - \frac{1}{\ell}\sum_{i=1}^{\ell} f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_i)\right)^2} \tag{2.5}$$

6: 更新:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \cdot \frac{\mathbf{g}(\boldsymbol{x}_t)}{\sigma_t} \tag{2.6}$$

7: **结束循环**8: **输出：** $\boldsymbol{x}_T$

2.4 自适应零阶方法的算法描述

设 $\boldsymbol{S} \in \mathbb{R}^{d\times\ell}$ 为一个预定义矩阵，$\boldsymbol{s}_i$ 表示其第 i列。我们通过以下方式近似给定点 $\boldsymbol{x}_t$ 处 $f : \mathbb{R}^d \to \mathbb{R}$ 的梯度:

$$\mathbf{g}(\boldsymbol{x}_t) = \sum_{i=1}^{\ell} \frac{f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_i) - f(\boldsymbol{x}_t)}{\alpha} \cdot \boldsymbol{s}_i, \tag{2.8}$$

是 $\alpha > 0$。对于随机最小化问题（2.2），我们只能访问带噪声的函数求值 $f(\cdot; \xi_t)$ 来近似随机梯度 $\nabla f(\boldsymbol{x}_t; \xi_t)$，其形式如下:

$$\hat{\mathbf{g}}(\boldsymbol{x}_t) = \sum_{i=1}^{\ell} \frac{f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_i; \xi_t) - f(\boldsymbol{x}_t; \xi_t)}{\alpha} \boldsymbol{s}_i. \tag{2.9}$$

此外，t 点 $\boldsymbol{x}_t + \alpha \boldsymbol{s}_i$ 处（含噪）函数评估的标准差为

$$\sigma_t = \sqrt{\frac{1}{\ell-1}\sum_{j=1}^{\ell}\left(f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_j) - \frac{1}{\ell}\sum_{i=1}^{\ell} f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_i)\right)^2}$$

算法 2 确定性优化的自适应零阶方法

1: **输入：** 初始向量 x_0、样本量 ℓ 和步长 η。2: **对于** $t = 0, 1, 2, \ldots\ldots, T-1$ **执行**3: 生成一个 (γ, k, δ)-无偏草图矩阵 $\boldsymbol{S}_t \in \mathbb{R}^{d\times\ell}$，用于与 $\boldsymbol{s}_i$ 进行矩阵乘积，其中 $\boldsymbol{S}_t$ 的第 i 列为 $\boldsymbol{s}_i$。4: 获取 $f(\cdot)$ 的值，并构造近似梯度

$$\hat{\mathbf{g}}(\boldsymbol{x}_t) = \sum_{i=1}^{\ell} \frac{f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_i;\ \xi_t) - f(\boldsymbol{x}_t;\ \xi_t)}{\alpha} \boldsymbol{s}_i.$$

5: 计算标准差:

$$\sigma_t = \sqrt{\frac{1}{\ell-1}\sum_{j=1}^{\ell}\left(f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_j;\ \xi_t) - \frac{1}{\ell}\sum_{i=1}^{\ell} f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_i;\ \xi_t)\right)^2} \tag{2.7}$$

6: 更新为:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \cdot \frac{\hat{\mathbf{g}}(\boldsymbol{x}_t)}{\sigma_t + \beta}$$

7: **结束循环**

8: **输出：** $\boldsymbol{x}_T$

对于确定性情形，以及

$$\sigma_t = \sqrt{\frac{1}{\ell-1}\sum_{j=1}^{\ell}\left(f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_j;\ \xi_t) - \frac{1}{\ell}\sum_{i=1}^{\ell} f(\boldsymbol{x}_t + \alpha \boldsymbol{s}_i;\ \xi_t)\right)^2}.$$

对于随机情形。

一旦获得标准 差，我们就可以对确定性情形进行更新，具体如下 s

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \cdot \frac{\mathbf{g}(\boldsymbol{x}_t)}{\sigma_t},$$

其中 η 为步长。类似地，我们对随机情形进行如下更新

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \cdot \frac{\mathbf{g}(\boldsymbol{x}_t)}{\sigma_t + \beta},$$

其中 β 为正常数。

确定性和随机优化的自适应零阶方法的详细算法描述分别见算法 1和算法2。

3 确定性问题的收敛性分析

本节介绍算法1 针对确定性问题（2.1）的收敛性分析。

3.1 事件

假设目标函数 $f(\boldsymbol{x})$ 满足假设 1，且 $\boldsymbol{S}_t$ 是一个无偏的 $\left(\frac{1}{4},\ k,\ \delta\right)$- 随机草图矩阵，定义与定义 1 中的相同，其中 $4 \leq k$ 且 $0 < \delta < 1$。接下来，我们将定义以下事件

$$\mathcal{E}_{t,1} := \left\{ \sum_{i=1}^{\ell} |\boldsymbol{s}_i^\top \boldsymbol{H} \boldsymbol{s}_i|^2 \leq \zeta_1 \mid \boldsymbol{s}_i \text{ is the } i\text{-th column of } \boldsymbol{S}_t \right\}, \tag{3.1}$$

$$\mathcal{E}_{t,2} := \left\{ \left(\frac{3}{4} - \frac{1}{k}\right) \|\nabla f(\boldsymbol{x}_t)\|^2 \leq \nabla^\top f(\boldsymbol{x}_t) \boldsymbol{S}_t \boldsymbol{S}_t^\top \nabla f(\boldsymbol{x}_t) \leq \left(\frac{5}{4} + \frac{1}{k}\right) \|\nabla f(\boldsymbol{x}_t)\|^2 \right\}, \tag{3.2}$$

$$\mathcal{E}_{t,3} := \left\{ \left\| \boldsymbol{S}_t^\top \boldsymbol{H} \boldsymbol{S}_t \right\|_2 \leq \zeta_2 \right\}, \tag{3.3}$$

$$\mathcal{E}_{t,4} := \left\{ \frac{1}{\ell} \left(\sum_{i=1}^{\ell} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_i \rangle \right)^2 \leq \frac{1}{4} \|\nabla f(\boldsymbol{x}_t)\|^2 \right\}, \tag{3.4}$$

其中 ζ_1和 ζ_2 分别定义于公式（3.5）和公式（3.6）中。

接下来，我们将给出上述事件发生概率的上界。

引理 1。 令 $\boldsymbol{S}_t$ 为一个满足定义*1* 的无偏（$1/4, k, \delta$）-随机草图矩阵，其中$k \geq 4$ 且 $\delta \in (0,1)$，则事件 $\mathcal{E}_{t,\ 1}$ 以至少 $1-\delta$ 的概率成立，其中 ζ_1 定义为

$$\zeta_1 := \|\boldsymbol{H}\|^2 \left(\frac{5\ell}{4} + \frac{d\ell}{4k} \right)^2. \tag{3.5}$$

证明。我们有

$$\sum_{i=1}^{\ell} |\boldsymbol{s}_i^\top \boldsymbol{H} \boldsymbol{s}_i|^2 \leq \|\boldsymbol{H}\|^2 \sum_{i=1}^{\ell} \|\boldsymbol{s}_i\|^4 \leq \|\boldsymbol{H}\|^2 \|\boldsymbol{S}_t\|_F^4 \overset{(A.2)}{\leq} \|\boldsymbol{H}\|^2 \left(\frac{5\ell}{4} + \frac{d\ell}{4k} \right)^2,$$

其中，最后一个不等式是由引理 28得出的。 □

引理 2. 令 $\boldsymbol{S} \in \mathbb{R}^{\ell \times d}$ 是满足定义*1*且参数为 $k \geq 4$ 和 $\delta \in (0,1)$的无偏$(1/4, k, \delta)$随机草图矩阵，则事件 $\mathcal{E}_{t,2}$ 以至少 $1-\delta$ 的概率成立。

证明。 根据公式 (2.4)，可得

$$\left\| \nabla^\top f(\boldsymbol{x}_t) \boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t) - \nabla^\top f(\boldsymbol{x}_t) \nabla f(\boldsymbol{x}_t) \right\| \leq \frac{1}{4} \left(\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{\|\nabla f(\boldsymbol{x}_t)\|^2}{k} \right).$$

此外，有

$$\begin{aligned}
\left\| \nabla^\top f(\boldsymbol{x}_t) \boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t) \right\| - \|\nabla f(\boldsymbol{x}_t)\|^2 &\leq \left\| \nabla^\top f(\boldsymbol{x}_t) \boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t) - \nabla^\top f(\boldsymbol{x}_t) \nabla f(\boldsymbol{x}_t) \right\|, \\
\|\nabla f(\boldsymbol{x}_t)\|^2 - \left\| \nabla^\top f(\boldsymbol{x}_t) \boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t) \right\| &\leq \left\| \nabla^\top f(\boldsymbol{x}_t) \boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t) - \nabla^\top f(\boldsymbol{x}_t) \nabla f(\boldsymbol{x}_t) \right\|.
\end{aligned}$$

因此，我们可以得到

$$\left(\frac{3}{4}-\frac{1}{k}\right)\|\nabla f(\boldsymbol{x}_t)\|^2 \leq \left\|\nabla^\top f(\boldsymbol{x}_t)\boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\| \leq \left(\frac{5}{4}+\frac{1}{k}\right)\|\nabla f(\boldsymbol{x}_t)\|^2 .$$

□

引理 3. 令 $\boldsymbol{S}\in\mathbb{R}^{d\times\ell}$ 为一个无偏的$(1/4, k, \delta)$-随机草图矩阵，则事件 $\mathcal{E}_{t,3}$以至少 $1-\delta$ 的概率成立，其中 ζ_2 定义为

$$\zeta_2 := \frac{5\|\boldsymbol{H}\|}{4}+\frac{\operatorname{tr}(\boldsymbol{H})}{4k}. \tag{3.6}$$

证明。 首先，我们有

$$\left\|\boldsymbol{S}^\top\boldsymbol{H}\boldsymbol{S}\right\| = \left\|\boldsymbol{H}^{1/2}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{H}^{1/2}\right\|.$$

根据 $\boldsymbol{S}$是一个无偏的 $(1/4, k, \delta)$-随机草图矩阵这一性质，我们可以得到

$$\left\|\boldsymbol{H}^{1/2}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{H}^{1/2}-\boldsymbol{H}\right\| \leq \frac{1}{4}\left(\|\boldsymbol{H}\|_2+\frac{\operatorname{tr}(\boldsymbol{H})}{k}\right),$$

这意味着

$$\left\|\boldsymbol{H}^{1/2}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{H}^{1/2}\right\| \leq \frac{5}{4}\|\boldsymbol{H}\|_2+\frac{\operatorname{tr}(\boldsymbol{H})}{4k}.$$

□

引理 4. 若 $\boldsymbol{S}\in\mathbb{R}^{d\times\ell}$ 是高斯、拉德马赫、子采样随机哈达玛变换或稀疏嵌入随机草图矩阵（参见第 *2.3* 节），则事件 $\mathcal{E}_{t,4}$ 以至少 $1-\delta'$ 的概率成立，其中$\delta'=\exp(-\mathcal{O}(\ell))$。

证明。 参见引理 30。 □

3.2 估计梯度和标准差的性质

首先，我们将介绍若干与公式（2.8）中定义的估计梯度相关的重要性质。

引理 5。 在假设 *1* 和 *2*的基础上，我们进一步假设公式 $\mathbf{g}(\boldsymbol{x})$ 所定义的梯度具有如下形式：

$$\mathbf{g}(\boldsymbol{x}) = \boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x})+\frac{\alpha}{4}\boldsymbol{S}\boldsymbol{v} \quad \textit{and} \quad |v^{(i)}| \leq 2\left(L_0+L_1\|\nabla f(\boldsymbol{x})\|\right)\|\boldsymbol{s}_i\|_{\boldsymbol{H}}^2, \tag{3.7}$$

其中 $\boldsymbol{S}=[\boldsymbol{s}_1,\dots\boldsymbol{s}_\ell]\in\mathbb{R}^{d\times\ell}$是给定矩阵， $\boldsymbol{v}=[v^{(i)},\dots,v^{(\ell)}]\in\mathbb{R}^\ell$是一个 ℓ维

向量，其坐标为 $v^{(i)}=4\alpha^{-1}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)$，且

$$\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)=f(\boldsymbol{x}+\alpha\boldsymbol{s}_i)-f(\boldsymbol{x})-\alpha\left\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\right\rangle. \tag{3.8}$$

证明。我们有

$$\begin{aligned}\sum_{i=1}^{\ell}\frac{f(\boldsymbol{x}+\alpha\boldsymbol{s}_i)-f(\boldsymbol{x})}{\alpha}\boldsymbol{s}_i &\overset{(3.8)}{=}\sum_{i=1}^{\ell}\frac{\alpha\left\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\right\rangle+\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)}{\alpha}\boldsymbol{s}_i\\ &=\sum_{i=1}^{\ell}\left\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\right\rangle\boldsymbol{s}_i+\alpha^{-1}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)\cdot\boldsymbol{s}_i\\ &=\boldsymbol{S}\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x})+\frac{\alpha}{4}\boldsymbol{S}\boldsymbol{v},\end{aligned}$$

其中 $v^{(i)}=4\alpha^{-1}\cdot\gamma(\boldsymbol{x},\ \alpha\boldsymbol{s}_i)$。

此外，根据公式（2.3），我们有

$$\left|4\alpha^{-1}\cdot\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)\right|\overset{(2.3)}{\leq}\frac{4}{\alpha}\cdot\frac{L_0+L_1\left\|\nabla f(\boldsymbol{x})\right\|}{2}\cdot\alpha^2\left\|\boldsymbol{s}_i\right\|_{\boldsymbol{H}}^2=2\left(L_0+L_1\left\|\nabla f(\boldsymbol{x})\right\|\right)\left\|\boldsymbol{s}_i\right\|_{\boldsymbol{H}}^2.$$

□

引理 6。 设 $\boldsymbol{v}=[v^{(i)},\ldots,v^{(\ell)}]\in\mathbb{R}^{\ell}$ 是一个 ℓ 维向量，其满足 $v^{(i)}=4\alpha^{-1}\cdot\gamma\left(\boldsymbol{x},\alpha\boldsymbol{s}_i\right)$，其中 $\gamma\left(\boldsymbol{x},\alpha\boldsymbol{s}_i\right)$ 遵循公式（3.8）中的定义。在事件 $\mathcal{E}_{t,\ 1}$ 条件下，$\|\boldsymbol{v}\|^2$ 的上界为

$$\|\boldsymbol{v}\|^2\leq 4(L_0+L_1\left\|\nabla f(\boldsymbol{x})\right\|)^2\zeta_1, \tag{3.9}$$

其中 ζ_1 定义见公式（3.5）。

证明。根据 $\boldsymbol{v}$ 的定义，我们可以得到

$$\|\boldsymbol{v}\|^2=\sum_{i=1}^{\ell}\left(v^{(i)}\right)^2\overset{(3.7)}{\leq}4(L_0+L_1\left\|\nabla f(\boldsymbol{x})\right\|)^2\sum_{i=1}^{\ell}\left\|\boldsymbol{s}_i\right\|_{\boldsymbol{H}}^4\overset{(3.1)}{\leq}4(L_0+L_1\left\|\nabla f(\boldsymbol{x})\right\|)^2\zeta_1,$$

由此得证。 □

引理 7. 令 $\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)$ 由公式 (3.8)定义，在事件 $\mathcal{E}_{t\text{下},\ 1}$ 由公式 (3.1)定义，则有

$$\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)^2\leq\frac{\alpha^4(L_0^2+L_1^2\left\|\nabla f(\boldsymbol{x})\right\|^2)}{4}\cdot\zeta_1. \tag{3.10}$$

证明。根据 d 的定义 γ 在公式(3.8)以及公式(2.3)中的假设，我们可以得到

$$|\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)|\leq\alpha^2\cdot\frac{L_0+L_1\left\|\nabla f(\boldsymbol{x})\right\|}{2}|\boldsymbol{s}_i^{\top}\boldsymbol{H}\boldsymbol{s}_i|.$$

因此，我们可以得到

$$\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)^2 \le \alpha^4\cdot\frac{(L_0+L_1\|\nabla f(\boldsymbol{x})\|)^2}{4}\cdot\sum_{i=1}^{\ell}|\boldsymbol{s}_i^\top\boldsymbol{H}\boldsymbol{s}_i|^2 \overset{(3.1)}{\le} \frac{\alpha^4(L_0+L_1\|\nabla f(\boldsymbol{x})\|)^2}{4}\cdot\zeta_1.$$

□

接下来，我们将证明公式（ σ 2.5）中定义的标准差的性质 .

