arXiv:2303.00883v2 [cs.LG] 2 Jun 2023

非凸优化的方差缩减裁剪

Amirhossein Reisizadeh*[1]、Haochuan Li*[1]、Subhro Das[2], 和 Ali Jadbabaie[1]

[1]麻省理工学院[2]MIT–IBM Watson人工智能实验室、IBM研究院

摘要

梯度裁剪是深度学习应用中的一种标准训练技术，例如大型语言模型训练中用于缓解梯度爆炸。近期实验研究表明，采用梯度裁剪进行训练时，训练目标函数沿其轨迹的光滑性呈现出一种相当特殊的行为：光滑性会随梯度范数增长 L 。这与非凸优化中广为人知的传统假设形成鲜明对比，该假设也称为 L–光滑性，即假定光滑性全局有界。近期提出的(L_0, L_1)–光滑性是一种更宽松的概念，能够刻画非凸优化中的此类行为。具体而言，研究已表明，在这一宽松的光滑性假设下，采用裁剪的随机梯度下降需要进行 $\mathcal{O}(\epsilon^{-4})$次随机梯度计算，才能找到一个 ϵ–驻点解。本文采用一种方差缩减技术，即 Spider，并证明，通过精心设计学习率，可将这一复杂度降低至 $\mathcal{O}(\epsilon^{-3})$，达到阶最优。我们设计的学习率结合了裁剪技术，以缓解光滑性的增长。此外，当目标函数是 n 个分量的平均值时，我们将现有的随机梯度复杂度 $\mathcal{O}(n\epsilon^{-2})$界改进至 $\mathcal{O}(\sqrt{n}\epsilon^{-2}+n)$，同样达到阶最优。除了理论上的最优性之外，在若干视觉任务中，采用我们设计参数的 Spider 与随机方差缩减梯度和 SARAH 等方差缩减方法相比，也展现出相当的实证性能。

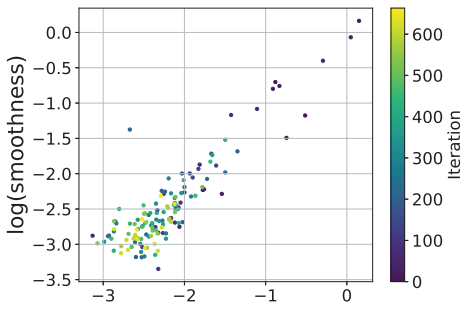
1 引言

我们研究最小化一个非凸 函数 $F\colon \mathbb{R}^d \to \mathbb{R}$ 的问题，该函数表示为一个随机函数的期望，即，

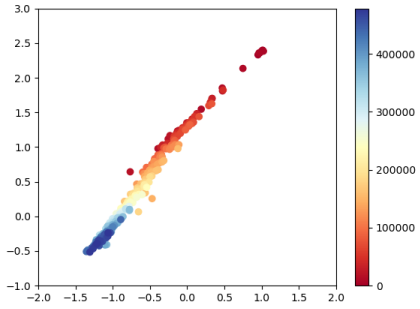
$$\underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize}} \;\; F(\mathbf{x}) = \mathbb{E}_{\xi}[f(\mathbf{x};\xi)], \tag{1}$$

其中，随机变量 ξ 根据分布 $\mathcal{D}$ 实现。在这种随机设置下，分布 $\mathcal{D}$ 通常是未知的，但可以获得若干实现的样本。在这种设置下，

*同等贡献。[1]
{amirr,haochuan,jadbabai}@mit.edu[2]
subhro.das@ibm.com



(a) 使用梯度裁剪训练 AWD-LSTM



(b) 使用 Adam 训练 Transformer

图 1：沿训练轨迹，光滑性随梯度范数增大而提升：(a) 使用裁剪后的 随机梯度下降 训练的 PTB 数据集上的 AWD-LSTM（图取自 Zhang 等人 (2019年)），以及 (b) 使用 Adam 训练的 WMT 2014 翻译数据集上的 Transformer（图取自 Wang 等人 (2022年)）。

称为有限和，目标函数 F 可以表示为 n 个分量函数 $f_1, \cdots, f_n$ 的平均值，即，

$$\underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize}} \; F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}). \tag{2}$$

这一表述概括了许多机器学习和深度学习应用中的标准训练框架：通过最小化大量带标签数据样本所产生的平均损失来训练模型参数（也称为经验风险最小化）。

基于梯度的算法（如随机梯度下降）因其简单易用，已广泛应用于深度学习模型的训练。然而，在自然语言处理（NLP）等特定应用中，自适应梯度方法的表现优于随机梯度下降。在这类方法中，梯度裁剪已成为训练语言模型时的一种标准做法，可用于缓解梯度爆炸问题。尽管梯度裁剪的这种优越表现尚未得到充分的理论论证，但 Zhang 等人，2019 提出了一些理由，有助于更好地理解这一现象可能的内在机制。为弥合这一特定应用中的理论与实践差距，Zhang 等人，2019 首先展示了大规模语言模型（例如使用梯度裁剪训练的长短期记忆网络）的优化景观具有一项有趣特征。如图 1 所示，沿训练轨迹，目标函数的光滑性随梯度范数增大而提升。这与非凸优化中的一个公认观点相悖：该观点假设光滑性在整个输入空间内受某个常数所界定，即 $\|\nabla^2 F(\mathbf{x})\| \leq L$。受实验结果启发，Zhang 等人，2019 引入了一种更宽松的光滑性概念，称为 (L_0, L_1)-光滑性，其中光滑性随梯度范数线性增长，即 $\|\nabla^2 F(\mathbf{x})\| \leq L_0 + L_1\|\nabla F(\mathbf{x})\|$，其中 L_0、L_1 为正常数。这类非凸函数包括许多不具有全局利普希茨连续梯度的实例，例如所谓的指数族。具体而言，所有次数至少为 3 的多项式都是 (L_0, L_1)–光滑的，但不存在能够全局界定其光滑性的常数（Zhang 等人，2019）。

非凸优化算法的标准性能指标是其为找到近似一阶平稳解所需的梯度计算次数。更准确地说，任何非凸优化算法的目标都是找到一个解$\mathbf{x}$ ，使得

$$\|\nabla F(\mathbf{x})\| \leq \epsilon,$$

对于给定的目标精度 ϵ。找到这样的驻点所需的梯度计算总次数，定义为相应算法的一阶预言机 或梯度复杂度 。采用梯度

裁剪，Zhang 等人(2019年)表明，经过裁剪的随机梯度下降能够找到一个 ϵ–驻点解，适用于任意$(L_0,\ L_1)$–光滑非凸目标函数，且梯度复杂度至多为 $\mathcal{O}(\epsilon^{-4})$。本文旨在回答以下问题：