引理 8。设标准差 σ 定义于公式（2.5），则有

$$\sigma^2 \le \frac{3\alpha^2}{2(\ell-1)}\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle^2+\frac{3}{\ell-1}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_j)^2 \tag{3.11}$$

and

$$\sigma^2 \ge \frac{\alpha^2}{2(\ell-1)}\sum_{i=1}^{\ell}\left(\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle^2-\frac{1}{\ell^2}\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle\right)^2\right)-\frac{1}{\ell-1}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_j)^2, \tag{3.12}$$

其中 γ（$\boldsymbol{x}$， $\alpha\boldsymbol{s}_i$）定义于公式（3.8）。

证明。根据公式（ γ3.8）中·， ·）的定义，我们有

$$\begin{aligned}
&f(\boldsymbol{x}+\alpha\boldsymbol{s}_j)-\frac{1}{\ell}\sum_{i=1}^{\ell}f(\boldsymbol{x}+\alpha\boldsymbol{s}_i)\\
=&f(\boldsymbol{x})+\alpha\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_j\rangle+\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_j)-\frac{1}{\ell}\sum_{i=1}^{\ell}\left(f(\boldsymbol{x})+\alpha\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle+\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)\right)\\
=&\alpha\left\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_j-\frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i\right\rangle+\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_j)-\frac{1}{\ell}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i).
\end{aligned}$$

定义 $\tau_j=\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_j)-\frac{1}{\ell}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_i)$，我们可以得到

$$\begin{aligned}
&\sum_{j=1}^{\ell}\left(f(\boldsymbol{x}+\alpha\boldsymbol{s}_j)-\frac{1}{\ell}\sum_{i=1}^{\ell}f(\boldsymbol{x}+\alpha\boldsymbol{s}_i)\right)^2\\
=&\sum_{j=1}^{\ell}\left(\alpha\left\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_j-\frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i\right\rangle+\tau_j\right)^2\\
=&\sum_{j=1}^{\ell}\left(\alpha^2\left\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_j-\frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i\right\rangle^2+\tau_j^2+2\alpha\left\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_j-\frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i\right\rangle\tau_j\right).
\end{aligned}$$

此外，

$$2\alpha\left|\left\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_j - \frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i \right\rangle \tau_j\right| \leq \frac{\alpha^2}{2}\left\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_j - \frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i \right\rangle^2 + 2\tau_j^2.$$

因此，成立如下

$$\sum_{j=1}^{\ell}\left(f(\boldsymbol{x}+\alpha\boldsymbol{s}_j) - \frac{1}{\ell}\sum_{i=1}^{\ell} f(\boldsymbol{x}+\alpha\boldsymbol{s}_i)\right)^2 \leq \sum_{j=1}^{\ell}\left(\frac{3\alpha^2}{2}\left\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_j - \frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i \right\rangle^2 + 3\tau_j^2\right), \tag{3.13}$$

and

$$\sum_{j=1}^{\ell}\left(\frac{\alpha^2}{2}\left\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_j - \frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i \right\rangle^2 - \tau_j^2\right) \leq \sum_{j=1}^{\ell}\left(f(\boldsymbol{x}+\alpha\boldsymbol{s}_j) - \frac{1}{\ell}\sum_{i=1}^{\ell} f(\boldsymbol{x}+\alpha\boldsymbol{s}_i)\right)^2. \tag{3.14}$$

接下来，我们将如下界定 $\left\langle \nabla f(\boldsymbol{x})\ \boldsymbol{s}_j - \frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i \right\rangle^2$ 与 τ_j^2的值

$$\begin{aligned}
&\sum_{j=1}^{\ell}\left\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_j - \frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i \right\rangle^2 \\
=&\sum_{j=1}^{\ell}\left(\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_j\rangle^2 - 2\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_j\rangle \cdot \left(\frac{1}{\ell}\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle\right) + \left(\frac{1}{\ell}\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle\right)^2\right) \\
=&\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle^2 - \frac{2}{\ell}\left(\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle\right)^2 + \frac{1}{\ell}\left(\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle\right)^2 \\
=&\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle^2 - \frac{1}{\ell}\left(\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle\right)^2.
\end{aligned} \tag{3.15}$$

此外，有

$$0 \leq \sum_{j=1}^{\ell}\tau_j^2 = \sum_{j=1}^{\ell}\left(\gamma(\boldsymbol{x}, \alpha\boldsymbol{s}_j) - \frac{1}{\ell}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x}, \alpha\boldsymbol{s}_i)\right)^2 \leq \sum_{i=1}^{\ell}\gamma(\boldsymbol{x}, \alpha\boldsymbol{s}_j)^2. \tag{3.16}$$

结合上述结果，我们可以得到

$$\begin{aligned}
\sigma^2 &\overset{(2.5)}{=} \frac{1}{\ell-1}\sum_{j=1}^{\ell}\left(f(\boldsymbol{x}+\alpha\boldsymbol{s}_j)-\frac{1}{\ell}\sum_{i=1}^{\ell}f(\boldsymbol{x}+\alpha\boldsymbol{s}_i)\right)^2 \\
&\overset{(3.13)}{\leq} \frac{1}{\ell-1}\sum_{j=1}^{\ell}\left(\frac{3\alpha^2}{2}\left\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_j-\frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i\right\rangle^2+3\tau_j^2\right) \\
&\overset{(3.15)}{=} \frac{3\alpha^2}{2(\ell-1)}\sum_{i=1}^{\ell}\left(\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle^2-\frac{1}{\ell^2}\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle\right)^2\right)+\frac{3}{\ell-1}\sum_{i=1}^{\ell}\tau_i^2 \\
&\overset{(3.16)}{\leq} \frac{3\alpha^2}{2(\ell-1)}\sum_{i=1}^{\ell}\left(\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle^2-\frac{1}{\ell^2}\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle\right)^2\right)+\frac{3}{\ell-1}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_j)^2 \\
&\leq \frac{3\alpha^2}{2(\ell-1)}\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle^2+\frac{3}{\ell-1}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_j)^2,
\end{aligned}$$

and

$$\begin{aligned}
\sigma^2 &= \frac{1}{\ell-1}\sum_{j=1}^{\ell}\left(f(\boldsymbol{x}+\alpha\boldsymbol{s}_j)-\frac{1}{\ell}\sum_{i=1}^{\ell}f(\boldsymbol{x}+\alpha\boldsymbol{s}_i)\right)^2 \\
&\overset{(3.14)}{\geq} \frac{1}{\ell-1}\sum_{j=1}^{\ell}\left(\frac{\alpha^2}{2}\left\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_j-\frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{s}_i\right\rangle^2-\tau_j^2\right) \\
&\overset{(3.15)}{=} \frac{\alpha^2}{2(\ell-1)}\sum_{i=1}^{\ell}\left(\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle^2-\frac{1}{\ell^2}\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle\right)^2\right)-\frac{1}{\ell-1}\sum_{i=1}^{\ell}\tau_i^2 \\
&\overset{(3.16)}{\geq} \frac{\alpha^2}{2(\ell-1)}\sum_{i=1}^{\ell}\left(\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle^2-\frac{1}{\ell^2}\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}),\boldsymbol{s}_i\rangle\right)^2\right)-\frac{1}{\ell-1}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_j)^2.
\end{aligned}$$

□

引理 9。 假设算法的平滑参数*1* 满足

$$\alpha^2 \leq \min\left\{\frac{1}{2L_1^2\zeta_1},\ \frac{3\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{8L_0^2\zeta_1}\right\}$$

其中 ζ_1 定义于公式 (3.5)中。在事件 $\mathcal{E}_{t,2}$ 和 $\mathcal{E}_{t,4}$ 分别定义于公式 (3.2) 和公式的条件下， (3.4) 有

$$\sigma_t \geq \frac{\alpha}{8\sqrt{(\ell-1)}}\left\|\nabla f(\boldsymbol{x}_t)\right\|. \tag{3.17}$$

证明。我们有

$$\begin{aligned}\sigma_t^2 \overset{(3.12)}{\geq} & \frac{\alpha^2}{2(\ell-1)}\sum_{i=1}^{\ell}\left(\langle\nabla f(\boldsymbol{x}_t),\boldsymbol{s}_i\rangle^2-\frac{1}{4\ell}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right)-\frac{1}{\ell-1}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_j)^2\\ \geq & \frac{\alpha^2}{2(\ell-1)}\cdot\frac{\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{4}-\frac{1}{\ell-1}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x},\alpha\boldsymbol{s}_j)^2,\end{aligned}$$

其中第一个不等式是因为条件 $\frac{1}{\ell}\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x})\ \boldsymbol{s}_i\rangle\right)^2\leq\frac{1}{4}\left\|\nabla f(\boldsymbol{x})\right\|^2$ 而最后一个不等式是因为引理 2。

由引理 7，我们可以得到

$$\begin{aligned}\sigma_t^2 \overset{(3.10)}{\geq} & \frac{\alpha^2}{8(\ell-1)}\cdot\left\|\nabla f(\boldsymbol{x}_t)\right\|^2-\frac{\alpha^4(L_0^2+L_1^2\left\|\nabla f(\boldsymbol{x}_t)\right\|^2)\zeta_1}{4(\ell-1)}\\ = & \frac{\alpha^2}{8(\ell-1)}\left(1-\alpha^2L_1^2\zeta_1\right)\left\|\nabla f(\boldsymbol{x}_t)\right\|^2-\frac{\alpha^4L_0^2\zeta_1}{4(\ell-1)}\\ \geq & \frac{\alpha^2}{16(\ell-1)}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2-\frac{\alpha^4L_0^2\zeta_1}{4(\ell-1)}\\ \geq & \frac{\alpha^2}{64(\ell-1)}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2,\end{aligned}$$

其中最后两个不等式是因为关于 α^2 的假设。 □

引理 10. 假设算法*1* 的平滑参数满足

$$\alpha^2\leq\min\{\frac{1}{2L_1^2\zeta_1},\ \frac{3\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{8L_0^2\zeta_1}\}.$$

在事件 $\mathcal{E}_{t,\ 2}$ （定义于公式 (3.2)）条件下，则有

$$\sigma_t\leq\frac{13\left\|\nabla f(\boldsymbol{x}_t)\right\|}{2\sqrt{\ell-1}}.\tag{3.18}$$

证明。 由引理 8，我们有

$$\begin{aligned}\sigma_t^2 \overset{(3.11)}{\leq} & \frac{3\alpha^2}{2(\ell-1)}\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}_t),\boldsymbol{s}_i\rangle^2+\frac{3}{\ell-1}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x}_t,\alpha\boldsymbol{s}_j)^2\\ \overset{(3.2)}{\leq} & \frac{9\alpha^2}{4(\ell-1)}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2+\frac{3}{\ell-1}\sum_{i=1}^{\ell}\gamma(\boldsymbol{x}_t,\alpha\boldsymbol{s}_j)^2\\ \leq & \frac{9\alpha^2}{4(\ell-1)}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2+\frac{3\alpha^4(L_0^2+L_1^2\left\|\nabla f(\boldsymbol{x}_t)\right\|^2)\zeta_1}{4(\ell-1)},\end{aligned}$$

其中最后一个不等式是由引理 7 得出的。

结合关于 α^2 的条件，我们可以得到

$$\sigma_t^2 \leq \left(\frac{9}{4}+\frac{3}{4}\right)\frac{\alpha^2\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{\ell-1}+\frac{3\alpha^4 L_0^2\zeta_1}{4(\ell-1)} \leq \left(\frac{9}{4}+\frac{3}{4}+\frac{9}{64}\right)\frac{\alpha^2\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{\ell-1},$$

这意味着

$$\sigma_t \leq \frac{13\alpha\left\|\nabla f(\boldsymbol{x}_t)\right\|}{2\sqrt{\ell-1}}.$$

□

引理 11。 令事件 $\mathcal{E}_{t,1}$, $\mathcal{E}_{t,2}$, $\mathcal{E}_{t,3}$, 以及 $\mathcal{E}_{t,4}$ 均成立，且算法 *1*的平滑参数满足

$$\alpha^2 \leq \min\{\frac{1}{2L_1^2\zeta_1}, \frac{3\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{8L_0^2\zeta_1}\},$$

则有

$$\frac{\left\|\mathbf{g}(\boldsymbol{x}_t)\right\|_{\boldsymbol{H}}}{\sigma_t} \leq \frac{12\sqrt{\zeta_2}\sqrt{\ell-1}}{\alpha}+\frac{4\sqrt{2}\sqrt{\zeta_2}}{\alpha}. \tag{3.19}$$

证明。我们有

$$\begin{aligned}
\frac{\left\|\mathbf{g}(\boldsymbol{x}_t)\right\|_{\boldsymbol{H}}}{\sigma_t} \overset{(3.7)}{\leq} & \frac{\left\|\boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\|_{\boldsymbol{H}}+\frac{\alpha}{4}\left\|\boldsymbol{S}\boldsymbol{v}\right\|_{\boldsymbol{H}}}{\sigma_t} \leq \frac{\sqrt{\zeta_2}\left(\left\|\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\|+\frac{\alpha}{4}\left\|\boldsymbol{v}\right\|\right)}{\sigma_t} \\
\leq & \frac{\sqrt{\zeta_2}\left(\frac{3}{2}\left\|\nabla f(\boldsymbol{x}_t)\right\|+\frac{\alpha}{4}\sqrt{4(L_0+L_1\left\|\nabla f(\boldsymbol{x})\right\|)^2\zeta_1}\right)}{\sigma_t} \\
\overset{(3.17)}{\leq} & \frac{\sqrt{\zeta_2}\left(\frac{3}{2}\left\|\nabla f(\boldsymbol{x}_t)\right\|+\frac{\alpha}{4}\sqrt{4(L_0+L_1\left\|\nabla f(\boldsymbol{x})\right\|)^2\zeta_1}\right)}{\frac{\alpha}{8\sqrt{(\ell-1)}}\left\|\nabla f(\boldsymbol{x}_t)\right\|} \\
= & \frac{12\sqrt{\zeta_2}\sqrt{\ell-1}}{\alpha}+\frac{4\sqrt{\zeta_2}\sqrt{\zeta_1}(L_0+L_1\left\|\nabla f(\boldsymbol{x})\right\|)}{\left\|\nabla f(\boldsymbol{x}_t)\right\|} \\
= & \frac{12\sqrt{\zeta_2}\sqrt{\ell-1}}{\alpha}+\frac{4\sqrt{\zeta_2}\sqrt{\zeta_1}L_1\alpha}{\alpha}+\frac{4\sqrt{\zeta_2}\sqrt{\zeta_1}L_0\alpha}{\alpha\left\|\nabla f(\boldsymbol{x}_t)\right\|} \\
\leq & \frac{12\sqrt{\zeta_2}\sqrt{\ell-1}}{\alpha}+\frac{2\sqrt{2}\sqrt{\zeta_2}}{\alpha}+\frac{2\sqrt{2}\sqrt{\zeta_2}}{\alpha} \\
= & \frac{12\sqrt{\zeta_2}\sqrt{\ell-1}}{\alpha}+\frac{4\sqrt{2}\sqrt{\zeta_2}}{\alpha},
\end{aligned}$$

其中，最后一个不等式是由关于 α^2的条件得出的。 □

3.3 下降分析与主要结果

引理12。 假设函数 $f:\ \mathbb{R}^d \to \mathbb{R}$ 满足假设 *1* 且事件 $\mathcal{E}_{t,\ 3}$ （定义于公式 (3.3)）成立。算法 *1* 以步长 $\eta \leq 3\alpha/(64L_1\zeta_2\sqrt{\ell}-1)$进行一次更新后，有

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) \leq & f(\boldsymbol{x}_t) - \frac{7\eta}{8\sigma_t}\left\|\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{\eta\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8\sigma_t} \\ & + \frac{\eta^2}{\sigma_t^2}\cdot\frac{L_0+L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|}{2}\cdot\left(2\zeta_2\cdot\left\|\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{\alpha^2\zeta_2}{8}\left\|\boldsymbol{v}_t\right\|^2\right), \end{aligned} \tag{3.20}$$

其中 ζ_2 定义于公式 (3.6)。

证明。 首先，根据更新规则，我们有

$$\begin{aligned} \left\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\right\|_H =& \eta\cdot\frac{\left\|\mathbf{g}(\boldsymbol{x}_t)\right\|_{\boldsymbol{H}}}{\sigma_t} \overset{(3.19)}{\leq} \eta\left(\frac{12\sqrt{\zeta_2}\sqrt{\ell-1}}{\alpha}+\frac{4\sqrt{2}\sqrt{\zeta_2}}{\alpha}\right) \leq \frac{9}{16L_1\sqrt{\zeta_2}}+\frac{3\sqrt{2}}{16L_1\sqrt{\ell-1}\sqrt{\zeta_2}} \\ \leq & \frac{1}{L_1\sqrt{\zeta_2}} \leq \frac{1}{L_1\sqrt{\left\|\boldsymbol{H}\right\|}}, \end{aligned}$$

其中，最后一个不等式是由 ζ_2 在公式 (3.6) 中的定义得出的。因此，假设 1 中的条件成立。然后，根据公式 (2.6) 中的更新步，以及假设 L_0, L_1-光滑的 $f(\boldsymbol{x})$，我们可以得到

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) \overset{(2.3)}{\leq} & f(\boldsymbol{x}_t) + \langle\nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1}-\boldsymbol{x}_t\rangle + \frac{L_0+L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|}{2}(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t)^\top\boldsymbol{H}(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t) \\ = & f(\boldsymbol{x}_t) - \frac{\eta}{\sigma_t}\langle\nabla f(\boldsymbol{x}_t), \mathbf{g}(\boldsymbol{x}_t)\rangle + \frac{\eta^2}{\sigma_t^2}\frac{L_0+L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|}{2}\mathbf{g}(\boldsymbol{x}_t)^\top\boldsymbol{H}\mathbf{g}(\boldsymbol{x}_t) \\ \overset{(3.7)}{=} & f(\boldsymbol{x}_t) - \frac{\eta}{\sigma_t}\left\langle\nabla f(\boldsymbol{x}_t), \boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)+\frac{\alpha}{4}\boldsymbol{S}\boldsymbol{v}_t\right\rangle \\ & + \frac{\eta^2}{\sigma_t^2}\frac{L_0+L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|}{2}\left(\boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)+\frac{\alpha}{4}S\boldsymbol{v}_t\right)^\top\boldsymbol{H}\left(\boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)+\frac{\alpha}{4}S\boldsymbol{v}_t\right). \end{aligned}$$

此外，还有

$$\begin{aligned} & -\left\langle\nabla f(\boldsymbol{x}_t), \boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)+\frac{\alpha}{4}\boldsymbol{S}\boldsymbol{v}_t\right\rangle = -\left\|\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\|^2 - \frac{1}{4}\left\langle\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t), \alpha\boldsymbol{v}_t\right\rangle \\ \leq & -\left\|\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{1}{8}\left\|\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8} \\ = & -\frac{7}{8}\left\|\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8}. \end{aligned}$$

我们还有

$$
\begin{aligned}
&\left(\boldsymbol{S}\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)+\frac{\alpha}{4}\boldsymbol{S}\boldsymbol{v}_t\right)^{\top}\boldsymbol{H}\left(\boldsymbol{S}\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)+\frac{\alpha}{4}\boldsymbol{S}\boldsymbol{v}_t\right)\\
=&\nabla^{\top} f(\boldsymbol{x}_t)\boldsymbol{S}\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)+\frac{\alpha}{2}\boldsymbol{v}_t^{\top}\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)+\frac{\alpha^2}{16}\boldsymbol{v}_t^{\top}\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\boldsymbol{v}_t\\
\leq&\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\cdot\left\|\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\|^2+\frac{\alpha}{2}\left\|\boldsymbol{v}_t\right\|\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\left\|\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\|+\frac{\alpha^2}{16}\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\left\|\boldsymbol{v}_t\right\|^2\\
\leq&\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\cdot\left\|\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\|^2+\left(\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\cdot\left\|\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\|^2+\frac{\alpha^2}{16}\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\left\|\boldsymbol{v}_t\right\|^2\right)\\
&+\frac{\alpha^2}{16}\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\left\|\boldsymbol{v}_t\right\|^2\\
=&2\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\cdot\left\|\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\|^2+\frac{\alpha^2}{8}\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\left\|\boldsymbol{v}_t\right\|^2.
\end{aligned}
$$

结合上述结果，我们可以得到

$$
\begin{aligned}
f(\boldsymbol{x}_{t+1})\leq& f(\boldsymbol{x}_t)-\frac{7\eta}{8\sigma_t}\left\|\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\|^2+\frac{\eta\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8\sigma_t}\\
&+\frac{\eta^2}{\sigma_t^2}\cdot\frac{L_0+L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|}{2}\cdot\left(2\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\cdot\left\|\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\|^2+\frac{\alpha^2}{8}\left\|\boldsymbol{S}^{\top}\boldsymbol{H}\boldsymbol{S}\right\|_2\left\|\boldsymbol{v}_t\right\|^2\right).
\end{aligned}
$$

结合事件 $\mathcal{E}_{t,\ 3}$，我们可以得到最终结果。 □

引理 13. 设函数 $f\colon\mathbb{R}^d\to\mathbb{R}$ 满足假设 *1*，且事件 $\mathcal{E}_{t,1}$, $\mathcal{E}_{t,2}$, $\mathcal{E}_{t,3}$和 $\mathcal{E}_{t,4}$均成立。我们为算法 $\eta\leq 3\alpha/(64L_1\zeta_2\sqrt{\ell}-1)$ 设定步长 $\alpha^2\leq\min\{\frac{1}{2L_1^2\zeta_1},\frac{3\|\nabla f(\boldsymbol{x}_t)\|^2}{8L_0^2\zeta_1}\}$和平滑参数 *1*，则有

$$
\begin{aligned}
f(\boldsymbol{x}_{t+1})\leq& f(\boldsymbol{x}_t)-\frac{\eta}{2\sigma_t}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2+\frac{3\eta^2L_0\zeta_2\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{2\sigma_t^2}+\frac{\eta\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8\sigma_t}\\
&+\frac{\eta^2\alpha^2\zeta_2(L_0+L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|)\left\|\boldsymbol{v}_t\right\|^2}{16\sigma_t^2}.
\end{aligned}
\tag{3.21}
$$

证明。 首先，我们有

$$
\begin{aligned}
f(\boldsymbol{x}_{t+1}) \overset{(3.20)}{\leq}& f(\boldsymbol{x}_t) - \frac{7\eta}{8\sigma_t}\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{\eta\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8\sigma_t} \\
&+ \frac{\eta^2}{\sigma_t^2}\cdot\frac{L_0 + L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|}{2}\cdot\left(2\zeta_2\cdot\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{\alpha^2\zeta_2}{8}\left\|\boldsymbol{v}_t\right\|^2\right) \\
\overset{(3.17)}{\leq}& f(\boldsymbol{x}_t) - \frac{7\eta}{8\sigma_t}\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{\eta^2\zeta_2 L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2}{\sigma_t\frac{\alpha\left\|\nabla f(\boldsymbol{x}_t)\right\|}{8\sqrt{\ell-1}}} \\
&+ \frac{\eta^2 L_0\zeta_2\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2}{\sigma_t^2} + \frac{\eta\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8\sigma_t} + \frac{\eta^2\alpha^2\zeta_2(L_0 + L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|)\left\|\boldsymbol{v}_t\right\|^2}{16\sigma_t^2} \\
\leq& f(\boldsymbol{x}_t) - \frac{7\eta}{8\sigma_t}\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{\eta^2\zeta_2 L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2}{\sigma_t\frac{\alpha\left\|\nabla f(\boldsymbol{x}_t)\right\|}{8\sqrt{\ell-1}}} \\
&+ \frac{\eta^2 L_0\zeta_2\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2}{\sigma_t^2} + \frac{\eta\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8\sigma_t} + \frac{\eta^2\alpha^2\zeta_2(L_0 + L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|)\left\|\boldsymbol{v}_t\right\|^2}{16\sigma_t^2} \\
=& f(\boldsymbol{x}_t) - \frac{\eta}{\sigma_t}\left(\frac{7}{8} - \frac{8\eta L_1\zeta_2\sqrt{\ell-1}}{\alpha}\right)\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2 \\
&+ \frac{\eta^2 L_0\zeta_2\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2}{\sigma_t^2} + \frac{\eta\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8\sigma_t} + \frac{\eta^2\alpha^2\zeta_2(L_0 + L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|)\left\|\boldsymbol{v}_t\right\|^2}{16\sigma_t^2} \\
\leq& f(\boldsymbol{x}_t) - \frac{\eta}{\sigma_t}\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{\eta^2 L_0\zeta_2\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2}{\sigma_t^2} + \frac{\eta\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8\sigma_t} \\
&+ \frac{\eta^2\alpha^2\zeta_2(L_0 + L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|)\left\|\boldsymbol{v}_t\right\|^2}{16\sigma_t^2},
\end{aligned}
$$

其中最后一个不等式是因为 $\eta \leq \frac{3\alpha}{64L_1\zeta_2\sqrt{\ell-1}}$。

结合公式（3.2），我们可以得到

$$
\begin{aligned}
f(\boldsymbol{x}_{t+1}) \leq& f(\boldsymbol{x}_t) - \frac{\eta}{2\sigma_t}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{3\eta^2 L_0\zeta_2\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{2\sigma_t^2} + \frac{\eta\alpha^2\left\|\boldsymbol{v}_t\right\|^2}{8\sigma_t} \\
&+ \frac{\eta^2\alpha^2\zeta_2(L_0 + L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|)\left\|\boldsymbol{v}_t\right\|^2}{16\sigma_t^2}.
\end{aligned}
$$

□

定理 *1*。假设函数 $f\colon \mathbb{R}^d \to \mathbb{R}$ 满足假设 *1* 和 *2*，且 $\boldsymbol{S}_t \in \mathbb{R}^{d\times\ell}$ 是一个 $(1/4, k, \delta)$-无偏草图矩阵。令 $\mathcal{E}_t = \cap_{i=1}^4 \mathcal{E}_{t,i}$，$i$ 为所有事件 $\mathcal{E}_{t,1}$、 $\mathcal{E}_{t,2}$、 $\mathcal{E}_{t,3}$ 和 $\mathcal{E}_{t,4}$（定义于公式 *(3.1)*-公式 *(3.4)*）均成立的事件。设定平滑参数 α 和步长 η 满足

$$
\alpha^2 \leq \min\left\{\frac{1}{2L_1^2\zeta_1 T}, \frac{3\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{8L_0^2\zeta_1}\right\} \qquad \textit{and} \qquad \eta = \frac{\alpha}{\sqrt{(\ell-1)\zeta_2 T}},
$$

且 $T>0$ 足够大，使得 $\eta \leq 3\alpha/(64L_1\zeta_2\sqrt{\ell}-1)$ 对于算法 1，则序列 $\{\boldsymbol{x}_t\}$生成的满足

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\| \leq & \frac{26\sqrt{\zeta_2}\Big(f(\boldsymbol{x}_0)-f(\boldsymbol{x}_T)\Big)}{\sqrt{T}}+\frac{96\cdot 26L_0\sqrt{\zeta_2}}{\sqrt{T}}+\frac{104+52\sqrt{3}}{L_1T} \\ & +\frac{26\cdot 32\cdot\sqrt{\zeta_2}}{T^{3/2}}+\frac{24\cdot 26\cdot\zeta_2}{T^{3/2}}.\end{aligned} \tag{3.22}$$

此外，成立如下

$$\mathbb{P}\left(\cap_{t=0}^{T-1}\mathcal{E}_t\right)\geq 1-4T\delta. \tag{3.23}$$

证明。 首先，我们有

$$\begin{aligned}f(\boldsymbol{x}_{t+1}) \overset{(3.21)(3.9)}{\leq} & f(\boldsymbol{x}_t)-\frac{\eta}{2\sigma_t}\|\nabla f(\boldsymbol{x}_t)\|^2+\frac{3\eta^2L_0\zeta_2\|\nabla f(\boldsymbol{x}_t)\|^2}{2\sigma_t^2} \\ & +\frac{\eta\alpha^2(L_0+L_1\|\nabla f(\boldsymbol{x}_t)\|)\zeta_1}{\sigma_t}+\frac{\eta^2\alpha^2\zeta_2(L_0+L_1\|\nabla f(\boldsymbol{x}_t)\|)^2\zeta_1}{2\sigma_t^2} \\ \leq & f(\boldsymbol{x}_t)-\frac{\eta}{2\sigma_t}\|\nabla f(\boldsymbol{x}_t)\|^2+\frac{3\eta^2L_0\zeta_2\|\nabla f(\boldsymbol{x}_t)\|^2}{2\sigma_t^2}+\frac{\eta\alpha^2L_1\|\nabla f(\boldsymbol{x}_t)\|\zeta_1}{\sigma_t} \\ & +\frac{\eta^2\alpha^2\zeta_1\zeta_2L_1^2\|\nabla f(\boldsymbol{x}_t)\|^2}{\sigma_t^2}+\frac{\eta L_0\zeta_1\alpha^2}{\sigma_t}+\frac{\eta^2\alpha^2\zeta_1\zeta_2L_0^2}{\sigma_t^2} \\ \overset{(3.17)}{\leq} & f(\boldsymbol{x}_t)-\frac{\eta}{2\sigma_t}\|\nabla f(\boldsymbol{x}_t)\|^2+\frac{96(\ell-1)\eta^2L_0\zeta_2}{\alpha^2}+8\eta\sqrt{\ell-1}L_1\zeta_1\alpha \\ & +64\eta^2(\ell-1)\zeta_1\zeta_2L_1^2+\frac{\eta L_0\zeta_1\alpha^2}{\sigma_t}+\frac{\eta^2\alpha^2\zeta_1\zeta_2L_0^2}{\sigma_t^2} \\ \overset{(3.18)}{\leq} & f(\boldsymbol{x}_t)-\frac{\eta\sqrt{\ell-1}}{26\alpha}\|\nabla f(\boldsymbol{x}_t)\|+\frac{96(\ell-1)\eta^2L_0\zeta_2}{\alpha^2}+8\eta\sqrt{\ell-1}L_1\zeta_1\alpha \\ & +64\eta^2(\ell-1)\zeta_1\zeta_2L_1^2+\frac{\eta L_0\zeta_1\alpha^2}{\sigma_t}+\frac{\eta^2\alpha^2\zeta_1\zeta_2L_0^2}{\sigma_t^2}.\end{aligned}$$

将上述方程改写为

$$\begin{aligned}\frac{\eta\sqrt{\ell-1}}{26\alpha}\|\nabla f(\boldsymbol{x}_t)\| \leq & f(\boldsymbol{x}_t)-f(\boldsymbol{x}_{t+1})+\frac{96(\ell-1)\eta^2L_0\zeta_2}{\alpha^2}+8\eta\sqrt{\ell-1}L_1\zeta_1\alpha \\ & +64\eta^2(\ell-1)\zeta_1\zeta_2L_1^2+\frac{\eta L_0\zeta_1\alpha^2}{\sigma_t}+\frac{\eta^2\alpha^2\zeta_1\zeta_2L_0^2}{\sigma_t^2}.\end{aligned}$$

对上述方程进行裂项求和，可得

$$\begin{aligned}\frac{\eta\sqrt{\ell-1}}{26\alpha}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\| \leq & f(\boldsymbol{x}_0)-f(\boldsymbol{x}_T)+\frac{96(\ell-1)\eta^2L_0\zeta_2}{\alpha^2}\cdot T+8\eta\sqrt{\ell-1}L_1\zeta_1\alpha\cdot T\\ &+64\eta^2(\ell-1)\zeta_1\zeta_2L_1^2\cdot T+\sum_{t=0}^{T-1}\left(\frac{\eta L_0\zeta_1\alpha^2}{\sigma_t}+\frac{\eta^2\alpha^2\zeta_1\zeta_2L_0^2}{\sigma_t^2}\right).\end{aligned}$$

因此，

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\| \leq & \frac{26\alpha\Big(f(\boldsymbol{x}_0)-f(\boldsymbol{x}_T)\Big)}{\eta\sqrt{\ell-1}}\cdot\frac{1}{T}+\frac{96\cdot26\sqrt{\ell-1}\eta L_0\zeta_2}{\alpha}+8\cdot26\cdot L_1\zeta_1\alpha^2\\ &+26\cdot64\cdot\eta\sqrt{\ell-1}\zeta_1\zeta_2L_1^2\alpha+\frac{26\alpha}{\sqrt{\ell-1}T}\sum_{t=0}^{T-1}\left(\frac{L_0\zeta_1\alpha^2}{\sigma_t}+\frac{\eta\alpha^2\zeta_1\zeta_2L_0^2}{\sigma_t^2}\right)\\ = & \frac{26\Big(f(\boldsymbol{x}_0)-f(\boldsymbol{x}_T)\Big)\sqrt{\zeta_2}}{\sqrt{T}}+\frac{96\cdot26L_0\sqrt{\zeta_2}}{\sqrt{T}}+8\cdot26\cdot L_1\zeta_1\alpha^2\\ &+\frac{26\cdot64\cdot\zeta_1\sqrt{\zeta_2}L_1^2\alpha^2}{\sqrt{T}}+\frac{26\alpha}{\sqrt{\ell-1}T}\sum_{t=0}^{T-1}\left(\frac{L_0\zeta_1\alpha^2}{\sigma_t}+\frac{\eta\alpha^2\zeta_1\zeta_2L_0^2}{\sigma_t^2}\right),\end{aligned}$$

其中，最后一个等式是由设定 $\eta=\frac{\alpha}{\sqrt{\ell-1}\sqrt{T}}$ 得出的。此外，

$$\begin{aligned}&\frac{\alpha}{\sqrt{\ell-1}}\cdot\left(\frac{L_0\zeta_1\alpha^2}{\sigma_t}+\frac{\eta\alpha^2\zeta_1\zeta_2L_0^2}{\sigma_t^2}\right)\\ \overset{(3.17)}{\leq} & \frac{\alpha}{\sqrt{\ell-1}}\cdot\left(\frac{8\sqrt{\ell-1}L_0\zeta_1\alpha}{\|\nabla f(\boldsymbol{x}_t)\|}+\frac{8^2(\ell-1)\eta\zeta_1\zeta_2L_0^2}{\|\nabla f(\boldsymbol{x}_t)\|^2}\right)\\ \leq & \frac{8L_0\zeta_1\alpha^2}{\|\nabla f(\boldsymbol{x}_t)\|}+\frac{8^2\alpha^2\zeta_1\zeta_2L_0^2}{\|\nabla f(\boldsymbol{x}_t)\|^2\sqrt{T}}\\ \leq & \frac{8L_0\zeta_1\sqrt{\frac{1}{2L_1^2\zeta_1}}\sqrt{\frac{3\|\nabla f(\boldsymbol{x}_t)\|^2}{8L_0^2\zeta_1}}}{\|\nabla f(\boldsymbol{x}_t)\|}+\frac{8^2\frac{3\|\nabla f(\boldsymbol{x}_t)\|^2}{8L_0^2\zeta_1}\zeta_1\zeta_2L_0^2}{\|\nabla f(\boldsymbol{x}_t)\|^2\sqrt{T}}\\ = & \frac{2\sqrt{3}}{L_1}+\frac{24\zeta_2}{\sqrt{T}},\end{aligned}$$

其中，最后一个不等式是由假设 $\alpha^2\leq\min\{\frac{1}{2L_1^2\zeta_1},\frac{3\|\nabla f(\boldsymbol{x}_t)\|^2}{8L_0^2\zeta_1}\}$ 得出的。

因此，

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\| \leq & \frac{26\sqrt{\zeta_2}v\Big(f(\boldsymbol{x}_0)-f(\boldsymbol{x}_T)\Big)}{\sqrt{T}}+\frac{96\cdot 26L_0\sqrt{\zeta_2}}{\sqrt{T}}+8\cdot 26\cdot L_1\zeta_1\alpha^2 \\ & +\frac{26\cdot 64\cdot\zeta_1\sqrt{\zeta_2}L_1^2\alpha^2}{\sqrt{T}}+\frac{52\sqrt{3}}{L_1T}+\frac{24\cdot 26\cdot\zeta_2}{T^{3/2}} \\ \leq & \frac{26\sqrt{\zeta_2}\Big(f(\boldsymbol{x}_0)-f(\boldsymbol{x}_T)\Big)}{\sqrt{T}}+\frac{96\cdot 26L_0\sqrt{\zeta_2}}{\sqrt{T}}+\frac{104}{L_1T} \\ & +\frac{26\cdot 32\cdot\sqrt{\zeta_2}}{T^{3/2}}+\frac{52\sqrt{3}}{L_1T}+\frac{24\cdot 26\cdot\zeta_2}{T^{3/2}} \\ = & \frac{26\sqrt{\zeta_2}\Big(f(\boldsymbol{x}_0)-f(\boldsymbol{x}_T)\Big)}{\sqrt{T}}+\frac{96\cdot 26L_0\sqrt{\zeta_2}}{\sqrt{T}}+\frac{104+52\sqrt{3}}{L_1T} \\ & +\frac{26\cdot 32\cdot\sqrt{\zeta_2}}{T^{3/2}}+\frac{24\cdot 26\cdot\zeta_2}{T^{3/2}},\end{aligned}$$