> 我们能否 改进在 (ϵ^{-4}) 光滑非凸优化中的梯度复杂度？

我们对此问题给予肯定的回答。具体而言，我们采用方差缩减 技术，并证明在常规条件下，寻找一个 ϵ–驻点所需的梯度复杂度可以改进至$\mathcal{O}(\epsilon^{-3})$，该复杂度达到了阶最优。

方差缩减一直是加速L–光滑目标函数的非凸优化算法的一种有前景的方法。与我们工作最相关的是 Fang 等人，2018 提出的 Spider 算法。Spider 的核心思想是设计并沿迭代点维护一个真实梯度的精确估计量。令 $\mathbf{v}_k$ 表示在迭代点 k处对真实梯度 $\nabla F(\mathbf{x}_k)$ 的 Spider 估计量。梯度估计器 $\mathbf{v}_k$在每次迭代中使用一小批随机梯度和前一个 $\mathbf{v}_{k-1}$进行更新。为了消除随机梯度噪声的不利影响，每隔一段时间使用一个大批次更新 $\mathbf{v}_k$ 。已有研究表明，对于适当选择的步长，Spider 能够在每次迭代中将梯度估计器的方差控制在 ϵ^2 以内。结合标准的下降引理，Fang 等人，2018 已证明，对于任意（平均） L–光滑目标函数，Spider 至多需要 $\mathcal{O}(\epsilon^{-3})$ 次随机梯度计算即可达到一个 ϵ– 驻点。与这一阶最优速率对应的自适应步长选取为 $\mathcal{O}(\min\{1, \epsilon/\|\mathbf{v}_k\|\})$。

在本文中，我们旨在将 Spider 中的方差缩减思想用于加速针对 (L_0, L_1)–光滑目标函数的裁剪随机梯度下降算法。这样做面临的首要挑战是，对于L–光滑目标函数，标准的下降引理在较宽松的 (L_0, L_1)–光滑条件下并不 成立。我们证明，步长中的裁剪项，即 $\epsilon/\|\mathbf{v}_k\|$，使我们能够在新的光滑性条件下确立下降性质。

在我们的设定中采用 Spider 方法所面临的第二个、也是更关键的挑战，是控制梯度估计器的方差，即 $\mathbb{E}\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\|^2$。如前所述，Fang 等人（2018）证明，对于 L–光滑目标函数，若步长选为 $\mathcal{O}(\min\{1, \epsilon/\|\mathbf{v}_k\|\})$，则沿迭代点的梯度估计器方差始终有界，由 ϵ^2 界定。然而，在我们的 (L_0, L_1)–光滑设定中，这一步长无法保证方差有界。具体而言，我们证明，要控制方差，需要采用更小的步长。也就是说，当步长为 $\mathcal{O}(\min\{1, \epsilon/\|\mathbf{v}_k\|, \epsilon/\|\mathbf{v}_k\|^2\})$ 时，可以证明梯度估计器 $\mathbf{v}_k$ 的方差受到控制，且由 ϵ^2 界定。步长中的附加项 $\epsilon/\|\mathbf{v}_k\|^2$ 对于缓解宽松的 (L_0, L_1)–光滑设定下随梯度范数增长的光滑性至关重要。我们将采用这一特定步长、处于此光滑性设定下的 Spider 称为 (L_0, L_1)–Spider。

结合（新的）下降引理，我们证明，采用上述设计的学习率，(L_0, L_1)–Spider 能以高概率找到任意非凸且(L_0, L_1)–光滑函数的一个 ϵ–驻点。更重要的是，我们证明，找到这样的驻点所需的随机梯度计算总次数至多为 $\mathcal{O}(\epsilon^{-3})$。在分析中，我们放宽了（Zhang 等人，2019）中更严格的随机梯度假设——噪声几乎必然有界，并改为采用常见且相当通用的有界噪声方差假设。此外，我们假设目标函数在其随机分量上满足平均(L_0, L_1)–光滑性。在较弱且典型的假设基础上，再施加更强的平均光滑性假设，确实是方差缩减优化文献中的标准做法。在此类通用假设下，已有研究表明，对于（平均） L–光滑目标函数，$\Omega(\epsilon^{-3})$ 速率确实是随机梯度复杂度的下界（Fang 等人，2018）。显然，每个 L–光滑函数也都是$(L, 0)$–光滑的。因此，我们的 $\mathcal{O}(\epsilon^{-3})$梯度复杂度界对于更广泛的$(L_0, L_1)$–光滑非凸函数类也同样是紧的。

我们将结果推广到有限和设定（2），并证明，对于我们设计的步长取值

算法	参考文献	随机	有限和
ClippedSGD	Zhang 等人 (2019年)	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(n\epsilon^{-2})$
$(L_0,\ L_1)$–Spider	**本文**	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\sqrt{n}\epsilon^{-2}+n)$
下界	Arjevani 等人 (2022年); Fang 等人(2018年)	$\Omega(\epsilon^{-3})$	$\Omega(\sqrt{n}\epsilon^{-2})^{\dagger}$

表 1：$(L_0,\ L_1)$–光滑非凸优化的复杂度。 †该下界适用于 $n \leq \mathcal{O}(\epsilon^{-4})$。

光滑性	参考文献	随机	有限和	学习率
L–光滑	Fang 等人(2018年)	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\sqrt{n}\epsilon^{-2}+n)$	$\mathcal{O}\left(\text{最小化}\left\{1, \frac{\epsilon}{\|\mathbf{v}_k\|}\right\}\right)$
$(L_0,\ L_1)$–光滑	**本文**	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\sqrt{n}\epsilon^{-2}+n)$	$\mathcal{O}\left(\text{最小值}\left\{1, \frac{\epsilon}{\|\mathbf{v}_k\|}, \frac{\epsilon}{\|\mathbf{v}_k\|^2}\right\}\right)$

表2：Spider 和 (L_0, L_1)–Spider 的学习率。所有梯度复杂度均为阶最优。

$\mathcal{O}(\min\{1, \epsilon/\|\mathbf{v}_k\|, \epsilon/\|\mathbf{v}_k\|^2\})$，Spider 中的方差缩减方法能够以高概率找到任意非凸（L_0, L_1）光滑函数的一个 ϵ–驻点，且至多需要 $\mathcal{O}(\sqrt{n}\epsilon^{-2}+n)$ 次梯度计算。这显著降低了裁剪随机梯度下降已有的梯度复杂度（Zhang 等人，2019），即 $\mathcal{O}(n\epsilon^{-2})$。此外，推导出的 $\mathcal{O}(\sqrt{n}\epsilon^{-2}+n)$ 复杂度界在（L_0, L_1）光滑设定下自然是紧的，因为在限制更严格的 L–光滑性条件下也是如此。表1和表2总结了我们的讨论。