其中，第二个不等式是因为 $\alpha^2 \leq \frac{1}{2L_1^2\zeta_1 T}$。

接下来，我们将证明公式 (3.23) 中的结果。根据并集界，我们有

$$\mathbb{P}\left(\mathcal{E}_t\right) \geq 1-\sum_{i=1}^{4}\mathbb{P}\left(\overline{\mathcal{E}_{t,i}}\right) \geq 1-(\delta+\delta+\delta+\exp(-\mathcal{O}\left(\ell\right))) \geq 1-4\delta,$$

其中 $\overline{\mathcal{E}_{t,\ i}}$ 表示 $\mathcal{E}_t$ 的补集，i，第二个不等式是由引理 1-引理4 得出的，最后一个不等式是因为 $\ell=\mathcal{O}\left(k\log\frac{1}{\delta}\right)$。

再次使用并集界，可得

$$\mathbb{P}\left(\cap_{t=0}^{T-1}\mathcal{E}_t\right) \geq 1-\sum_{t=0}^{T-1}\mathbb{P}\left(\overline{\mathcal{E}_t}\right) \geq 1-4T\delta,$$

由此得证。 □

推论 2. 假设函数 $f:\mathbb{R}^d\to\mathbb{R}$ 满足假设 *1* 和 *2*。令 $\boldsymbol{S}_t\in\mathbb{R}^{d\times\ell}$为满足 $\ell=16k\log(4T/\delta)$ 且 $\delta\in(0,1)$ 的高斯或拉德马赫草图矩阵。设平滑参数 α 和步长 η 满足

$$\alpha^2 \leq \min\left\{\frac{1}{2L_1^2\zeta_1 T},\ \frac{3\varepsilon^2}{8L_0^2\zeta_1}\right\} \qquad and \qquad \eta=\frac{\alpha}{\sqrt{(\ell-1)\zeta_2 T}},$$

其中 $T>0$ 足够大，使得 $\eta\leq 3\alpha/(64L_1\zeta_2\sqrt{\ell-1})$ 并且 $\varepsilon>0$ 是算法 *1*的目标精度，那么如果总迭代次数 T 满足

$$T=\mathcal{O}\left(\frac{\|\boldsymbol{H}\|}{\varepsilon^2}+\frac{\mathrm{tr}(\boldsymbol{H})}{k\cdot\varepsilon^2}\right) \tag{3.24}$$

以至少 $1-\delta$的概率，序列 $\{\boldsymbol{x}_t\}$ 其中 $t=0$，…，$T-1$ 由算法 1生成，满足

$$\min_{t=0,\ldots,T-1}\|\nabla f(\boldsymbol{x}_t)\|\leq\varepsilon.$$

此外，总查询复杂度为

$$Q=\mathcal{O}\left(\left(\frac{k\,\|\boldsymbol{H}\|}{\varepsilon^2}+\frac{\mathrm{tr}(\boldsymbol{H})}{\varepsilon^2}\right)\log\frac{1}{\varepsilon\delta}\right). \tag{3.25}$$

证明。根据第2.3节中高斯和拉德马赫草图矩阵的性质，我们可以得到 $\boldsymbol{S}_t\in\mathbb{R}^{d\times\ell}$ 取 $\ell=16k\log\frac{T}{\delta}$ 将使公式（3.23）以 $1-\delta$ 成立。–

将公式（3.22）右侧设为 ε，可得

$$T=\mathcal{O}\left(\frac{\zeta_2}{\varepsilon^2}\right)\overset{(3.6)}{=}\mathcal{O}\left(\frac{\|\boldsymbol{H}\|}{\varepsilon^2}+\frac{\mathrm{tr}(\boldsymbol{H})}{k\cdot\varepsilon^2}\right).$$

总查询复杂度为

$$Q=T\cdot\ell=\mathcal{O}\left(\frac{\zeta_2}{\varepsilon^2}\cdot\ell\right)=\mathcal{O}\left(\left(\frac{k\,\|\boldsymbol{H}\|}{\varepsilon^2}+\frac{\mathrm{tr}(\boldsymbol{H})}{\varepsilon^2}\right)\log\frac{1}{\varepsilon\delta}\right),$$

其中，最后一个等式是因为 $\ell=\mathcal{O}\left(k\log\frac{T}{\delta}\right)$。 – □

注 1。定理 1 和推论 2 表明，算法 1 可以达到 $\mathcal{O}\left(\|\boldsymbol{H}\|/\varepsilon^2+\mathrm{tr}(\boldsymbol{H})/(k\varepsilon^2)\right)$迭代复杂度。如果增大 k，即增大批量大小 ℓ，算法 1 可以用更少的步数找到一个 ε-驻点。然而，公式 (3.25) 表明，增大批量大小 ℓ完全不会降低查询复杂度。

注释2。 公式 (3.25) 中的查询复杂度实现了弱维度依赖性。为清楚起见，我们可以将公式 (3.25) 重新表述为

$$Q=\mathcal{O}\left(\left(\frac{k\,\|\boldsymbol{H}\|}{\varepsilon^2}+\frac{\mathrm{tr}(\boldsymbol{H})}{\|\boldsymbol{H}\|}\cdot\frac{\|\boldsymbol{H}\|}{\varepsilon^2}\right)\log\frac{1}{\varepsilon\delta}\right).$$

在许多实际应用中，我们可以选择合适的矩阵 $\boldsymbol{H}$ 使得 $\mathrm{tr}(\boldsymbol{H})/\|\boldsymbol{H}\|\ll d$。因此，算法 1 即使在假设 (L_0,L_1) 中扩展的光滑条件下，也能内在地实现弱维度依赖性1。我们的工作将采用草图矩阵的零阶算法对经典光滑函数所具有的弱维度依赖性，扩展到了(L_0,L_1)-光滑函数。

注 3. 请回想，即使取 $\hat{g}(\boldsymbol{x}_t)=\nabla f(\boldsymbol{x}_t)$，并在 $\boldsymbol{H}=\boldsymbol{I}$ 下满足假设 1，通过固定步长更新 $\boldsymbol{x}_{t+1}=\boldsymbol{x}_t-\eta\,\hat{g}(\boldsymbol{x}_t)$，要找到一个 ε-驻点且满足 $\eta_t>0$，所需的迭代次数仍取决于 $M=\sup\{\|\nabla f(\boldsymbol{x})\|\mid\boldsymbol{x},\ f(\boldsymbol{x})\leq f(\boldsymbol{x}_0)\}$（Zhang 等人，2020b，定理 4）。相比之下，我们的推论 2 表明，算法 1 所需的迭代次数

仅取决于 $\|\boldsymbol{H}\|$ 和 $\mathrm{tr}(\boldsymbol{H})$，这意味着对于较差的初始化，自适应零阶方法的收敛速度可以任意快于采用固定步长的算法。

4 随机问题的收敛性分析

本节给出算法2 （2.2）在确定性问题中的收敛性分析。

假设4. 我们假设 $f(\boldsymbol{x};\xi)$二次可微，且 $\left\|\nabla^2 f(\boldsymbol{x};\xi)\right\| \leq\hat{} L$，若 $\boldsymbol{x}\in\mathcal{K}$，其中集合 $\mathcal{K}$ 定义为

$$\mathcal{K}:=\left\{\boldsymbol{x}\mid\|\boldsymbol{x}-\boldsymbol{x}_0\|\leq\frac{3\sqrt{T}}{350L_1\sqrt{\zeta_2}}\right\}. \tag{4.1}$$

并且向量 $\boldsymbol{x}_0$ 是算法 2 的初始点， T 是总迭代次数，且 ζ_2 在公式 (3.6)中定义。

上述假设仅用于界定平滑参数 α 的取值。

4.1 事件

假设目标函数 $f(\boldsymbol{x})$ 满足假设 1 且 $\boldsymbol{S}_t$ 是一个无关（$1/4$， k， δ）-随机草图矩阵，正如定义 1 中所定义，其中 $k\geq 4$ 且 $\delta\in(0,\ \mathbf{1})$。然后，我们将定义以下两个事件：

$$\hat{\mathcal{E}}_{t,1}:=\left\{\sum_{i=1}^{\ell}\|\boldsymbol{s}_i\|^4\leq\hat{\zeta}_1\mid\boldsymbol{s}_i\text{ is the }i\text{-th column of }\boldsymbol{S}_t\right\}, \tag{4.2}$$

$$\hat{\mathcal{E}}_{t,2}:=\left\{\left(\frac{3}{4}-\frac{1}{k}\right)\|\nabla f(\boldsymbol{x}_t;\ \xi_t)\|^2\leq\left\|\boldsymbol{S}_t^\top\nabla f(\boldsymbol{x}_t;\ \xi_t)\right\|^2\leq\left(\frac{5}{4}+\frac{1}{k}\right)\|\nabla f(\boldsymbol{x}_t;\ \xi_t)\|^2\right\}, \tag{4.3}$$

$$\hat{\mathcal{E}}_{t,3}:=\left\{\frac{1}{\ell}\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}_t;\ \xi_t),\boldsymbol{s}_i\rangle\right)^2\leq\frac{1}{4}\|\nabla f(\boldsymbol{x}_t;\ \xi_t)\|^2\right\}, \tag{4.4}$$

$$\hat{\mathcal{E}}_{t,4}:=\left\{\left\|\boldsymbol{S}_t^\top\left(\nabla f(\boldsymbol{x}_t;\ \xi_t)-\nabla f(\boldsymbol{x}_t)\right)\right\|^2\leq\left(\frac{5}{4}+\frac{1}{k}\right)\|\nabla f(\boldsymbol{x}_t;\ \xi_t)-\nabla f(\boldsymbol{x}_t)\|^2\right\}, \tag{4.5}$$

ˆ其中 ζ_1 在公式（4.6）中定义。

接下来，我们将界定上述事件发生的概率。

引理 14. 设 $\boldsymbol{S}_t$ 为一个无关$(1/4,k,\delta)$-随机草图矩阵，其定义正如定ˆ
ˆ 义 1 中所述，且 $k\geq 4$ 与 $\delta\in(0,1)$，则事件 $\mathcal{E}_{t,1}$ 以至少 $1-\delta$ 的概率成立，其中 ζ_1 定义为

$$\hat{\zeta}_1:=\left(\frac{5\ell}{4}+\frac{d\ell}{4k}\right)^2. \tag{4.6}$$

证明。我们有

$$\sum_{i=1}^{\ell}\|\boldsymbol{s}_i\|^4 \leq \|\boldsymbol{S}_t\|_F^4 \overset{(A.2)}{\leq} \left(\frac{5\ell}{4}+\frac{d\ell}{4k}\right)^2,$$

其中，最后一个不等式是由引理 28 得出的。 □

引理 15。 令 $\boldsymbol{S}\in\mathbb{R}^{\ell\times d}$ 为一个无偏的$(1/4,k,\delta)$-随机草图矩阵，定义与定义^*1* 中的相同，其中 $k\geq 4$ 和 $\delta\in(0,1)$，则事件 $\mathcal{E}_{t,2}$ 以至少 $1-\delta$的概率成立。

证明。 证明几乎与引理 2的证明相同，只需将 $\nabla f(\boldsymbol{x})$ 替换为 $\nabla f(\boldsymbol{x};\ \xi)$。 □

引理 16。 若 $\boldsymbol{S}\in\mathbb{R}^{d\times\ell}$ 是高斯、拉德马赫、子采样随机哈达玛变换或^稀疏嵌入（参见第 *2.3*节），则在公式 $\mathcal{E}_{t,3}$ （(4.4)中定义的事件以至少 $1-\delta'$ 的概率成立，其中 $\delta'=\exp(-\mathcal{O}(\ell))$。

证明。 参见引理 30 ，其中将 $\nabla f(\boldsymbol{x})$ 替换为 $\nabla f(\boldsymbol{x};\ \xi)$。 □

引理 17. 设 $\boldsymbol{S}\in\mathbb{R}^{\ell\times d}$ 是一个无关的 $(1/4,k,\delta)$-随机草图矩阵，正如所定义^于定义 *1* 中，其中 $k\geq 4$ 且 $\delta\in(0,1)$，则事件 $\mathcal{E}_{t,4}$ 以至少 $1-\delta$ 的概率成立。

证明。 该证明与引理 2 的证明几乎相同，只是将 $\nabla f(\boldsymbol{x})$ 替换为 $\nabla f(\boldsymbol{x};\xi)\ -\nabla f(\boldsymbol{x})$。

□

4.2 估计梯度与标准差的性质

引理 18. 给定 $\boldsymbol{S}\in\mathbb{R}^{d\times\ell}$ 为一个预定义矩阵，令每个函数 $f(\boldsymbol{x};\xi)$ 满足假设^ *4* ，且所有 $\boldsymbol{x}$ 与 $\boldsymbol{x}+\alpha\boldsymbol{s}_i$ 均属于 $\mathcal{K}$。在事件 $\mathcal{E}_{t,1}$ 条件下，该事件定义于公式 (4.2)，近似随机梯度 ^$\mathbf{g}(\boldsymbol{x})$ 定义于公式 (2.9) 满足

$$\hat{\mathbf{g}}(\boldsymbol{x})=\boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x};\xi)+\frac{\alpha}{4}\boldsymbol{S}\hat{\boldsymbol{v}}\ \ and\ \ \|\hat{\boldsymbol{v}}\|^2\leq\hat{L}^2\hat{\zeta}_1, \tag{4.7}$$

其中 ^$\boldsymbol{v}$ 是一个 ℓ-维向量，其中 ^$v^{(i)}=\boldsymbol{s}_i^\top\nabla^2 f(\boldsymbol{x}_{\boldsymbol{s}_i};\xi)\boldsymbol{s}_i$ 为其第 i 个分量， $\boldsymbol{s}_i$ 是第 i^列，属于 $\boldsymbol{S}$，$\boldsymbol{x}_{\boldsymbol{s}_i}$ 是位于 $\boldsymbol{x}$ 与 $\boldsymbol{x}+\boldsymbol{s}_i$ 之间连线上的一点，且 ζ_1 定义于公式 (4.6)。

证明。 首先，通过泰勒展开，我们可以得到

$$f(\boldsymbol{x}+\alpha\boldsymbol{s}_i;\ \xi)=f(\boldsymbol{x};\ \xi)+\alpha\left\langle\nabla f(\boldsymbol{x};\ \xi),\boldsymbol{s}_i\right\rangle+\frac{\alpha^2}{2}\boldsymbol{s}_i^\top\nabla^2 f(\boldsymbol{x}_{\boldsymbol{s}_i};\ \xi)\boldsymbol{s}_i.$$

因此，我们有

$$\begin{aligned}
&\sum_{i=1}^{\ell} \frac{f(\boldsymbol{x}+\alpha \boldsymbol{s}_i;\ \xi)-f(\boldsymbol{x};\ \xi)}{\alpha} \boldsymbol{s}_i \\
=&\sum_{i=1}^{\ell} \frac{\alpha\left\langle \nabla f(\boldsymbol{x};\ \xi), \boldsymbol{s}_i\right\rangle+\frac{\alpha^2}{2} \boldsymbol{s}_i^{\top} \nabla^2 f(\boldsymbol{x}_{\boldsymbol{s}_i};\ \xi) \boldsymbol{s}_i}{\alpha} \boldsymbol{s}_i \\
=&\sum_{i=1}^{\ell}\left\langle \nabla f(\boldsymbol{x};\ \xi), \boldsymbol{s}_i\right\rangle \boldsymbol{s}_i+\frac{\alpha}{2} \sum_{i=1}^{\ell}\left(\boldsymbol{s}_i^{\top} \nabla^2 f(\boldsymbol{x}_{\boldsymbol{s}_i};\ \xi) \boldsymbol{s}_i\right) \boldsymbol{s}_i \\
=&\boldsymbol{S} \boldsymbol{S}^{\top} \nabla f(\boldsymbol{x};\ \xi)+\frac{\alpha}{4} \boldsymbol{S} \hat{\boldsymbol{v}}
\end{aligned}$$

此外，根据 $\hat{\boldsymbol{v}}$ 的定义，我们可以得到

$$|\hat{\boldsymbol{v}}^{(i)}|=|\boldsymbol{s}_i^{\top} \nabla^2 f(\boldsymbol{x}_{\boldsymbol{s}_i};\ \xi) \boldsymbol{s}_i| \leq \hat{L}\left\|\boldsymbol{s}_i\right\|^2,$$

其中，最后一个不等式是由假设 4得出的。

因此，

$$\|\hat{\boldsymbol{v}}\|^2=\sum_{i=1}^{\ell}|\hat{\boldsymbol{v}}^{(i)}|^2 \leq \hat{L}^2 \sum_{i=1}^{\ell}\left\|\boldsymbol{s}_i\right\|^4 \leq \hat{L}^2 \hat{\zeta}_1,$$

其中，最后一个不等式是因为事件$\mathcal{E}_{t,\ 1}$ 成立这一条件。 □

引理19。 令 σ_t 由公式 (2.7)定义，则有

$$\sigma_t^2 \leq \frac{3 \alpha^2}{2(\ell-1)} \sum_{i=1}^{\ell}\left\langle \nabla f(\boldsymbol{x}_t;\ \xi_t), \boldsymbol{s}_i\right\rangle^2+\frac{3}{\ell-1} \sum_{i=1}^{\ell} \hat{\gamma}(\boldsymbol{x}_t, \alpha \boldsymbol{s}_j)^2,$$

and

$$\sigma_t^2 \geq \frac{\alpha^2}{2(\ell-1)} \sum_{i=1}^{\ell}\left(\left\langle \nabla f(\boldsymbol{x}_t;\ \xi_t), \boldsymbol{s}_i\right\rangle^2-\frac{1}{\ell^2}\left(\sum_{i=1}^{\ell}\left\langle \nabla f(\boldsymbol{x}_t;\ \xi_t), \boldsymbol{s}_i\right\rangle\right)^2\right)-\frac{1}{\ell-1} \sum_{i=1}^{\ell} \hat{\gamma}(\boldsymbol{x}_t, \alpha \boldsymbol{s}_j)^2,$$

其中定义

$$\hat{\gamma}(\boldsymbol{x}, \alpha \boldsymbol{s}_i)=f(\boldsymbol{x}+\alpha \boldsymbol{s}_i ; \xi)-f(\boldsymbol{x} ; \xi)-\alpha\left\langle \nabla f(\boldsymbol{x} ; \xi), \boldsymbol{s}_i\right\rangle. \tag{4.8}$$

证明。 该证明与引理 8 的证明相同。 □

引理20。 令 $\hat{\gamma}(\boldsymbol{x}, \alpha \boldsymbol{s}_i)$ 由公式 (4.8)定义。在事件 $\mathcal{E}_{t,1}$ 和假设 4下，且所有 $\boldsymbol{x}$ 和 $\boldsymbol{x}+\alpha \boldsymbol{s}_i$ 均属于 $\mathcal{K}$，则有

$$\sum_{i=1}^{\ell} \hat{\gamma}(\boldsymbol{x}, \alpha \boldsymbol{s}_i)^2 \leq \frac{\alpha^4 \hat{L}^2 \hat{\zeta}_1}{4},$$

ˆ其中 ζ_1 由公式 (4.6)定义。

证明。根据$\gamma(\cdot,\ \cdot)$ 在公式 (4.8中的定义，我们有

$$\hat{\gamma}(\boldsymbol{x}, \alpha \boldsymbol{s}_i) = f(\boldsymbol{x} + \alpha \boldsymbol{s}_i; \xi) - f(\boldsymbol{x}; \xi) - \alpha \left\langle \nabla f(\boldsymbol{x}; \xi), \boldsymbol{s}_i \right\rangle = \frac{\alpha^2}{2} \boldsymbol{s}_i^\top \nabla^2 f(\boldsymbol{x}'_{\boldsymbol{s}_i}; \xi) \boldsymbol{s}_i$$

其中， $\xi'_{\boldsymbol{s}_i}$ 是位于 $\boldsymbol{x}$ 与 $\boldsymbol{x} + \alpha \boldsymbol{s}_i$之间直线上的一点。根据假设 4，我们可以得到

$$|\boldsymbol{s}_i^\top \nabla^2 f(\boldsymbol{x}'_{\boldsymbol{s}_i}) \boldsymbol{s}_i| \leq \hat{L} \left\| \boldsymbol{s}_i \right\|^2.$$

因此，我们可以得到

$$\sum_{i=1}^{\ell} \hat{\gamma}(\boldsymbol{x}, \alpha \boldsymbol{s}_i)^2 \leq \frac{\alpha^4 \hat{L}^2}{4} \sum_{i=1}^{\ell} \left\| \boldsymbol{s}_i \right\|^4 = \frac{\alpha^4 \hat{L}^2 \hat{\zeta}_1}{4}.$$

□

引理21。假设 $\beta = \alpha\nu/(4\sqrt{\ell - 1})$和 $\alpha^2 \leq \nu^2/(8\hat{}L^2\hat{}\zeta_1)$满足ˆ ζ_1在公式 (4.6)中定义，且ˆ在事件 $\mathcal{E}_t$条件下，3 在公式 (4.4)中定义，则有

$$\sigma_t + \beta \geq \sqrt{\frac{\alpha^2}{8(\ell - 1)} \cdot \left\| \nabla f(\boldsymbol{x}_t;\ \xi_t) \right\|^2 + \frac{\beta^2}{2}}, \tag{4.9}$$

$$\sigma_t \leq \frac{3\alpha}{2\sqrt{\ell - 1}} \left\| \nabla f(\boldsymbol{x}_t;\ \xi_t) \right\| + \sqrt{\frac{3\alpha^4 \hat{L}^2 \hat{\zeta}_1}{4(\ell - 1)}}. \tag{4.10}$$

ˆ证明。结合引理 19 与事件$\mathcal{E}_t$, 3，我们可以得到

$$\sigma_t^2 \geq \frac{\alpha^2}{2(\ell - 1)} \cdot \frac{\left\| \nabla f(\boldsymbol{x}_t;\ \xi_t) \right\|^2}{4} - \frac{1}{\ell - 1} \sum_{i=1}^{\ell} \hat{\gamma}(\boldsymbol{x}, \alpha \boldsymbol{s}_j)^2.$$

根据引理 20，我们可以得到

$$\sigma_t^2 \geq \frac{\alpha^2}{8(\ell - 1)} \cdot \left\| \nabla f(\boldsymbol{x}_t;\ \xi_t) \right\|^2 - \frac{\alpha^4 \hat{L}^2 \hat{\zeta}_1}{4(\ell - 1)}.$$

此外，

$$\begin{aligned}
(\sigma_t + \beta)^2 \geq & \sigma_t^2 + \beta^2 \geq \frac{\alpha^2}{8(\ell - 1)} \cdot \left\| \nabla f(\boldsymbol{x}_t;\ \xi_t) \right\|^2 - \frac{\alpha^4 \hat{L}^2 \hat{\zeta}_1}{4(\ell - 1)} + \beta^2 \\
= & \frac{\alpha^2}{8(\ell - 1)} \cdot \left\| \nabla f(\boldsymbol{x}_t;\ \xi_t) \right\|^2 - \frac{\alpha^4 \hat{L}^2 \hat{\zeta}_1}{4(\ell - 1)} + \frac{\alpha^2 \nu^2}{16(\ell - 1)} \\
\geq & \frac{\alpha^2}{8(\ell - 1)} \cdot \left\| \nabla f(\boldsymbol{x}_t;\ \xi_t) \right\|^2 + \frac{\beta^2}{2},
\end{aligned}$$

其中，最后一个不等式是因为 $\alpha^2 \leq \frac{\nu^2}{8L^2\hat{\zeta}_1}$。

类似地，结合引理 19 与引理 20，我们有

$$\begin{aligned}\sigma_t^2 \leq & \frac{3\alpha^2}{2(\ell-1)}\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}_t;\,\xi_t),\boldsymbol{s}_i\rangle^2+\frac{3}{\ell-1}\sum_{i=1}^{\ell}\hat{\gamma}(\boldsymbol{x}_t,\alpha\boldsymbol{s}_j)^2\\ \overset{(4.3)}{\leq} & \frac{9\alpha^2}{4(\ell-1)}\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|^2+\frac{3}{\ell-1}\sum_{i=1}^{\ell}\hat{\gamma}(\boldsymbol{x}_t,\alpha\boldsymbol{s}_j)^2\\ \leq & \frac{9\alpha^2}{4(\ell-1)}\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|^2+\frac{3\alpha^4\hat{L}^2\hat{\zeta}_1}{4(\ell-1)},\end{aligned}$$

这意味着

$$\sigma_t \leq \frac{3\alpha}{2\sqrt{\ell-1}}\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|+\sqrt{\frac{3\alpha^4\hat{L}^2\hat{\zeta}_1}{4(\ell-1)}}.$$

□

引理22. 设事件 $\hat{\mathcal{E}}_{t,1}$、$\hat{\mathcal{E}}_{t,2}$ 和 $\mathcal{E}_{t,3}$ 成立。令 $\beta=\frac{\alpha\nu}{4\sqrt{\ell-1}}$ 和 $\alpha^2\leq\frac{\nu^2}{8L^2\hat{\zeta}_1}$，则有

$$\frac{\left\|\hat{\mathbf{g}}(\boldsymbol{x}_t)\right\|_{\boldsymbol{H}}}{\sigma_t+\beta}\leq\frac{2\sqrt{6\zeta_2(\ell-1)}}{\alpha}+\frac{2\hat{L}\sqrt{(\ell-1)\zeta_2\hat{\zeta}_1}}{\nu}. \tag{4.11}$$

证明。 首先，我们有

$$\begin{aligned}\frac{\left\|\hat{\mathbf{g}}(\boldsymbol{x}_t)\right\|_{\boldsymbol{H}}}{\sigma_t+\beta} \overset{(4.9)}{\leq} & \frac{\left\|\hat{\mathbf{g}}(\boldsymbol{x}_t)\right\|_{\boldsymbol{H}}}{\sqrt{\frac{\alpha^2}{8(\ell-1)}\cdot\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|^2+\frac{\beta^2}{2}}}\leq\frac{4\sqrt{\ell-1}\left\|\hat{\mathbf{g}}(\boldsymbol{x}_t)\right\|_{\boldsymbol{H}}}{\alpha\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|+4\sqrt{\ell-1}\beta}\\ \overset{(4.7)}{\leq} & \frac{4\sqrt{\ell-1}\left(\left\|\boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|_{\boldsymbol{H}}+\frac{\alpha}{2}\left\|\boldsymbol{S}\hat{\boldsymbol{v}}\right\|_{\boldsymbol{H}}\right)}{\alpha\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|+4\sqrt{\ell-1}\beta}\\ \leq & \frac{\left\|\boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|_{\boldsymbol{H}}}{\frac{\alpha}{4\sqrt{(\ell-1)}}\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|}+\frac{\frac{\alpha}{2}\left\|\boldsymbol{S}\hat{\boldsymbol{v}}\right\|_{\boldsymbol{H}}}{\beta}.\end{aligned}$$

此外，

$$\begin{aligned}\left\|\boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|_{\boldsymbol{H}}=&\sqrt{\nabla^\top f(\boldsymbol{x}_t;\,\xi_t)\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{H}\boldsymbol{S}\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t;\,\xi_t)}\\ \overset{(3.3)}{\leq}&\sqrt{\zeta_2}\left\|\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|\overset{(4.3)}{\leq}\sqrt{\frac{3\zeta_2}{2}}\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|.\end{aligned}$$

类似地，

$$\left\|\boldsymbol{S}\hat{\boldsymbol{v}}\right\|_{\boldsymbol{H}}=\sqrt{\hat{\boldsymbol{v}}^\top\boldsymbol{S}^\top\boldsymbol{H}\boldsymbol{S}\hat{\boldsymbol{v}}}\leq\sqrt{\zeta_2}\left\|\hat{\boldsymbol{v}}\right\|\overset{(4.7)}{\leq}\hat{L}\sqrt{\zeta_2\hat{\zeta}_1}.$$

结合上述结果，我们可以得到

$$\begin{aligned}\frac{\|\hat{\mathbf{g}}(\boldsymbol{x}_t)\|_{\boldsymbol{H}}}{\sigma_t+\beta} \leq & \frac{\sqrt{\frac{3\zeta_2}{2}}\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\|}{\frac{\alpha}{4\sqrt{(\ell-1)}}\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\|}+\frac{\frac{\alpha}{2}\hat{L}\sqrt{\zeta_2\hat{\zeta}_1}}{\beta} \\ = & \frac{2\sqrt{6\zeta_2(\ell-1)}}{\alpha}+\frac{2\hat{L}\sqrt{(\ell-1)\zeta_2\hat{\zeta}_1}}{\nu},\end{aligned}$$

其中，最后一个等式是因为 $\beta=\frac{\alpha\nu}{4\sqrt{\ell-1}}$。 □

4.3 下降分析与主要结果

引理23。 设函数 $f(\boldsymbol{x})$ 和 $f(\boldsymbol{x};\ \xi)$ 满足假设 1 和假设4。矩阵$\boldsymbol{S}\in\mathbb{R}^{d\times\ell}$是一个无偏（1 4, k,δ）随机草图矩阵。假设事件 $\hat{\mathcal{E}}_t$,1– $\hat{\mathcal{E}}_t$,3 和 $\mathcal{E}_t$,2– $\mathcal{E}_t$,3 成立。令 $\beta=\frac{\alpha\nu}{4\sqrt{\ell-1}}$, $\alpha^2\leq\frac{\nu^2}{8L^2\hat{}\hat{\zeta}_1}$，步长为 $\eta\leq\frac{\alpha}{700\cdot\sqrt{\ell-1}\zeta_2L_1}$。经过算法 2 的一步更新后，有

$$\begin{aligned}f(\boldsymbol{x}_{t+1})\leq & f(\boldsymbol{x}_t)-\frac{\eta\|\nabla f(\boldsymbol{x}_t)\|^2}{4(\sigma_t+\beta)}-\eta\cdot\frac{\left\langle\boldsymbol{S}^\top\left(\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x};\,\xi_t)\right),\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\rangle}{\sigma_t+\beta} \\ & +\frac{5\eta\alpha^2\hat{L}^2\hat{\zeta}_1}{32\beta}+\frac{\eta^2\zeta_2(L_0+L_1\|\nabla f(\boldsymbol{x}_t)\|)}{2}\cdot\left(\frac{48(\ell-1)}{\alpha^2}+\frac{\alpha^2\hat{L}^2\hat{\zeta}_1}{8\beta^2}\right).\end{aligned}\tag{4.12}$$

证明。 首先，根据算法 2 的更新规则，我们有

$$\begin{aligned}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|_H= & \eta\frac{\|\hat{\mathbf{g}}(\boldsymbol{x}_t)\|_{\boldsymbol{H}}}{\sigma_t+\beta}\overset{(4.11)}{\leq}\eta\left(\frac{2\sqrt{6\zeta_2(\ell-1)}}{\alpha}+\frac{2\hat{L}\sqrt{(\ell-1)\zeta_2\hat{\zeta}_1}}{\nu}\right) \\ = & \frac{\sqrt{6}}{350L_1\sqrt{\zeta_2}}+\frac{\alpha\hat{L}\sqrt{\hat{\zeta}_1}}{350L_1\sqrt{\zeta_2}\nu}\leq\frac{\sqrt{6}}{350L_1\sqrt{\zeta_2}}+\frac{1}{350\sqrt{8}L_1\sqrt{\zeta_2}}\leq\frac{1}{L_1\sqrt{\zeta_2}}\leq\frac{1}{L_1\sqrt{\|\boldsymbol{H}\|}},\end{aligned}$$

其中，第二个不等式是因为条件 $\alpha^2\leq\frac{\nu^2}{8L^2\hat{}\hat{\zeta}_1}$ 最后一个不等式则是因为 ζ_2 在公式（3.6）中的定义。因此，假设 1 中的条件成立。接着，根据

$(L_0,\ L_1)$-光滑性，以及算法 2的一步更新 $f(\boldsymbol{x})$，我们有

$$
\begin{aligned}
& f(\boldsymbol{x}_{t+1}) \\
\overset{(2.3)}{\leq} & f(\boldsymbol{x}_t)+\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1}-\boldsymbol{x}_t\rangle + \frac{L_0+L_1\|\nabla f(\boldsymbol{x}_t)\|}{2}(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t)^\top \boldsymbol{H}(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t) \\
= & f(\boldsymbol{x}_t)-\frac{\eta}{\sigma_t+\beta}\langle \nabla f(\boldsymbol{x}_t), \hat{\mathbf{g}}(\boldsymbol{x}_t)\rangle + \frac{\eta^2}{(\sigma_t+\beta)^2}\frac{L_0+L_1\|\nabla f(\boldsymbol{x}_t)\|}{2}\hat{\mathbf{g}}(\boldsymbol{x}_t)^\top \boldsymbol{H}\hat{\mathbf{g}}(\boldsymbol{x}_t) \\
\overset{(4.7)}{=} & f(\boldsymbol{x}_t)-\frac{\eta}{\sigma_t+\beta}\left\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t;\ \xi_t)+\frac{\alpha}{4}\boldsymbol{S}\hat{\boldsymbol{v}}_k\right\rangle \\
& + \frac{\eta^2}{(\sigma_t+\beta)^2}\frac{L_0+L_1\|\nabla f(\boldsymbol{x}_t)\|}{2}\left(\boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t;\ \xi_t)+\frac{\alpha}{4}\boldsymbol{S}\hat{\boldsymbol{v}}_k\right)^\top \boldsymbol{H}\left(\boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t;\ \xi_t)+\frac{\alpha}{4}\boldsymbol{S}\hat{\boldsymbol{v}}_k\right).
\end{aligned}
$$

此外，

$$
\begin{aligned}
& -\left\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t;\ \xi_t)+\frac{\alpha}{4}S\hat{\boldsymbol{v}}_k\right\rangle \\
= & -\left\langle \boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t), \boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\rangle - \left\langle \boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t), \boldsymbol{S}^\top\left(\nabla f(\boldsymbol{x}_t;\ \xi_t)-\nabla f(\boldsymbol{x}_t)\right)\right\rangle - \frac{\alpha}{4}\left\langle \boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t), \hat{\boldsymbol{v}}_k\right\rangle \\
\leq & -\frac{\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2}{2} - \left\langle \boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t), \boldsymbol{S}^\top\left(\nabla f(\boldsymbol{x}_t;\ \xi_t)-\nabla f(\boldsymbol{x}_t)\right)\right\rangle + \frac{\alpha^2\left\|\hat{\boldsymbol{v}}_k\right\|^2}{8}.
\end{aligned}
$$

此外，还有

$$
\left(\boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t;\ \xi_t)+\frac{\alpha}{4}S\hat{\boldsymbol{v}}_k\right)^\top \boldsymbol{H}\left(\boldsymbol{S}\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t;\ \xi_t)+\frac{\alpha}{4}S\hat{\boldsymbol{v}}_k\right) \leq \left\|\boldsymbol{S}^\top \boldsymbol{H}\boldsymbol{S}\right\|_2 \cdot \left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t;\ \xi_t)+\frac{\alpha}{4}\hat{\boldsymbol{v}}_k\right\|^2
$$

因此，

$$
\begin{aligned}
f(\boldsymbol{x}_{t+1}) \leq & f(\boldsymbol{x}_t) - \frac{\eta\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\|^2}{2(\sigma_t+\beta)} - \eta\cdot\frac{\left\langle \boldsymbol{S}^\top\left(\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x};\ \xi_t)\right), \boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t)\right\rangle}{\sigma_t+\beta} \\
& + \frac{\eta\alpha^2}{8(\sigma_t+\beta)}\left\|\hat{\boldsymbol{v}}_k\right\|^2 + \frac{\eta\alpha^2}{32(\sigma_t+\beta)}\left\|\hat{\boldsymbol{v}}_k\right\|^2 \\
& + \frac{\eta^2}{(\sigma_t+\beta)^2}\frac{L_0+L_1\|\nabla f(\boldsymbol{x}_t)\|}{2}\left\|\boldsymbol{S}^\top \boldsymbol{H}\boldsymbol{S}\right\|_2 \cdot \left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t;\ \xi_t)+\frac{\alpha}{4}\hat{\boldsymbol{v}}_k\right\|^2
\end{aligned}
$$