贡献。综上所述，本文研究了在 (L_0, L_1)– 光滑性条件下的非凸优化，并改进了随机和有限和设定下达到一阶驻点解的现有梯度复杂度。本文采用方差缩减技术 SPIDER，并设计了可实现阶最优梯度复杂度的学习率。表 1 比较了本文的结果与现有及最优梯度复杂度。表 2比较了传统的 L–光滑条件和新的 (L_0, L_1)–光滑性条件下实现阶最优复杂度的学习率。此外，我们使用设计的学习率实现了 SPIDER，并在采用 MNIST、CIFAR10 和 CIFAR100 数据集的不同图像分类任务上，将其实际性能与若干基准方法进行比较，例如随机梯度下降、SARAH 和随机方差缩减梯度。

记号。本文中，我们用 $\|\mathbf{a}\|$ 表示向量 a 的 ℓ_2–范数。我们还用 $\|\mathbf{A}\|$ 表示矩阵 A 的谱范数。对于定义在同一定义域上的非负函数 f, g：$\mathcal{X} \to [0, \infty)$，标准的大 O 记号 $f = \mathcal{O}(g)$ 概括了这样一个事实：存在正常数 $c > 0$，使得对于所有 $\mathbf{x} \in \mathcal{X}$，$f(\mathbf{x}) \leq c \cdot g(\mathbf{x})$。此外，如果存在正常数 $c > 0$，使得对于所有 $\mathbf{x} \in \mathcal{X}$，$f(\mathbf{x}) \geq c \cdot g(\mathbf{x})$，我们就记作 $f = \Omega(g)$。最后，我们使用简写记号 $\mathbf{x}_{i:j}$ 表示序列 $\mathbf{x}_i, \cdots, \mathbf{x}_j$。

相关工作。

(L_0, L_1)–光滑性与梯度裁剪。如前所述，梯度裁剪已广泛用于训练大规模语言模型等深度学习模型，以规避梯度爆炸问题（Merity 等人，2017年；Gehring 等人，2017年；Peters 等人，2018年）。Zhang 等人（2019）提出了一个理论框架，以更好地理解裁剪算法相较于传统非自适应梯度方法的优越性能。随后，若干研究对 Zhang 等人（2019）引入的 (L_0, L_1)–光滑性概念进行了研究。Zhang 等人（2020）采用动量技术，并改进了 Zhang 等人（2019）此前推导的裁剪随机梯度下降的收敛速率常数依赖性。在这一宽松的光滑性假设下，Qian 等人（2021）研究了裁剪在增量梯度方法中的作用。在采用全批量梯度计算的确定性设置下，裁剪梯度下降和

归一化梯度下降（NGD）在本质上仅相差一个常数。Zhao 等人（2021）为 (L_0, L_1)–光滑非凸函数下的随机归一化梯度下降提供了收敛保证。Faw 等人（2023）基于 AdaGrad 类方法，放宽了现有关于随机梯度噪声的严格假设。裁剪方法也可以实现可扩展性（Liu 等人，2022）并满足隐私考量（Yang 等人，2022；Xia 等人，2022）。在相同设置下，研究表明，非裁剪方法，尤其是广义 SignSGD 算法，可以达到与裁剪随机梯度下降相同的速率（Crawshaw 等人，2022）。有趣的是，这种光滑性行为并不限于 Zhang 等人（2019）所实现的裁剪式随机梯度下降优化器；使用 Adam 训练语言模型同样表现出这类光滑性现象（Wang 等人，2022）。除了优化领域之外，(L_0, L_1)–光滑性也被用于研究变分推断问题（Sun 等人，2022）。

非凸优化中的方差缩减。众所周知，方差缩减是加速凸和非凸优化算法的有效途径。为回顾方差缩减技术在寻找具有全局 Lipschitz 梯度（即 L–光滑）的非凸函数驻点时所带来的收敛速率改进，我们回想一下随机梯度下降/梯度下降在有限和与随机设定下广为人知的速率 $\mathcal{O}(\min\{n\epsilon^{-2}, \epsilon^{-4}\})$（Nesterov，2003）。随机方差缩减梯度（SVRG）和随机控制随机梯度（SCSG）将梯度复杂度改进至 $\tilde{\mathcal{O}}(\min\{n^{2/3}\epsilon^{-2}, \epsilon^{-10/3}\})$（Allen-Zhu 和 Hazan，2016；Reddi 等，2016；Lei 等人，2017）。为进一步降低梯度复杂度，Fang 等人（2018）提出了一种更准确且成本更低的方法，用于跟踪各迭代点处的真实梯度，即随机路径积分微分估计器（Spider）。这种方差缩减梯度方法的成本至多为 $\mathcal{O}(\min\{\sqrt{n}\epsilon^{-2}, \epsilon^{-3}\})$，在有限和设定（Fang 等人，2018）和随机设定（Arjevani 等，2022）下均达到下界复杂度。因此，Spider 是寻找非凸光滑函数驻点的阶最优算法。与此同时，Nguyen 等人（2017）提出了 SARAH，它与 Spider 采用相同的递归随机梯度更新框架。此外，Zhou 等人（2020）提出了 SNVRG，其复杂度界同样紧致。其他具有阶最优收敛速率的方法包括 Wang 等人（2019）、Pham 等人（2020）以及 Li 等人（2021a,b）提出的方法。这些方法还可采用 AdaSpider（Kavis 等人，2022）等自适应学习率。

深度学习中的方差缩减。尽管方差缩减方法在理论上颇具优势，但在加速现代深度神经网络的训练方面，其表现却不尽如人意（Defazio 和 Bottou，2019年；Defazio 等人，2014；Roux 等人，2012；Shalev-Shwartz 和 Zhang，2012）。由于这一理论与实践差距的成因尚未得到解决，人们对方差缩减方法（以及动量方法）在深度神经网络应用中的无效性提出了多种猜测。有观点认为，这类方法（例如随机方差缩减梯度）的非自适应学习率可能导致参数调优难以进行（Cutkosky 和 Orabona，2019年）。此外，正如文献（Zhang 等人，2019年）所暗示的，理论假设与实际情况不一致，也可能显著加剧这一差距。上述大多数方差缩减方法都高度依赖全局 Lipschitz 梯度假设，而研究发现，至少在现代自然语言处理应用中，这一假设并不成立（Zhang 等人，2019年）。