结合 $\hat{\mathbf{g}}$ 引理18，引理21以及事件$\mathcal{E}_{t,\ 2}$ 由公式（4.3）定义，我们可以得到

$$
\begin{aligned}
& \frac{\left(\left\|\boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t;\ \xi_t)\right\|+\frac{\alpha}{4}\left\|\hat{\boldsymbol{v}}_k\right\|\right)}{\sigma_t+\beta} \overset{(4.3)(4.9)(4.7)}{\leq} \frac{\sqrt{\frac{3}{2}}\left\|\nabla f(\boldsymbol{x}_t;\ \xi_t)\right\|+\frac{\alpha\hat{L}\sqrt{\hat{\zeta}_1}}{4}}{\sqrt{\frac{\alpha^2}{8(\ell-1)}\cdot\left\|\nabla f(\boldsymbol{x}_t;\ \xi_t)\right\|^2+\frac{\beta^2}{2}}} \\
\leq & \frac{\sqrt{\frac{3}{2}}\left\|\nabla f(\boldsymbol{x}_t;\ \xi_t)\right\|}{\frac{\alpha\|\nabla f(\boldsymbol{x}_t;\ \xi_t)\|}{4\sqrt{\ell-1}}} + \frac{\alpha\hat{L}\sqrt{\hat{\zeta}_1}}{4\beta} = \frac{2\sqrt{6}\sqrt{\ell-1}}{\alpha} + \frac{\alpha\hat{L}\sqrt{\hat{\zeta}_1}}{4\beta}.
\end{aligned}
$$

因此，我们可以得到

$$\begin{aligned}&\frac{\left(\boldsymbol{S}\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t;\,\xi_t)+\frac{\alpha}{4}S\hat{\boldsymbol{v}}_k\right)^{\top}H\left(\boldsymbol{S}\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t;\,\xi_t)+\frac{\alpha}{4}S\hat{\boldsymbol{v}}_k\right)}{(\sigma_t+\beta)^2}\\ \leq&\zeta_2\left(\frac{2\sqrt{6}\sqrt{\ell-1}}{\alpha}+\frac{\alpha\hat{L}\sqrt{\hat{\zeta}_1}}{4\beta}.\right)^2\leq\zeta_2\left(\frac{48(\ell-1)}{\alpha^2}+\frac{\alpha^2\hat{L}^2\hat{\zeta}_1}{8\beta^2}\right),\end{aligned}$$

其中第一个不等式同样是因为事件$\mathcal{E}_{t,3}$ 定义于公式 (3.3)。

此外，

$$\frac{\eta\alpha^2}{\sigma_t+\beta}\left\|\hat{\boldsymbol{v}}_k\right\|^2\overset{(4.7)}{\leq}\frac{\eta\alpha^2\hat{L}^2\hat{\zeta}_1}{\beta}.$$

因此，

$$\begin{aligned}f(\boldsymbol{x}_{t+1})\leq&f(\boldsymbol{x}_t)-\frac{\eta\left\|\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\|^2}{2(\sigma_t+\beta)}-\eta\cdot\frac{\left\langle\boldsymbol{S}^{\top}\left(\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x};\,\xi_t)\right),\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\rangle}{\sigma_t+\beta}\\&+\frac{5\eta\alpha^2\hat{L}^2\hat{\zeta}_1}{32\beta}+\frac{\eta^2\zeta_2(L_0+L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|)}{2}\cdot\left(\frac{48(\ell-1)}{\alpha^2}+\frac{\alpha^2\hat{L}^2\hat{\zeta}_1}{8\beta^2}\right)\\\leq&f(\boldsymbol{x}_t)-\frac{\eta\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{4(\sigma_t+\beta)}-\eta\cdot\frac{\left\langle\boldsymbol{S}^{\top}\left(\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x};\,\xi_t)\right),\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\rangle}{\sigma_t+\beta}\\&+\frac{5\eta\alpha^2\hat{L}^2\hat{\zeta}_1}{32\beta}+\frac{\eta^2\zeta_2(L_0+L_1\left\|\nabla f(\boldsymbol{x}_t)\right\|)}{2}\cdot\left(\frac{48(\ell-1)}{\alpha^2}+\frac{\alpha^2\hat{L}^2\hat{\zeta}_1}{8\beta^2}\right),\end{aligned}$$

其中最后一个不等式是因为事件$\mathcal{E}_{t,2}$。 □

引理*24.* 设 $\nabla f(\boldsymbol{x};\,\xi)$满足假设3。记$\hat{}\,\mathcal{E}_t:=\left(\cap_{i=1}^4\hat{}\,\mathcal{E}_{t,i}\right)\cap\mathcal{E}_{t,2}\cap\mathcal{E}_{t,3}$，这意味着$\hat{}\,\hat{}\,\hat{}$事件$\mathcal{E}_{t,1}$–$\mathcal{E}_{t,4}$与$\mathcal{E}_{t,2}$–$\mathcal{E}_{t,3}$成立。假设 $\mathcal{E}_t$ 对 $t=0,\ldots,T-1$成立。令$\beta=\frac{\alpha\nu}{4\sqrt{\ell-1}}$与$\alpha^2\leq\frac{\nu^2}{8L^2}\hat{}\,\hat{}\,\zeta_1$，其中$\hat{}\,\zeta_1$ 定义于公式 *(4.6)*，则以至少为 $1-\delta$的概率，有

$$\frac{1}{T}\sum_{k=1}^{T}\frac{\left\langle\boldsymbol{S}^{\top}\left(\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t;\,\xi_t)\right),\boldsymbol{S}^{\top}\nabla f(\boldsymbol{x}_t)\right\rangle}{\sigma_t+\beta}\leq\frac{1}{\sqrt{T}}\cdot\frac{9\sqrt{2}\sqrt{\ell-1}\nu}{\alpha}\sqrt{\log\frac{2}{\delta}}.\tag{4.13}$$

证明。 首先，由引理 21，我们有

$$\begin{aligned}\frac{1}{\sigma_t+\beta}\overset{(4.9)}{\leq}&\frac{1}{\sqrt{\frac{\alpha^2}{8(\ell-1)}\cdot\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|^2+\frac{\beta^2}{2}}}\leq\frac{\sqrt{2}}{\sqrt{\frac{\alpha^2}{8(\ell-1)}\cdot\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|^2}+\sqrt{\frac{\beta^2}{2}}}\\=&\frac{4\sqrt{\ell-1}}{\alpha\left\|\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|+4\sqrt{\ell-1}\beta}\leq\frac{4\sqrt{\ell-1}}{\alpha(\left\|\nabla f(\boldsymbol{x}_t)\right\|-\nu+4\sqrt{\ell-1}\beta/\alpha)}=\frac{4\sqrt{\ell-1}}{\alpha\left\|\nabla f(\boldsymbol{x}_t)\right\|},\end{aligned}\tag{4.14}$$

其中最后一个等式是因为 $\beta=\frac{\alpha\nu}{4\sqrt{\ell-1}}$，第二个不等式是因为如下事实：

$a^2+b^2 \geq \frac{(a+b)^2}{2}$ 对所有 $a,\ b \in \mathbb{R}$，且最后一个不等式是因为假设 3。于是，我们有

$$\begin{aligned}
&\left|\frac{\left\langle \boldsymbol{S}^\top\left(\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t;\,\xi_t)\right), \boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\rangle}{\sigma_t+\beta}\right| \\
\leq &\frac{\left\|\boldsymbol{S}^\top\left(\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t;\,\xi_t)\right)\right\|\cdot\left\|\boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\|}{\sigma_t+\beta} \\
\overset{(3.2)(4.5)}{\leq} &\frac{9\left\|\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|\left\|\nabla f(\boldsymbol{x}_t)\right\|}{4(\sigma_t+\beta)} \\
\overset{(4.14)}{\leq} &\frac{9\sqrt{\ell-1}\left\|\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t;\,\xi_t)\right\|}{\alpha} \leq \frac{9\sqrt{\ell-1}\nu}{\alpha},
\end{aligned}$$

其中，最后一个不等式是由假设 3得出的。

根据引理 25 中的 Azuma-Hoeffding 不等式，结合 $E_t = \frac{\left\langle \boldsymbol{S}^\top\left(\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t;\,\xi_t)\right), \boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\rangle}{\sigma_t+\beta}$、$B = \frac{9\sqrt{\ell-1}\nu}{\alpha}$ 和 $\zeta = \frac{1}{\sqrt{T}}\cdot\frac{9\sqrt{2\ell-1}\nu}{\alpha}\sqrt{\log\frac{2}{\delta}}$，我们可以得到

$$\mathbb{P}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{\left\langle \boldsymbol{S}^\top\left(\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x};\,\xi_t)\right), \boldsymbol{S}^\top\nabla f(\boldsymbol{x}_t)\right\rangle}{\sigma_t+\beta} \geq \zeta\right] \leq \exp\left(-\frac{T\zeta^2}{2\left(\frac{9\sqrt{\ell-1}\nu}{\alpha}\right)^2}\right) = \delta,$$

由此完成证明。 □

定理 *3.* 设函数 $f(x)$ 和 $f(x;\ \xi)$ 满足假设 *1–4*。矩阵 $\boldsymbol{S}_t \in \mathbb{R}^{d\times\ell}$ 是一个无偏的 $(1/4, k, \delta)$ 随机草图矩阵。记 $\hat{}\ \mathcal{E}_t := \left(\cap_{i=1}^4\ \hat{}\mathcal{E}_{t,i}\right)\cap\mathcal{E}_{t,2}\cap\mathcal{E}_{t,3}$，这意味着 $\hat{}\ \hat{}\ \hat{}\ \hat{}$ 事件 $\mathcal{E}_t$*,1*, $\mathcal{E}_t$*,1*, $\mathcal{E}_t$*,4* 和 $\mathcal{E}_t$*,4*（定义见式 *(4.2)*–式 *(4.5)*）以及 $\mathcal{E}_t$*,2* 和 $\mathcal{E}_t$*,3*（定义见式 *(3.2)*– 式 *(3.3)*）均成立。将算法 *2* 的参数设置如下：

$$\alpha = \min\left\{\frac{1}{T^{1/2}}, \sqrt{\frac{\nu^2}{8\hat{L}^2\hat{\zeta}_1}}\right\}, \qquad \eta = \frac{\alpha}{700\sqrt{T}\cdot\sqrt{\ell-1}\zeta_2 L_1}, \qquad \textit{and} \qquad \beta = \frac{\alpha\nu}{4\sqrt{\ell-1}}, \tag{4.15}$$

那么序列 $\{\boldsymbol{x}_t\}$（其中 $t = 0,\ldots,T-1$ 由算法 *2* 生成）以至少 $1-\delta$ 的概率 $0<\delta<1$满足以下性质，

$$\begin{aligned}
\min_{t=0,\ldots,T-1}\left\|\nabla f(\boldsymbol{x}_t)\right\| \leq \max\Bigg\{&\frac{1}{T^{1/4}}, \\
&\frac{28\cdot 700\cdot\nu\sqrt{\zeta_2}L_1\cdot\left(f(\boldsymbol{x}_0)-f(\boldsymbol{x}^*)\right)}{T^{1/4}} + \frac{252\sqrt{2}\nu^2}{T^{1/4}}\sqrt{\log\frac{2}{\delta}} + \frac{35\hat{L}^2\hat{\zeta}_1}{2T^{1/4}} + \frac{L_0\nu\sqrt{\zeta_2}}{T^{1/4}L_1}\Bigg\}.
\end{aligned} \tag{4.16}$$

此外，还有

$$\mathbb{P}\left(\cap_{t=0}^{T-1}\hat{\mathcal{E}}_t\right) \geq 1-6T\delta. \tag{4.17}$$

证明。 我们首先考虑 $\min\left\{\left\|\nabla f(\boldsymbol{x}_0)\right\|,\ldots,\left\|\nabla f(\boldsymbol{x}_{T-1})\right\|\right\} \leq \frac{1}{T^{1/4}}$ 的情形。如果这个结果

成立，那么我们就证明了公式 (4.16) 中的结果。如果不成立，我们将证明以下情形:

$$\min\left\{\left\|\nabla f(\boldsymbol{x}_0)\right\|, \ldots, \left\|\nabla f(\boldsymbol{x}_{T-1})\right\|\right\} \geq \frac{1}{T^{1/4}}. \tag{4.18}$$

为方便记号，我们用 ϕ 表示 $1/T^{1/4}$。于是，根据假设3，可得

$$\sigma_t + \beta \overset{(4.10)}{\leq} \sqrt{\frac{5}{2(\ell-1)}}\alpha \left\|\nabla f(\boldsymbol{x}_t;\, \xi_t)\right\| + \beta \leq \sqrt{\frac{5}{2(\ell-1)}}\alpha \left(\left\|\nabla f(\boldsymbol{x}_t)\right\| + \nu\right) + \beta,$$

其中，最后一个不等式是由假设3得出的。

因此，我们可以得到

$$-\frac{\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{\sigma_t + \beta} \leq -\frac{\left\|\nabla f(\boldsymbol{x}_t)\right\|^2}{\sqrt{\frac{5}{2(\ell-1)}}\alpha \left(\left\|\nabla f(\boldsymbol{x}_t)\right\| + \nu\right) + \beta}.$$

此外，

$$\frac{\left\|\nabla f(\boldsymbol{x}_t)\right\|}{\sqrt{\frac{5}{2(\ell-1)}}\left(\left\|\nabla f(\boldsymbol{x}_t)\right\| + \nu\right) + \beta/\alpha} \overset{(4.18)}{\geq} \frac{\phi}{\sqrt{\frac{5}{2(\ell-1)}}\left(\phi + \nu\right) + \beta/\alpha} \geq \frac{2\sqrt{\ell-1}\phi}{7\nu},$$

其中，最后一个不等式是由于 β 的设置满足

$$\beta = \frac{\alpha\nu}{4\sqrt{\ell-1}} \leq \alpha \cdot \left(\frac{7-\sqrt{10}}{2\sqrt{\ell-1}}\nu - \frac{\sqrt{10}}{2\sqrt{\ell-1}}\phi\right), \text{ with } \phi \leq \sqrt{\frac{5}{2}} - 1.$$

此外，有

$$\frac{48(\ell-1)}{\alpha^2} + \frac{\alpha^2\hat{L}^2\hat{\zeta}_1}{8\beta^2} = \frac{48(\ell-1)}{\alpha^2} + \frac{2(\ell-1)\hat{L}^2\hat{\zeta}_1}{\nu^2} \leq \frac{50(\ell-1)}{\alpha^2},$$

其中第一个等式 性是因为 β，最后一个不等式是因为 $\alpha^2 \leq {}^{\nu^2}\hat{L}^{2}\hat{\zeta}_1$ ——.

结合公式（4.12）以及 β的值，我们可以得到

$$
\begin{aligned}
f(\boldsymbol{x}_{t+1}) \leq & f(\boldsymbol{x}_t) - \frac{\eta}{4\alpha} \cdot \left(\frac{2\sqrt{\ell-1}\phi}{7\nu} - \frac{100\eta(\ell-1)\zeta_2 L_1}{\alpha} \right) \|\nabla f(\boldsymbol{x}_t)\| \\
& - \eta \cdot \frac{\left\langle \boldsymbol{S}^\top \left(\nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x};\, \xi_t) \right), \boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t) \right\rangle}{\sigma_t + \beta} \\
& + \frac{5\eta\sqrt{\ell-1}\alpha \hat{L}^2 \hat{\zeta}_1}{8\nu} + \frac{50\eta^2(\ell-1)\zeta_2 L_0}{\alpha^2} \\
= & f(\boldsymbol{x}_t) - \frac{\eta}{4\alpha} \cdot \frac{\sqrt{\ell-1}\phi}{7\nu} \cdot \|\nabla f(\boldsymbol{x}_t)\| \\
& - \eta \cdot \frac{\left\langle \boldsymbol{S}^\top \left(\nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x};\, \xi_t) \right), \boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t) \right\rangle}{\sigma_t + \beta} \\
& + \frac{5\eta\sqrt{\ell-1}\alpha \hat{L}^2 \hat{\zeta}_1}{8\nu} + \frac{25\eta^2(\ell-1)\zeta_2 L_0}{\alpha^2},
\end{aligned}
$$

其中，最后一个不等式是由设定 $\eta \leq \frac{\alpha\phi}{700\sqrt{\ell-1}\zeta_2 L_1}$ 得出的。

对上面的等式进行裂项求和，可得

$$
\begin{aligned}
& \frac{\eta}{4\alpha} \cdot \frac{\sqrt{\ell-1}\phi}{7\nu} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\| \\
\leq & \frac{f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)}{T} - \frac{\eta}{T} \cdot \sum_{t=0}^{T-1} \frac{\left\langle \boldsymbol{S}^\top \left(\nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x};\, \xi_t) \right), \boldsymbol{S}^\top \nabla f(\boldsymbol{x}_t) \right\rangle}{\sigma_t + \beta} \\
& + \frac{5\eta\sqrt{\ell-1}\alpha \hat{L}^2 \hat{\zeta}_1}{8\nu} + \frac{25\eta^2(\ell-1)\zeta_2 L_0}{\alpha^2} \\
\overset{(4.13)}{\leq} & \frac{f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)}{T} + \frac{\eta}{\sqrt{T}} \cdot \frac{9\sqrt{2}\sqrt{\ell-1}\nu}{\alpha} \sqrt{\log \frac{2}{\delta}} \\
& + \frac{5\eta\sqrt{\ell-1}\alpha \hat{L}^2 \hat{\zeta}_1}{8\nu} + \frac{25\eta^2(\ell-1)\zeta_2 L_0}{\alpha^2}.
\end{aligned}
$$