2 预备知识

在本节中，我们首先回顾以往研究中引入的 (L_0, L_1)–光滑函数的基本性质，并给出本文所关注的情形（即随机和有限和）下所采用的假设。

2.1 $(L_0,\ L_1)$–光滑性

如（Zhang 等人，2019年）中所述，函数 F 称为 (L_0, L_1)–光滑函数，若存在常数$L_0 > 0$ 和 $L_1 \geq 0$，使得对所有 $\mathbf{x} \in \mathbb{R}^d$,

$$\|\nabla^2 F(\mathbf{x})\| \leq L_0 + L_1\|\nabla F(\mathbf{x})\|. \tag{3}$$

正如 Zhang 等人，2020年 所指出并在下文阐述的，这一定义中的二阶可微条件可以放宽。

定义1 ((L_0, L_1)–光滑)。 若存在常数 $L_0 > 0$ 和 $L_1 \geq 0$，使得当 $\|\mathbf{x} - \mathbf{y}\| \leq 1/L_1$时，满足以下条件，则称可微函数 F为(L_0, L_1)–光滑：

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq (L_0 + L_1\|\nabla F(\mathbf{x})\|)\|\mathbf{x} - \mathbf{y}\|. \tag{4}$$

下文将证明，这两个条件在常数因子意义下基本等价。因此，接下来我们将定义1 作为 $(L_0,\ L_1)$–光滑性的主要条件。

命题 1。 若 F 二阶可微，则条件 (4) 蕴含 (3)。此外，条件 (3) 蕴含(4)，其中常数为 $(2L_0,\ 2L_1)$。

证明见 F.1 节。定义1给出了主要目标函数 F上的光滑性条件。显然，这种光滑性概念放宽了非凸优化中传统的全局 Lipschitz 梯度假设；该假设要求对所有 x 和 y，$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ 均成立；或者等价地，若函数二阶可微，则对所有 x，$\|\nabla^2 F(\mathbf{x})\| \leq L$ 成立。在本文关注的设定中，目标函数 F 表示为分量函数的（随机或有限和）平均值，如式 (1) 和 (2) 所示。然而，光滑性条件 (4) 仅施加于主要目标函数 F，而与其分量无关。作为方差缩减优化中的标准假设（Fang 等人，2018年），我们对 F 及其分量施加以下平均光滑性条件。

假设 1 (平均 (L_0, L_1)–光滑)。 存在常数 $L_0 > 0$和 $L_1 \geq 0$，使得若 $\|\mathbf{x} - \mathbf{y}\| \leq 1/L_1$，则

(i) 在随机设定下 (1)，

$$\mathbb{E}\Big[\|\nabla f(\mathbf{x};\xi) - \nabla f(\mathbf{y};\xi)\|^2\Big]^{1/2} \leq (L_0 + L_1\|\nabla F(\mathbf{x})\|)\|\mathbf{x} - \mathbf{y}\|,$$

其中期望是对随机 ξ 求取的；或者，

(ii) 在有限和设定下 (2)，

$$\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2\right)^{1/2} \leq (L_0 + L_1\|\nabla F(\mathbf{x})\|)\|\mathbf{x} - \mathbf{y}\|.$$

值得注意的是，在随机设定和有限和设定下，假设1 中的条件均意味着主要目标函数 F 根据定义1具有光滑性。特别是在有限和情形下，由詹森不等式可得

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq \mathbb{E}\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq \mathbb{E}\Big[\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2\Big]^{1/2} \leq (L_0 + L_1\|\nabla F(\mathbf{x})\|)\|\mathbf{x} - \mathbf{y}\|,$$

其中，最后一个等式对满足 $\|\mathbf{x}$，$\mathbf{y}$ 的任意 $\mathbf{x} - \mathbf{y}\| \leq 1/L_1$均成立。在随机梯度无偏的情况下，类似的论证也适用于随机设定。在给出了主要的光滑性假设后，我们将在下一节回顾现有的梯度裁剪方法及其收敛特性。

2.2 梯度裁剪

与传统的（未裁剪）梯度方法类似，通用的梯度裁剪算法通过多次迭代运行，我们将这些迭代记为 $k = 0$、1、$\cdots$，并以 $\mathbf{x}_0$ 初始化。在每个迭代点 k，对随机选取的大小为 $|\mathcal{S}| = S$ 的小批量 $\mathcal{S}$ 计算随机梯度 $\mathbf{g}_k := \nabla f(\mathbf{x}_k; \mathcal{S})$。随后，迭代点 $\mathbf{x}_k$ 按照 $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{g}_k$ 进行更新，其中 η_k 表示学习率（或步长）。例如，对于目标精度 ϵ，自适应步长可表示为 $\eta_k = \mathcal{O}(\min\{1, \epsilon/\|\mathbf{g}_k\|\})$，由常数部分和裁剪部分组成。算法 1 总结了这一过程，我们将其称为 ClippedSGD。下面，我们说明该算法在寻找 (L_0, L_1)–光滑函数的 ϵ–驻点时的梯度复杂度。

算法 1 CLIPPEDSGD

输入：光滑性参数 L_0、L_1，精度 ϵ，批量大小 $S = |\mathcal{S}|$，迭代次数 K 初始化 $\mathbf{x}_0$ 对于 $k = 0$、$\cdots$、$K-1$，执行 抽取样本 $\mathcal{S}$ 并计算 $\mathbf{g}_k = \nabla f(\mathbf{x}_k; \mathcal{S})$ 更新 $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{g}_k$ ▷ $\eta_k = \min\left\{\frac{1}{2L_0}, \frac{1}{L_0}\frac{\epsilon}{\|\mathbf{g}_k\|}\right\}$ 结束循环 从 $\{\mathbf{x}_0$、$\cdots$、$\mathbf{x}_{K-1}\}$ 中随机且均匀地选取并返回 $\tilde{\mathbf{x}}$

让我们从随机设定（1）开始。我们注意到（Zhang 等人，2019年）刻画了 CLIPPEDSGD 在随机梯度噪声几乎处处有界时的梯度复杂度，而我们在分析中放宽了这一条件。下面的假设精确陈述了随机梯度噪声所需满足的条件。

假设 2. 随机梯度 $f(\cdot; \xi)$ 是无偏且方差有界的，也就是说，

$$\mathbb{E}[\nabla f(\mathbf{x}; \xi)] = \nabla F(\mathbf{x}), \quad \text{and} \quad \mathbb{E}\|\nabla f(\mathbf{x}; \xi) - \nabla F(\mathbf{x})\|^2 \leq \sigma^2.$$

上述假设在随机优化中是标准且相当一般的。现在我们可以陈述 ClippedSGD 的迭代复杂度。

定理1 (随机设定). 设假设 *1 (i)* 和 *2* 成立，且 $\epsilon \leq \frac{L_0}{20L_1}$。选取以下步长和参数

$$\eta_k = \min\left\{\frac{1}{2L_0}, \frac{1}{L_0}\frac{\epsilon}{\|\mathbf{g}_k\|}\right\}, \quad S = \frac{\sigma^2}{\epsilon^2}, \quad K = \left\lceil \frac{16\Delta L_0}{\epsilon^2} \right\rceil.$$

那么，对于算法 *CLIPPEDSGD 1* 中的 *1* 的输出，即 $\tilde{\mathbf{x}}$ 从 $\{\mathbf{x}_{0:K-1}\}$ 中随机均匀选取，我们有 $\|\nabla F(\tilde{\mathbf{x}})\| \leq 12\epsilon$，其成立概率至少为 $1/2$。此外，随机梯度复杂度总量有界于 $\Delta L_0 \sigma^2 \mathcal{O}(\epsilon^{-4})$。

证明。证明推迟到第 D 节。 □

这里有几点说明。首先，本文始终用 $\Delta := F(\mathbf{x}_0) - F^*$ 表示初始次优性，其中全局最优值 F^* 假设为有限常数，即 $F^* := \min_{\mathbf{x}} F(\mathbf{x}) > -\infty$。其次，放宽假设 1 （i）为定义 1 中的条件（4），定理 1仍然成立。最后，定理 1 改进了 Zhang 等人(2019年) 的结果（定理7），并将几乎处处有界的梯度噪声条件放宽为有界方差。

在有限和设定（2）下，ClippedSGD 退化为裁剪梯度下降 梯度下降，此时梯度 $\mathbf{g}_k = \nabla f(\mathbf{x}_k; \mathcal{S})$ 通过大小为 $S = |\mathcal{S}| = n$ 的全批次计算得到。（Zhang 等人，2019年）刻画了裁剪梯度下降的迭代复杂度，并证明其有界于 $\mathcal{O}(\Delta L_0 \epsilon^{-2} + \Delta L_1^2 / L_0)$。为使本文内容完整，下面给出这一设定下 ClippedSGD 的收敛速率。

定理 2 (有限和设定)。设假设 *1 (ii)* 和 *2* 成立，且 $\epsilon \leq \frac{L_0}{20 L_1}$。按如下方式选取步长和参数

$$\eta_k = \min\left\{\frac{1}{2L_0}, \frac{1}{L_0}\frac{\epsilon}{\|\mathbf{g}_k\|}\right\}, \quad S = n, \quad K = \left\lceil \frac{16 \Delta L_0}{\epsilon^2} \right\rceil.$$

那么，那么，对于算法 *1* 中的 *ClippedSGD* 的输出，即从 $\tilde{\mathbf{x}}$ 中随机且均匀地选取的$\{\mathbf{x}_{0:K-1}\}$，我们有 $\|\nabla F(\tilde{\mathbf{x}})\| \leq 5\epsilon$ 以至少 $1/2$ 的概率成立。此外，随机梯度复 - 杂度由 $\mathcal{O}(\Delta L_0 n \epsilon^{-2})$ 所界定。

证明。我们将证明推迟至第 E 节。 □

定理 1 和 2 分别给出了寻找非凸且 (L_0, L_1)–光滑函数的 ϵ–驻点时，随机 (1) 和有限和 (2) 优化问题的梯度复杂度上界，分别为 $\mathcal{O}(\epsilon^{-4})$ 和 $\mathcal{O}(n\epsilon^{-2})$。正如第 1 节中简要提及的，针对全局 Lipschitz 梯度目标函数，即 L–光滑函数，复杂度下界已得到刻画。已有研究表明，对于给定的问题参数，存在 L–光滑函数，寻找其 ϵ–驻点至少需要 $\mathcal{O}(\epsilon^{-3})$ 和 $\mathcal{O}(\sqrt{n}\epsilon^{-2} + n)$ 次随机梯度访问（Arjevani 等，2022；Fang 等人，2018）。由于 L–光滑函数是 (L_0, L_1)–光滑函数的子集，因此这类下界对更大的函数类同样成立。因此，本文其余部分的目标是回答以下问题:

对于（ $\mathcal{O}$）和（ $\mathcal{O}$），在（L_0, L_1）–光滑的非凸函数上，下界是否也可达到?

在下一节中，我们将采用方差缩减技术，并正式证明，对于更广泛的非凸函数类别，即也就是（L_0， L_1）–光滑函数，这样的下界速率确实可以达到。

3 主要结果

本节将介绍我们的主要结果，并正式证明，在随机和有限和设定下，求取（L_0， L_1）–光滑函数驻点的梯度复杂度下界均可达到，这些下界已在第2 节中提及。具体而言，我们将说明，最初为 L–光滑目标函数设计的裁剪方法和方差缩减方法，如何应用于我们关注的更一般的函数类别。首先回顾一种最初为具有全局 Lipschitz 梯度的非凸目标函数设计的方差缩减方法。

3.1 Spider

Spider （Fang 等人，2018年）是一种一阶方差缩减算法，能够以近乎最优的梯度复杂度高效地求解光滑非凸目标函数的驻点（见算法 2）。特别是在随机设定下，已有研究表明，当实例函数 $f(\cdot; \xi)$ 具有平均 L–利普希茨连续梯度和 σ–有界方差时，Spider 能够找到一个 ϵ–驻点，其（随机）梯度复杂度为 F （ $\mathcal{O}$） $L\sigma\epsilon^{-3}$。同样，在有限和设定下，Spider 需要 $\mathcal{O}$（$L\sqrt{n}\epsilon^{-2} + n$）梯度