因此，

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\| \leq & \frac{28\nu\alpha\cdot(f(\boldsymbol{x}_0)-f(\boldsymbol{x}^*))}{\eta\sqrt{\ell-1}\phi T}+\frac{252\sqrt{2}\nu^2}{\phi\sqrt{T}}\sqrt{\log\frac{2}{\delta}} \\ & +\frac{35\alpha^2\hat{L}^2\hat{\zeta}_1}{2\phi}+\frac{700\eta\sqrt{(\ell-1)}\zeta_2 L_0\nu}{\alpha\phi} \\ = & \frac{28\cdot 700\cdot\nu\sqrt{\zeta_2}L_1\cdot(f(\boldsymbol{x}_0)-f(\boldsymbol{x}^*))}{\phi\sqrt{T}}+\frac{252\sqrt{2}\nu^2}{\phi\sqrt{T}}\sqrt{\log\frac{2}{\delta}} \\ & +\frac{35\alpha^2\hat{L}^2\hat{\zeta}_1}{2\phi}+\frac{L_0\nu\sqrt{\zeta_2}}{\phi\sqrt{T}L_1} \\ = & \frac{28\cdot 700\cdot\nu\sqrt{\zeta_2}L_1\cdot(f(\boldsymbol{x}_0)-f(\boldsymbol{x}^*))}{T^{1/4}}+\frac{252\sqrt{2}\nu^2}{T^{1/4}}\sqrt{\log\frac{2}{\delta}} \\ & +\frac{35\alpha^2\hat{L}^2\hat{\zeta}_1 T^{1/4}}{2}+\frac{L_0\nu\sqrt{\zeta_2}}{T^{1/4}L_1} \\ \leq & \frac{28\cdot 700\cdot\nu\sqrt{\zeta_2}L_1\cdot(f(\boldsymbol{x}_0)-f(\boldsymbol{x}^*))}{T^{1/4}}+\frac{252\sqrt{2}\nu^2}{T^{1/4}}\sqrt{\log\frac{2}{\delta}} \\ & +\frac{35\hat{L}^2\hat{\zeta}_1}{2T^{1/4}}+\frac{L_0\nu\sqrt{\zeta_2}}{T^{1/4}L_1},\end{aligned}$$

其中，第一个等号是因为 $\eta=\frac{\alpha}{700\sqrt{T}\cdot\sqrt{\ell-1}\sqrt{\zeta_2}L_1}$，第二个等号是因为 $\phi=\frac{1}{T^{1/4}}$，最后一个不等式是因为 $\alpha\leq T^{-1/2}$。

由于有

$$\min_{t=0,\ldots,T-1}\|\nabla f(\boldsymbol{x}_t)\|\leq\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|,$$

因此有

$$\min_{t=0,\ldots,T-1}\|\nabla f(\boldsymbol{x}_t)\|\leq\frac{28\cdot 700\cdot\nu\sqrt{\zeta_2}L_1\cdot(f(\boldsymbol{x}_0)-f(\boldsymbol{x}^*))}{T^{1/4}}+\frac{252\sqrt{2}\nu^2}{T^{1/4}}\sqrt{\log\frac{2}{\delta}}+\frac{35\hat{L}^2\hat{\zeta}_1}{2T^{1/4}}+\frac{L_0\nu\sqrt{\zeta_2}}{T^{1/4}L_1}.$$

^接下来，我们将证明公式 (4.17)。根据并集界以及 $\mathcal{E}_t$ 的定义，我们有

$$\mathbb{P}\left(\mathcal{E}_t\right)\geq 1-\left(\sum_{i=1}^{4}\mathbb{P}\left(\overline{\hat{\mathcal{E}}_{t,i}}\right)+\sum_{i=2}^{3}\mathbb{P}\left(\overline{\mathcal{E}_{t,i}}\right)\right)\geq 1-(\delta+\delta+\delta+\exp(-\mathcal{O}\left(\ell\right))+\delta+\delta)\geq 1-6\delta,$$

其中 $\overline{\mathcal{E}_{t,\ i}}$ 表示$\mathcal{E}_{t}$的补集i，第二个不等式由引理 14–引理 17和引理 2–引理 3得出，最后一个不等式是因为 $\ell=\mathcal{O}\left(k\text{ 对数 }\frac{1}{\delta}\right)$。

再次使用并集界，可得

$$\mathbb{P}\left(\cap_{t=0}^{T-1}\,\hat{\mathcal{E}}_t\right) \geq 1-\sum_{t=0}^{T-1}\mathbb{P}\left(\overline{\hat{\mathcal{E}}_t}\right) \geq 1-6T\delta,$$

至此，证明完毕。 □

推论 4. 设函数 $f(\boldsymbol{x})$ 和 $f(\boldsymbol{x};\ \xi)$ 满足定理 3中描述的性质。令$\boldsymbol{S}_t \in \mathbb{R}^{d\times\ell}$ 为高斯或拉德马赫草图矩阵，且满足 $\ell = 16k\log(4T/\delta)$ 以及 $\delta\in(0,1)$。设平滑参数 α 和步长 η 满足

$$\alpha^2 \leq \min\left\{\frac{1}{2L_1^2\zeta_1 T},\ \frac{3\varepsilon^2}{8L_0^2\zeta_1}\right\} \qquad and \qquad \eta = \frac{\alpha}{\sqrt{(\ell-1)\zeta_2 T}},$$

其中 $T>0$ 足够大，使得 $\eta \leq 3\alpha/(64L_1\zeta_2\sqrt{\ell}-1)$ 成立，并且 $0<\varepsilon$ 为目标精度。若总迭代次数 T 满足

$$T = \mathcal{O}\left(\max\left\{\frac{1}{\varepsilon^4},\ \frac{\|\boldsymbol{H}\|^2\nu^4}{\varepsilon^4},\ \frac{\left(\mathrm{tr}(\boldsymbol{H})\right)^2}{k^2\varepsilon^4}\right\}\right), \tag{4.19}$$

则以至少 $1-\delta$的概率，序列 $\{\boldsymbol{x}_t\}$ 满足 $t=0,\ldots,T-1$ 由算法 1生成，并满足

$$\min_{t=0,\ldots,T-1}\|\nabla f(\boldsymbol{x}_t)\| \leq \varepsilon.$$

此外，总查询复杂度为

$$Q = \mathcal{O}\left(\max\left\{\frac{k}{\varepsilon^4},\ \frac{k\,\|\boldsymbol{H}\|^2\nu^4}{\varepsilon^4},\ \frac{\left(\mathrm{tr}(\boldsymbol{H})\right)^2\nu^4}{k\varepsilon^4}\right\}\cdot\log\frac{1}{\delta\varepsilon}\right). \tag{4.20}$$

证明。 首先，我们将证明 $\boldsymbol{x}_t\in\mathcal{K}$ 在假设 4中定义。根据 $\boldsymbol{x}_t$ 的定义，我们可以得到

$$\|\boldsymbol{x}_t-\boldsymbol{x}_0\|_{\boldsymbol{H}} = \left\|\boldsymbol{x}_{t-1}-\eta\frac{\hat{\mathbf{g}}(\boldsymbol{x}_{t-1})}{\sigma_{t-1}}-\boldsymbol{x}_0\right\|_{\boldsymbol{H}} \leq \|\boldsymbol{x}_{t-1}-\boldsymbol{x}_0\|_{\boldsymbol{H}}+\eta\left\|\frac{\hat{\mathbf{g}}(\boldsymbol{x}_{t-1})}{\sigma_{t-1}}\right\|_{\boldsymbol{H}} \leq \eta\sum_{i=0}^{t-1}\left\|\frac{\hat{\mathbf{g}}(\boldsymbol{x}_{t-1})}{\sigma_{t-1}}\right\|_{\boldsymbol{H}}.$$

结合公式 (4.11) 和步长的设置，我们可以得到

$$\begin{aligned}\|\boldsymbol{x}_t-\boldsymbol{x}_0\|_{\boldsymbol{H}} \leq & \eta \sum_{i=0}^{t-1}\left(\frac{2\sqrt{6\zeta_2(\ell-1)}}{\alpha}+\frac{2\hat{L}\sqrt{(\ell-1)\zeta_2\hat{\zeta}_1}}{\nu}\right)\\ \leq & T\eta\left(\frac{2\sqrt{6\zeta_2(\ell-1)}}{\alpha}+\frac{2\hat{L}\sqrt{(\ell-1)\zeta_2\hat{\zeta}_1}}{\nu}\right)\\ \leq & T\cdot\frac{\alpha}{700\sqrt{T}\cdot\sqrt{\ell-1}\zeta_2 L_1}\left(\frac{2\sqrt{6\zeta_2(\ell-1)}}{\alpha}+\frac{2\hat{L}\sqrt{(\ell-1)\zeta_2\hat{\zeta}_1}}{\nu}\right)\\ = & \frac{\sqrt{6T}}{350L_1\sqrt{\zeta_2}}+\frac{\hat{L}\alpha\sqrt{\hat{\zeta}_1 T}}{350\sqrt{\zeta_2}L_1\nu}\\ \leq & \frac{\sqrt{6T}}{350L_1\sqrt{\zeta_2}}+\frac{\sqrt{T}}{350\sqrt{8}L_1\sqrt{\zeta_2}}\\ \leq & \frac{3\sqrt{T}}{350L_1\sqrt{\zeta_2}},\end{aligned}$$

其中，第五个不等式是因为 $\alpha=\min\{\frac{1}{T^{1/2}},\sqrt{\frac{\nu^2}{8L^2\hat{\zeta}_1}}\}$。因此，我们得到所有 $\boldsymbol{x}_t$ 都属于集合 $\mathcal{K}$。

接下来，我们将证明 $\boldsymbol{x}_t+\alpha\boldsymbol{s}_i$ 也属于集合 $\mathcal{K}$。

$$\|\boldsymbol{x}_t+\alpha\boldsymbol{s}_i-\boldsymbol{x}_0\|_{\boldsymbol{H}} \leq \|\boldsymbol{x}_t-\boldsymbol{x}_0\|_{\boldsymbol{H}}+\alpha\|\boldsymbol{s}_i\|_{\boldsymbol{H}} \overset{(A.4)}{\leq} \|\boldsymbol{x}_t-\boldsymbol{x}_0\|_{\boldsymbol{H}}+2\alpha\sqrt{\mathrm{tr}(\boldsymbol{H})} \leq \frac{3\sqrt{T}}{350L_1\sqrt{\zeta_2}}+2\alpha\sqrt{\mathrm{tr}(\boldsymbol{H})}.$$

结合 α ≤的设定，我们可以得出 $\boldsymbol{x}_t+\alpha\boldsymbol{s}_i$ 属于 $\mathcal{K}$。

通过设置 公式（4.16）的右侧为 ε，那么我们只需要求 T 满足 t

$$T=\mathcal{O}\left(\max\left\{\frac{1}{\varepsilon^4},\frac{\zeta_2^2\nu^4}{\varepsilon^4}\right\}\right)=\mathcal{O}\left(\max\left\{\frac{1}{\varepsilon^4},\frac{\|\boldsymbol{H}\|^2\nu^4}{\varepsilon^4},\frac{\left(\mathrm{tr}(\boldsymbol{H})\right)^2}{k^2\varepsilon^4}\right\}\right),$$

其中，最后一个等式是根据 ζ_2 公式（3.6）中的定义得出的。

那么，总查询复杂度为

$$\begin{aligned}Q=T\ell=&\mathcal{O}\left(\max\left\{\frac{k}{\varepsilon^4},\frac{k\|\boldsymbol{H}\|^2\nu^4}{\varepsilon^4},\frac{\left(\mathrm{tr}(\boldsymbol{H})\right)^2\nu^4}{k\varepsilon^4}\right\}\cdot\log\frac{T}{\delta}\right)\\ =&\mathcal{O}\left(\max\left\{\frac{k}{\varepsilon^4},\frac{k\|\boldsymbol{H}\|^2\nu^4}{\varepsilon^4},\frac{\left(\mathrm{tr}(\boldsymbol{H})\right)^2\nu^4}{k\varepsilon^4}\right\}\cdot\log\frac{1}{\delta\varepsilon}\right).\end{aligned}$$

□

注 4. 定理 3 和推论 4表明算法2 可以达到

$$\mathcal{O}\left(\max\left\{\frac{1}{\varepsilon^4}, \frac{\|\boldsymbol{H}\|^2\nu^4}{\varepsilon^4}, \frac{\left(\mathrm{tr}(\boldsymbol{H})\right)^2}{k^2\varepsilon^4}\right\}\right)$$

迭代复杂度。如果我们增大 k，也就是增大批量大小 ℓ，算法2 就能以更少的步数找到一个 ε-驻点。同时，公式(4.20) 表明，如果我们增大样本量 ℓ，直到 $\ell = 16k\log\frac{T}{\delta}$ 且 $k = \mathrm{tr}(\boldsymbol{H})\|\boldsymbol{H}\|$，算法1 则有所不同，其查询复杂度无法 从增大批量大小中获益，正如注1中所讨论的。

注 5. 如果我们选择批量大小 $\ell = 16k\log(T/\delta)$ 且 $k = \mathrm{tr}(\boldsymbol{H})/\|\boldsymbol{H}\|$，那么公式 (4.20) 将简化为

$$\begin{aligned} Q =& \mathcal{O}\left(\max\left\{\frac{\mathrm{tr}(\boldsymbol{H})}{\|\boldsymbol{H}\|\varepsilon^4}, \frac{\|\boldsymbol{H}\|\cdot\mathrm{tr}(\boldsymbol{H})\cdot\nu^4}{\varepsilon^4}\right\}\cdot\log\frac{1}{\delta\varepsilon}\right) \\ =& \mathcal{O}\left(\max\left\{\frac{\mathrm{tr}(\boldsymbol{H})}{\|\boldsymbol{H}\|\varepsilon^4}, \frac{\mathrm{tr}(\boldsymbol{H})}{\|\boldsymbol{H}\|}\cdot\frac{\|\boldsymbol{H}\|^2\cdot\nu^4}{\varepsilon^4}\right\}\cdot\log\frac{1}{\delta\varepsilon}\right). \end{aligned}$$

因此，算法2 的查询复杂度也实现了弱维度依赖性，与算法 1 一样。在许多实际应用中，我们可以选择合适的矩阵 $\boldsymbol{H}$，使得 $\frac{\mathrm{tr}(\boldsymbol{H})}{\|\boldsymbol{H}\|} \ll d$。

注 6. 与注 3 中的陈述类似，即使令 $\hat{g}(x_t) = \nabla f(x_t;\xi_t)$，并采用满足 $H = I$ 的假设 1，通过满足 $\eta_t > 0$ 的固定步长更新 $x_{t+1} = x_t - \eta\hat{g}(x_t)$ 来寻找 ε-驻点，所需迭代次数仍取决于 M（Zhang 等人，2020b，定理 8）。相比之下，我们的推论 4 表明，算法 2 所需的迭代次数仅取决于 $\|H\|$ 和 $tr(H)$。

5 结论

在本工作中，我们为自适应零阶优化方法的有效性建立了理论解释。通过刻画函数值变化性与梯度幅值之间的关系，我们表明，采样函数值的经验标准差以高概率可作为（随机）梯度范数的可靠代理。这一结果为将标准差归一化作为零阶优化中一种有原则的自适应步长控制机制提供了严格依据。

基于这一洞见，我们在关于半正定矩阵范数定义的广义（L_0，L_1）光滑性条件下，分析了自适应零阶方法。该设定超越了经典光滑性假设，并捕捉了现代深度神经网络（包括大语言模型）中常见的各向异性几何。我们的分析为确定性目标和随机目标函数分别给出了明确的收敛率和查询复杂度界。

我们的研究结果表明，自适应零阶方法能够通过自动适应优化景观的局部几何，相较于固定步长原始零阶方法实现严格更快的收敛，并显著降低查询复杂度。这些发现弥合了自适应零阶算法的实证实践与理论理解之间的鸿沟。展望未来，该框架为分析更复杂的自适应机制，以及将零阶优化理论扩展到更广泛的非光滑和大规模学习问题提供了新的方向。

参考文献

Borodich, E. & Kovalev, D.（2025）。Nesterov finds graal：面向凸优化的最优自适应梯度法。*arXiv* 预印本 *arXiv:2507.09823*。Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C.-J.（2017）。Zoo：一种无需训练替代模型、基于零阶优化的深度神经网络黑盒攻击。载于第十届 *ACM*人工智能与安全研讨会论文集（第 15–26 页）。：ACM。Chen, Z., Zhou, Y., Liang, Y., & Lu, Z.（2023）。广义光滑非凸优化与光滑非凸优化同样高效。载于 国际机器学习会议（第 5396–5427 页）。Chezhegov, S., Beznosikov, A., Horv´ath, S., & Gorbunov, E.（2025）。重尾噪声下凸（L_0, L_1）-光滑优化中裁剪随机梯度下降的收敛性。*arXiv* 预印本 *arXiv:2505.20817*。Chrabaszcz, P., Loshchilov, I., & Hutter, F.（2018）。回归基础：对用于 Atari 游戏的经典进化策略进行基准测试。*arXiv* 预印本 *arXiv:1802.08842*。Cohen, M. B.（2016）。对稀疏无关子空间嵌入进行更简单、更紧致的分析。载于第 *27* 届 *ACM-SIAM*离散算法研讨会论文集（*SODA*）。Cohen, M. B., Nelson, J., & Woodruff, D. P.（2016）。基于稳定秩的最优近似矩阵乘积。载于 国际自动机、语言与编程研讨会：Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, 达格施图尔出版社。Cooper, Y.（2022）。对（L_0, L_1）-光滑性条件的实证研究。载于 现代机器学习数学研讨会。Cortinovis, A. & Kressner, D.（2021）。不定矩阵的随机迹估计及其在行列式中的应用。计算数学基础, （第 1–29 页）。Crawshaw, M. & Liu, M.（2025）。宽松光滑性条件下非凸随机优化自适应梯度算法的复杂度下界。*arXiv* 预印本 *arXiv:2505.04599*。Crawshaw, M., Liu, M., Orabona, F., Zhang, W., & Zhuang, Z.（2022）。广义SignSGD 对无界光滑性的鲁棒性。载于神经信息处理系统进展（第 9955–9968 页）。Dang, S., Guo, Y., Zhao, Y., Ye, H., Zheng, X., Dai, G., & Tsang, I.（2025）。FZOO：用于微调大语言模型、达到 Adam 级速度的快速零阶优化器。*arXiv* 预印本 *arXiv:2506.09034*。Gaash, O., Levy, K. Y., & Carmon, Y.（2025）。凸（L_0, L_1）-光滑函数上裁剪随机梯度下降的收敛性。*arXiv* 预印本 *arXiv:2502.16492*。