算法 2 (L_0, L_1)–SPIDER

1：输入：光滑性参数 L_0、L_1，精度 ϵ，批次大小 $S_1 = |\mathcal{S}_1|$、$S_2 = |\mathcal{S}_2|$，迭代次数q、K2：初始化 $\mathbf{x}_0$3：对于 $k = 0$、$\cdots$、$K-1$，执行 4：如果 $k \equiv 0$ 对 q 取模，则 5：抽取样本 $\mathcal{S}_1$ 并计算 $\mathbf{v}_k = \nabla f(\mathbf{x}_k;\ \mathcal{S}_1)$ 6：否则 7：抽取样本 $\mathcal{S}_2$ 并计算 $\mathbf{v}_k = \nabla f(\mathbf{x}_k;\ \mathcal{S}_2) - \nabla f(\mathbf{x}_{k-1};\ \mathcal{S}_2) + \mathbf{v}_{k-1}$8：结束 if 9：更新 $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{v}_k$ ▷ $\eta_k = \min\left\{\frac{1}{2L_0}, \frac{1}{L_0}\frac{\epsilon}{\|\mathbf{v}_k\|}, \frac{1}{L_1}\frac{\epsilon}{\|\mathbf{v}_k\|^2}\right\}$10：结束 for

11：从 $\{\mathbf{x}0, \cdots, \mathbf{x}_{K-1}\}$ 中随机且均匀地选取并返回 $\tilde{\mathbf{x}}$

当分量函数 f_i 的平均梯度满足 L–利普希茨连续性时的计算量。下面简要介绍 Spider 的核心思想。

在每次迭代 k 中，Spider（算法 2）都会维护真实梯度 $\nabla F(\mathbf{x}_k)$ 的一个估计值，记为 $\mathbf{v}_k$，并使用大小为 $S_2 = |\mathcal{S}_2|$ 的一小批样本 $\mathcal{S}_2$ 对其进行更新。每经过 q 次迭代，便使用大小为 $S_1 = |\mathcal{S}_1| \geq S_2$ 的大批次 $\mathcal{S}_1$ 刷新 $\mathbf{v}_k$。无论哪种情况，随后都会将迭代点更新为 $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{v}_k$。Spider 的一个重要参数是步长 η_k。最初，对于 L–光滑函数（Fang 等人，2018），学习率选为 $\eta_k = \mathcal{O}(\min\{L^{-1}, \epsilon L^{-1}/\|\mathbf{v}_k\|\})$。我们将 Spider 的核心思想应用于更广泛的光滑性设定，并在下文说明其中的挑战。

挑战。在 Fang 等人，2018 中，作者证明了，对于 L–光滑目标函数，在采用适当的参数S_1、S_2、q、K 以及学习率 $\eta_k = \mathcal{O}(\min\{1, \epsilon/\|\mathbf{v}_k\|\})$时，Spider（算法 2）能够在每次迭代中控制估计量 $\mathbf{v}_k$ 的方差。更准确地说，有 $\mathbb{E}\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\|^2 \leq \epsilon^2$成立。这是 Spider 达到近最优收敛速率的关键。然而，当 L–光滑性假设放宽为（L_0，L_1）–光滑性时，在相同条件下，之前的相同步长将不再适用。具体而言，我们证明，较小的学习率 $\mathcal{O}(\min\{1, \epsilon/\|\mathbf{v}_k\|,\ \epsilon/\|\mathbf{v}_k\|^2\})$可通过 ϵ^2维持估计量的方差。我们将采用这一特定学习率取值的 Spider 方法称为（L_0，L_1）–Spider。进一步细节请参见证明概要和附录。

接下来，我们将给出随机情形和有限和情形下的主要收敛结果，并简要概述证明。

3.2 随机设定

下面的定理给出了在随机设定下寻找 ϵ–驻点解的梯度复杂度上界（1）。

定理 3（随机设定）。考虑(1) 中的随机最小化问题，并假设条件*1* （*i*）和*2* 成立。此外，假设$\epsilon < \frac{L_0}{20L_1}$ 并选取以下步长和参数

$$\eta_k = \min\left\{\frac{1}{2L_0}, \frac{1}{L_0}\frac{\epsilon}{\|\mathbf{v}_k\|}, \frac{1}{L_1}\frac{\epsilon}{\|\mathbf{v}_k\|^2}\right\}, \quad S_1 = \frac{4\sigma^2}{\epsilon^2}, \quad S_2 = 48\frac{L_0}{L_1}\frac{\sigma}{\epsilon}, \quad q = 2\frac{L_0}{L_1}\frac{\sigma}{\epsilon}, \quad K = \left\lceil \frac{16\Delta L_0}{\epsilon^2} \right\rceil.$$

那么，对于 (L_0, L_1)–SPIDER 算法 *2*的输出，即 $\tilde{\mathbf{x}}$ 从 $\{\mathbf{x}_{0:K-1}\}$中随机均匀选取，我们有 $\|\nabla F(\tilde{\mathbf{x}})\| \leq 24\epsilon$ 的概率至少为 $1/2$。此外，随机梯度复杂度为

上界为

$$32\Delta\sigma\left(L_1+24\frac{L_0^2}{L_1}\right)\frac{1}{\epsilon^3}+\frac{4\sigma^2}{\epsilon^2}+2\sigma\left(\frac{L_1}{L_0}+24\frac{L_0}{L_1}\right)\frac{1}{\epsilon}.$$

证明。证明详见第 B 节。 □

定理 3 表明，SPIDER 算法中的方差缩减技术，结合指定的参数和学习率选择，将定理 $\mathcal{O}(\epsilon^{-4})$ 中刻画的 CLIPPEDSGD的梯度复杂度从定理 1 中的 $\mathcal{O}(\epsilon^{-3})$降低至 $1/2$ 。值得注意的是，定理 3 给出的 $1-p$ 概率保证可以提升至任意（常数）概率 $\mathcal{O}(\epsilon^{-3}/\mathrm{poly}(p))$，但这会相应导致更大的随机梯度复杂度。

3.3 有限和设定

接下来，我们考虑有限和设定（2），并给出 SPIDER 寻找驻点的收敛保证。

定理 4（有限和设定）。 考虑 (2) 中的有限和最小化问题，并假设条件 *1* （*ii*）成立。此外，假设 $\epsilon<\frac{L_0}{20L_1}$ 并选取如下的步长和参数

$$\eta_k=\min\left\{\frac{1}{2L_0},\frac{1}{L_0}\frac{\epsilon}{\|\mathbf{v}_k\|},\frac{1}{L_1}\frac{\epsilon}{\|\mathbf{v}_k\|^2}\right\},\quad S_1=n,\quad S_2=12\sqrt{n},\quad q=\sqrt{n},\quad K=\left\lceil\frac{16\Delta L_0}{\epsilon^2}\right\rceil.$$

那么，对于算法 (L_0,L_1)–*SPIDER 2*的输出，即 $\tilde{\mathbf{x}}$ 从 $\{\mathbf{x}_{0:K-1}\}$中随机均匀选取的点，我们有 $\|\nabla F(\tilde{\mathbf{x}})\|\leq 24\epsilon$ ，且此结论以至少 $1/2$的概率成立。此外，找到这样一个驻点的随机梯度复杂度上界为

$$208\Delta L_0\sqrt{n}\frac{1}{\epsilon^2}+n+13\sqrt{n}.$$

证明。证明见第 C 节。 □

考虑主导项，定理 4 将 CLIPPEDSGD 的梯度复杂度从定理 $\mathcal{O}(n\epsilon^{-2})$中的 2 降低至 $\mathcal{O}(\sqrt{n}\epsilon^{-2}+n)$。与定理 3类似， $1/2$ 的概率保证可提高至任意概率 $1-p$，但随机梯度复杂度会增大至 $\mathcal{O}(\sqrt{n}\epsilon^{-2}/\mathrm{poly}(p)+n)$。还值得注意的是，与定理 3不同，定理 4 不要求有界随机梯度噪声，例如假设 2。

证明概要。 为证明定理 3 和 4的收敛速率，我们为随机设定建立了两个论证，概述如下。有限和情形可通过类似论证得出，详细内容见附录。

引理 1（证明概要）。 考虑定理 *3* 中所述的设置和参数，并令 $\epsilon<\frac{L_0}{20L_1}$。那么，对于任意迭代 $k=0,1,\cdots$，有

$$F(\mathbf{x}_{k+1})\leq F(\mathbf{x}_k)-\frac{1}{8}\eta_k\|\mathbf{v}_k\|^2+\frac{5}{8}\eta_k\left\|\mathbf{v}_k-\nabla F(\mathbf{x}_k)\right\|^2,\qquad\text{(descent inequality) (5)}$$

and

$$\mathbb{E}\left[\left\|\mathbf{v}_k-\nabla F(\mathbf{x}_k)\right\|^2\right]\leq\epsilon^2.\qquad\text{(bounded estimator's variance) (6)}$$

第一个论证（5）建立了一个下降不等式，这对于非凸方法的收敛证明通常至关重要。需要注意的是，广为流传的下降引理在 （L_0， L_1）–光滑情形下会失效。此外，（6）保证了当按照定理 3所规定的学习率进行选择时，在所有迭代中估计真实梯度的误差都是有界且较小的。定理 4 中的有限和情形也遵循相同的逻辑。

下界。 为了论证定理 3 和 4中推导出的收敛速率的最优性，我们需要刻画在相同条件下寻找驻点解的梯度复杂度下界。我们先考虑有限和设定以及定理 4。在 L–光滑性设定下，已有结果表明，对于任意 $L > 0$ 和 $n \leq \mathcal{O}(\epsilon^{-4})$，存在一个具有（ 2 ）形式的函数 F，使得

$$\mathbb{E}\Big[\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2\Big]^{1/2} \leq L\|\mathbf{x} - \mathbf{y}\|,$$

对于该函数，寻找一个 ϵ–驻点解至少需要 $\Omega(\sqrt{n}\epsilon^{-2})$ 次随机梯度访问（Fang 等人（2018），定理 3）。显然，根据定义1，任何这样的函数也都是 $(L, 0)$–光滑的，并且在 $L_0 = L$ 和 $L_1 = 0$ 的条件下满足假设 1 (ii)。因此，定理 4 中 Spider 的梯度复杂度对于 $n \leq \mathcal{O}(\epsilon^{-4})$ 是阶最优的，即 $\mathcal{O}(\sqrt{n}\epsilon^{-2} + n) = \mathcal{O}(\sqrt{n}\epsilon^{-2})$，与下界 $\underline{\Omega}(\sqrt{n}\epsilon^{-2})$ 相匹配。

类似地，对于 L-光滑随机设定，（Arjevani 等人，2022年，定理 2）表明，对于任意 $L > 0$和 $\sigma > 0$，存在一个函数 F，其形式为（1），且随机梯度为 g（·； ξ），使得

$$\mathbb{E}[g(\mathbf{x};\xi)] = \nabla F(\mathbf{x}), \qquad \mathbb{E}\Big[\|g(\mathbf{x};\xi) - \nabla F(\mathbf{x})\|^2\Big] \leq \sigma^2, \quad \text{and} \quad \mathbb{E}\Big[\|g(\mathbf{x};\xi) - g(\mathbf{y};\xi)\|^2\Big]^{1/2} \leq L\|\mathbf{x} - \mathbf{y}\|,$$

对于该函数，求得一个 ϵ-驻点解至少需要$\Omega(\sigma\epsilon^{-3} + \sigma^2\epsilon^{-2})$ 次随机梯度查询。因此，存在满足假设（L，0）-光滑条件的函数，满足假设1 （i）和2，其至少需要$\Omega(\sigma\epsilon^{-3} + \sigma^2\epsilon^{-2})$ 次随机梯度访问，才能使SPIDER在定理3 中的收敛速率达到阶最优。