Ghadimi, S. & Lan, G.（2013）。非凸随机规划的随机一阶与零阶方法。*SIAM* 优化期刊，23(4)，2341–2368。Gorbunov, E., Tupitsa, N., Choudhury, S., Aliev, A., Richt´arik, P., Horv´ath, S., & Tak´aˇc, M.（2024）。凸（L_0, L_1）-光滑优化方法：裁剪、加速与自适应性。*arXiv* 预印本 *arXiv:2409.14989*。Ilyas, A., Engstrom, L., Athalye, A., & Lin, J.（2018）。查询与信息有限条件下的黑盒对抗攻击。载于 J. Dy & A. Krause（编），第 *35* 届国际机器学习会议论文集，机器学习研究论文集第 80 卷（第 2137–2146 页）。Stockholmsm¨assan，瑞典斯德哥尔摩：PMLR。Jiang, Z., Balu, A., & Sarkar, S.（2025）。基于梯度下降方法的去中心化松弛光滑优化。*arXiv* 预印本 *arXiv:2508.08413*。Khirirat, S., Sadiev, A., Riabinin, A., Gorbunov, E., & Richt´arik, P.（2024a）。广义光滑性假设下的通信高效算法。$OpenReview$。Khirirat, S., Sadiev, A., Riabinin, A., Gorbunov, E., & Richt´arik, P.（2024b）。（L_0, L_1）-光滑性下的误差反馈：归一化与动量。*arXiv* 预印本 *arXiv:2410.16871*。Koloskova, A., Hendrikx, H., & Stich, S. U.（2023）。重新审视梯度裁剪：随机偏差与紧致的收敛保证。载于 国际机器学习会议 （第 17343–17363 页）。Laurent, B. & Massart, P.（2000）。通过模型选择自适应估计二次泛函。统计年刊，28(5)，1302–1338。Li, H., Rakhlin, A., & Jadbabaie, A.（2023）。宽松假设下 Adam 的收敛性。神经信息处理系统进展，（第 52166–52196 页）。Li, J., Chen, X., Ma, S., & Hong, M.（2024）。无问题参数的去中心化非凸随机优化。*arXiv* 预印本 *arXiv:2402.08821*。Lobanov, A. & Gasnikov, A.（2025）。随机凸优化中广义光滑性的优势：一阶与零阶算法。*arXiv* 预印本 *arXiv:2501.18198*。Lobanov, A., Gasnikov, A., Gorbunov, E., & Tak´aˇc, M.（2024）。凸情形下可以实现线性收敛率！基于（L_0, L_1）-光滑性的梯度下降方法变体。*arXiv* 预印本 *arXiv:2412.17050*。Lyu, Y. & Tsang, I. W.（2019）。具有隐式自然梯度的黑盒优化器。*arXiv* 预印本 *arXiv:1910.04301*。Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., & Arora, S.（2023）。仅通过前向传播微调语言模型。神经信息处理系统进展，36，53038–53075。

Mania, H., Guy, A., & Recht, B. (2018). 静态线性策略的简单随机搜索在强化学习中具有竞争力。发表于神经信息处理系统国际会议第 *32* 届会议论文集（第 1805–1814 页）。Nesterov, Y. & Spokoiny, V. (2017). 凸函数的随机无梯度最小化。计算数学基础，17(2), 527–566。Qiu, X., Gan, Y., Hayes, C. F., Liang, Q., Meyerson, E., Hodjat, B., & Miikkulainen, R. (2025). 大规模进化策略：超越强化学习的大语言模型微调。*arXiv* 预印本 *arXiv:2509.24372*。Reisizadeh, A., Li, H., Das, S., & Jadbabaie, A. (2025). 非凸优化的方差缩减裁剪。发表于国际声学、语音与信号处理会议（第 1–5 页）。Salimans, T., Ho, J., Chen, X., Sidor, S., & Sutskever, I. (2017). 作为强化学习可扩展替代方案的进化策略。*arXiv* 预印本 *arXiv:1703.03864*。Tovmasyan, Z., Malinovsky, G., Condat, L., & Richt´arik, P. (2025). 重新审视随机近端点方法：广义光滑性与相似性。*arXiv* 预印本 *arXiv:2502.03401*。Tyurin, A. (2025). 广义光滑性下梯度下降的统一理论。发表于国际学习表征会议。Vankov, D., Rodomanov, A., Nedich, A., Sankar, L., & Stich, S. U. (2025). 用梯度方法优化（L_0, L_1）-光滑函数。发表于国际学习表征会议。Woodruff, D. P. 等人 (2014). 草图化作为数值线性代数工具。理论计算机科学基础与趋势® 理论计算机科学，10(1–2), 1–157。Xie, C., Li, C., Zhang, C., Deng, Q., Ge, D., & Ye, Y. (2024). 超越 Lipschitz 光滑性的非凸随机优化信赖域方法。发表于*AAAI* 人工智能会议 （第 16049–16057 页）。Ye, H., Chang, X., & Chen, X. (2025). 基于非自适应随机草图化的统一零阶优化框架。*arXiv* 预印本 *arXiv:2510.10945*。Yu, D., Jiang, W., Wan, Y., & Zhang, L. (2025). 广义光滑性下的镜像下降。*arXiv* 预印本 *arXiv:2502.00753*。Zhang, B., Jin, J., Fang, C., & Wang, L. (2020a). 非凸优化裁剪算法的改进分析。发表于神经信息处理系统进展（第 15511–15521 页）。Zhang, J., He, T., Sra, S., & Jadbabaie, A. (2020b). 为什么梯度裁剪能加速训练：自适应性的理论依据。国际学习表征会议。

一个有用的引理

引理 25 (Azuma-Hoeffding 不等式)。设 $E_1, \ldots, E_T$ 为鞅差序列，且对所有 i 均有一致界 $|E_i| \leq B$。那么，

$$\Pr\left(\frac{1}{s}\sum_{t=1}^{s} E_t > \zeta\right) \leq \exp\left(-\frac{\zeta^2}{2B}\cdot s\right). \tag{A.1}$$

引理 26. 若 $\boldsymbol{S}$ 是高斯草图矩阵，即其各个元素独立同分布地取自 $\mathcal{N}(0, 1/\ell)$，且 $\ell \geq 4$，则以至少 $1-\delta'$ 的概率，有

$$\delta' = \exp\left(-\left(\frac{\ell}{8} - \sqrt{\frac{\ell-2}{8}}\right)\right),$$

满足 $\frac{1}{\ell}\left(\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x})\ \boldsymbol{s}_i\rangle\right)^2 \leq \frac{1}{4}\left\|\nabla f(\boldsymbol{x})\right\|^2$。

证明。首先，若 S 是高斯草图矩阵，即其各个元素独立同分布地取自 $\mathcal{N}(0, 1/\ell)$。根据高斯分布的性质，可得 $\langle \nabla f(\boldsymbol{x})\ \boldsymbol{s}_i\rangle \sim \mathcal{N}(0, \|\nabla f(\boldsymbol{x})\|^2/\ell)$。由于 $\boldsymbol{s}_i$相互独立，可得

$$\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle \sim \mathcal{N}(0, \|\nabla f(\boldsymbol{x})\|^2).$$

因此，$(\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x})\ \boldsymbol{s}_i\rangle)^2$ 服从一个经 $\|\nabla f(\boldsymbol{x})\|^2$缩放的 $\chi^2(1)$分布。

那么，根据引理 27，$\frac{1}{\ell}(\sum_{i=1}^{\ell}\langle \nabla f(\boldsymbol{x})\ \boldsymbol{s}_i\rangle)^2$大于 $1/4\left\|\nabla f(\boldsymbol{x})\right\|^2$的概率不超过

$$\mathbb{P}\left[q_1 \geq 1 + 2\sqrt{\tau} + 2\tau = \frac{\ell}{4}\right] \leq \exp(-\tau),$$

其中 $\sqrt{\tau} = \sqrt{\frac{\ell-2}{8}} - \frac{1}{2}$，这意味着 $\tau = \frac{\ell}{8} - \sqrt{\frac{\ell-2}{8}}$ □

引理 27 (χ^2 尾部界 Laurent & Massart (2000))。设 $q_1, \ldots, q_n$ 为相互独立的 χ^2 随机变量，每个变量的自由度为 *1*。对于任意向量 $\gamma = (\gamma_1, \ldots, \gamma_n) \in \mathbb{R}_+^n$，其分量均非负，以及任意 $\tau > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n}\gamma_i q_i \geq \|\gamma\|_1 + 2\sqrt{\|\gamma\|_2^2\,\tau} + 2\|\gamma\|_\infty \tau\right] \leq \exp(-\tau),$$

其中 $\|\gamma\|_1 = \sum_{i=1}^{n}|\gamma_i|$。

引理 28 (引理 第 11 页，来自 Ye等人 (2025))。如果 $\boldsymbol{S} \in \mathbb{R}^{d\times\ell}$是一个无偏的$(14, k, \delta)$-随机草图矩阵 g

，那么以至少 $1-\delta$ 的概率，有

$$\|\boldsymbol{S}\|_F^2 \leq \frac{5\ell}{4} + \frac{d\ell}{4k}. \tag{A.2}$$

引理 29. 设 $\boldsymbol{r} \in \mathbb{R}^d$ 为一个均值为0的向量，其分量为独立同分布的次高斯随机变量，且次高斯参数为常数 C。设 $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ 为一个矩阵。那么，存在一个仅依赖于 C 的常数 c，使得对任意 $\tau \geq 0$,

$$\mathbb{P}\left\{\left|\boldsymbol{r}^\top \boldsymbol{A}\boldsymbol{r} - \mathbb{E}\left[\boldsymbol{r}^\top \boldsymbol{A}\boldsymbol{r}\right]\right| > \tau\right\} \leq 2\exp\left(-c\cdot\min\left\{\frac{\tau^2}{\|\boldsymbol{A}\|_F^2}, \frac{\tau}{\|\boldsymbol{A}\|_2}\right\}\right) \tag{A.3}$$

引理 30. 如果 $\boldsymbol{S} \in \mathbb{R}^{d\times\ell}$ 是高斯、拉德马赫、子采样随机哈达玛变换或稀疏嵌入随机草图矩阵（参见第 2.3 节），那么以至少 $1-\delta'$ 的概率 $\delta' = \exp(-\mathcal{O}(\ell))$，有 $\frac{1}{\ell}\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}) \boldsymbol{s}_i\rangle\right)^2 \leq \frac{1}{4}\|\nabla f(\boldsymbol{x})\|^2$。

证明。 首先，我们

$$\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle\right)^2 = \nabla^\top f(\boldsymbol{x})\boldsymbol{S}\mathbf{1}_\ell\mathbf{1}_\ell^\top\boldsymbol{S}^\top\nabla f(\boldsymbol{x}) = \left(\boldsymbol{S}\mathbf{1}_\ell\right)^\top\nabla f(\boldsymbol{x})\nabla^\top f(\boldsymbol{x})\boldsymbol{S}\mathbf{1}_\ell.$$

我们对草图矩阵 $\boldsymbol{S}$ 的选择保证了 $\boldsymbol{S}\mathbf{1}_\ell$ 是一个均值为 0、分量独立同分布的次高斯随机变量向量。

此外，有

$$\mathbb{E}\left[\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle\right)^2\right] = \sum_{i=1}^{\ell}\mathbb{E}\left[\langle\nabla f(\boldsymbol{x}), \boldsymbol{s}_i\rangle^2\right] = \sum_{i=1}^{\ell}\mathbb{E}\left[\nabla^\top f(\boldsymbol{x})\boldsymbol{s}_i\boldsymbol{s}_i^\top\nabla f(\boldsymbol{x})\right] = \|\nabla f(\boldsymbol{x})\|^2,$$

其中，第一个等号是因为各个 $\boldsymbol{s}_i$彼此独立。

根据引理 29，结合 $\boldsymbol{r} = \boldsymbol{S}\mathbf{1}_\ell$，$\boldsymbol{A} = \nabla f(\boldsymbol{x})\nabla^\top f(\boldsymbol{x})$，以及 $\tau = \left(\frac{\ell}{4}-1\right)\|\nabla f(\boldsymbol{x})\|^2$，要得到 $\frac{1}{\ell}\left(\sum_{i=1}^{\ell}\langle\nabla f(\boldsymbol{x})\ \boldsymbol{s}_i\rangle\right)^2 \leq \frac{1}{4}\|\nabla f(\boldsymbol{x})\|^2$，其成立的概率至少为 $1-\delta'$，其中 $\delta' = \exp(-\mathcal{O}(\ell))$。

□

引理 31 (Cortinovis 与 Kressner 的推论 2(2021)) 。设 $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ 为正定矩阵，$\delta \in (0,\ 1/2]$，$\varepsilon \in (0,\ 1)$，$\ell \in \mathbb{N}$。设 $\tau_\ell(\boldsymbol{A}) = \sum_{i=1}^{\ell}\boldsymbol{s}_i^\top\boldsymbol{A}\boldsymbol{s}_i$。要使以下性质以至少 $1-\delta$ 的概率成立，

$$\left|\tau_\ell(\boldsymbol{A}) - \mathrm{tr}(\boldsymbol{A})\right| \leq \varepsilon\cdot\mathrm{tr}(\boldsymbol{A}),$$

那么，

1. 如果 $\boldsymbol{s}_i$ 是随机高斯草图矩阵的第 i 列，那么 $\ell = 8\varepsilon^{-2}\|A\|_2\mathrm{tr}(A)\log\frac{2}{\delta}$ 就足够了。

2. 如果 $s_i \in \mathbb{R}^d$ 是随机 *Rademacher* 矩阵的第 i 列，那么 $\ell = 8\varepsilon^{-2}(1+\varepsilon)\|A\|_2 \mathrm{tr}(A) \log \frac{2}{\delta}$ 就足够了。

3. 如果 $s_i \in \mathbb{R}^d$ 是子采样随机哈达玛变换的一列，则 $\ell = \left(1+\sqrt{8\ln \frac{2n}{\delta}}\right)^2 \varepsilon^{-2} \log \frac{2}{\delta}$ 就足够了。

引理 32. 设 $S_t \in \mathbb{R}^{d\times\ell}$ 为高斯或拉德马赫草图矩阵，满足 $\ell = 16k\log \frac{4T}{\delta}$ 和 $0 < \delta < 1$。那么，对于所有作为 S_t的一列的 s_i，以至少 $1-\delta$的概率，有

$$\|s_i\|_{\boldsymbol{H}} \leq 2\sqrt{\mathrm{tr}(\boldsymbol{H})} \tag{A.4}$$

证明。根据草图矩阵的性质以及 ℓ的值，我们可以得到

$$\|s_i\|_{\boldsymbol{H}} = \sqrt{\|s_i\|_{\boldsymbol{H}}^2} \leq \sqrt{\sum_{i=1}^{\ell} \|s_i\|_{\boldsymbol{H}}^2} = \sqrt{\sum_{i=1}^{\ell} s_i^\top \boldsymbol{H} s_i} \leq 2\sqrt{\mathrm{tr}(\boldsymbol{H})},$$

其中，最后一个不等式是因为引理 31以及 ℓ的值。 □

B 假设1与相关工作的联系

回顾原始的松弛光滑性条件（Zhang 等人，2020b）指出可微函数 $f\colon \mathbb{R}^d \to \mathbb{R}$ 成立

$$\left\|\nabla^2 f(\boldsymbol{x})\right\| \leq L_0 + L_1 \|\nabla f(\boldsymbol{x})\| \tag{B.1}$$

对所有 $\boldsymbol{x} \in \mathbb{R}^d$。后来 Zhang 等人（2020a，推论 A.3 和 A.4）证明了（B.1）意味着它成立

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq \left(AL_0 + BL_1 \|\nabla f(\boldsymbol{x})\|\right) \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{B.2}$$

and

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{AL_0 + BL_1 \|\nabla f(\boldsymbol{x})\|}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 \tag{B.3}$$

对于所有 $\|\boldsymbol{x}-\boldsymbol{y}\| \leq c/L_1$，其中 A， B， c 是一些常数。事实上，我们可以将（B.3）加强如下

$$
\begin{aligned}
&|f(\boldsymbol{y})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle| \\
=&\left|\int_0^1\langle\nabla f(\theta\boldsymbol{y}+(1-\theta)\boldsymbol{x})-\nabla f(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{y}\rangle \,\mathrm{d}\theta\right| \\
\leq&\int_0^1\|\nabla f(\theta\boldsymbol{y}+(1-\theta)\boldsymbol{x})-\nabla f(\boldsymbol{x})\|\,\|\boldsymbol{x}-\boldsymbol{y}\| \,\mathrm{d}\theta \\
\leq&\int_0^1(AL_0+BL_1\|\nabla f(\boldsymbol{x})\|)\,\|\theta\boldsymbol{y}-\theta\boldsymbol{x}\|\,\|\boldsymbol{x}-\boldsymbol{y}\| \,\mathrm{d}\theta \\
\leq&\int_0^1\theta(AL_0+BL_1\|\nabla f(\boldsymbol{x})\|)\,\|\boldsymbol{x}-\boldsymbol{y}\|^2 \,\mathrm{d}\theta \\
\leq&\frac{AL_0+BL_1\|\nabla f(\boldsymbol{x})\|}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 .
\end{aligned}
$$

自然地，我们可以将上述公式中的最后一项推广为 $\boldsymbol{H}$-范数，从而得到以下不等式

$$
|f(\boldsymbol{y})-f(\boldsymbol{x})-\langle\nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle| \leq \frac{AL_0+BL_1\|\nabla f(\boldsymbol{x})\|}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_{\boldsymbol{H}}^2 ,
$$

这就是我们的假设1 （忽略常数上的差异）。