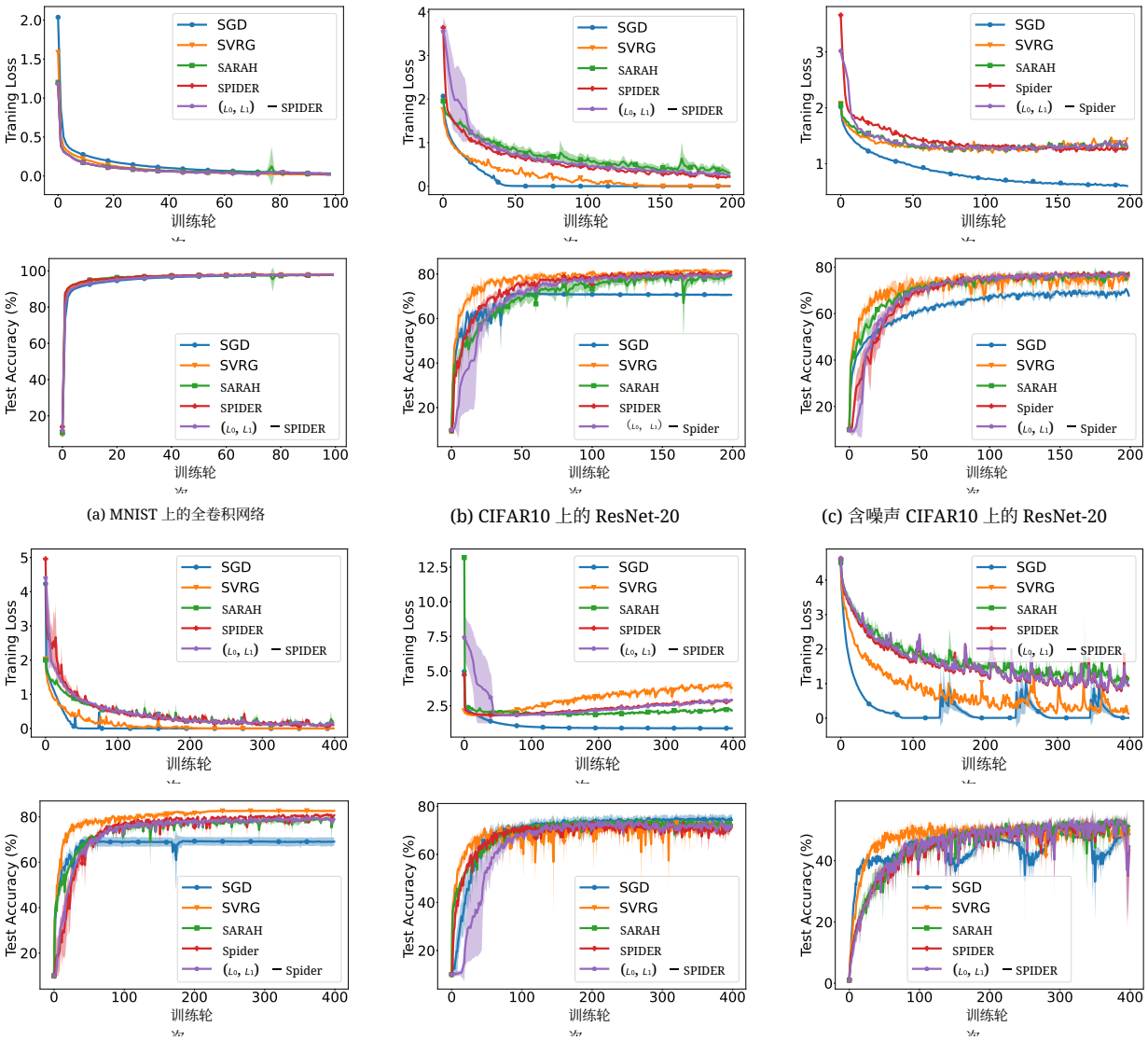
4 实验

本节通过实证展示不同方差缩减方法在神经网络的多种图像分类任务上的收敛表现。我们的实现代码已发布在以下 GitHub 链接[1]。

模型、数据集和基准方法： 我们在三个用于图像分类的标准数据集上训练三种不同的神经网络模型：三层全连接网络（全卷积网络）、ResNet-20 和 ResNet-56（He 等人，2016年）：MNIST（LeCun 等人，1998年）、CIFAR10 和 CIFAR100。在每项实验中，我们将（L_0， L_1）-SPIDER 与相关基准方法进行比较，具体包括随机梯度下降、随机方差缩减梯度（Reddi 等人，2016年）、SARAH（Li 等人，2021b）以及SPIDER （Fang 等人，2018年）。由于我们的主要目标是公平比较不同的方差缩减方法，因此不会采用动量、权重衰减、学习率降低等技巧来追求最先进的准确率。更多实现细节请参见附录。

如图 2(a) 所示，对于在 MNIST 上训练全卷积网络这一简单任务，所有方法都取得了相当高的测试准确率，且表现几乎完全相同。图 2(b) 展示了在干净（无噪声）CIFAR10 上训练的 ResNet-20 的训练和测试准确率曲线。尽管随机梯度下降的测试准确率低于方差缩减方法，但其在训练过程中收敛得快得多。这与 Defazio 和 Bottou（2019年）的结论一致，即方差缩减方法对于深度学习任务而言并不是十分高效。我们强调，我们提出的（L_0, L_1）–Spider 达到了与原始 Spider 相近的性能，并且步长中的附加 $\mathcal{O}(\|\mathbf{v}_k\|^{-2})$ 项，除了在初始步骤中之外，并不会显著减慢训练速度。从定量结果来看，（L_0, L_1）–Spider 的测试准确率达到 79.94%，与之相当

[1] github.com/haochuan-mit/varaince-reduced-clipping-for-non-convex-optimization



(a) MNIST 上的全卷积网络　(b) CIFAR10 上的 ResNet-20　(c) 含噪声 CIFAR10 上的 ResNet-20

(d) CIFAR10 上的 ResNet-56　(e) 含噪声 CIFAR10 上的 ResNet-56　(f) CIFAR100 上的 ResNet-20

图2：神经网络图像分类任务。

的其他方差缩减基准分别达到 81.72%、79.28% 和 81.03% 的测试准确率。我们在附录中提供了更多定量结果。

我们进一步在含噪的 CIFAR10 数据集上训练同一个模型。具体而言，我们向图像添加均值为零、方差为单位的高斯噪声，并以相同概率将标签更改为所有可能类别中的任意一个。0.1 如图2(c)所示，尽管随机方差缩减梯度仍然比其他方差缩减方法略快，但SPIDER和(L_0, L_1)–SPIDER 与图2(b)中的无噪声实验相比，取得了更高的测试准确率，这凸显了SPIDER 对噪声具有潜在的鲁棒性。

接下来，我们在无噪声和有噪声的 CIFAR10 数据集上训练更大的模型，即 ResNet-56。如图 2(d) 和 2(e) 所示，(L_0, L_1)–SPIDER 取得了与其他降方差基准方法相近的性能。它们在收敛方面也表现出与较小模型 ResNet-20 相似的其他特征，详见

图 2(b) 和 2(c)。最后，我们使用 CIFAR100 数据集上的一个更复杂任务，并训练 ResNet-20，结果如图 2(f) 所示。

5 结论

梯度裁剪已广泛用于训练深度神经网络，尤其是语言模型等特定应用。此类模型中梯度裁剪的训练轨迹与传统的L–光滑性假设相矛盾，这要求在非凸优化中放宽这一前提。更宽松的（L_0，L_1）–光滑性概念构建了一个用于研究梯度裁剪方法性能和复杂度的理论框架。在本文中，我们采用一种称为 Spider 的方差缩减技术，在这一更广泛的设定下改进了裁剪方法的梯度复杂度。我们证明，采用精心选择的学习率后，Spider 能够在寻找一阶驻点方面达到阶最优的梯度复杂度速率。然而，如何进一步增强该方法，使其能够逃离鞍点，并针对（L_0，L_1）–光滑非凸目标函数找到二阶 驻点解，仍有待研究。与一阶文献类似，现有的大多数二阶方法高度依赖限制性较强的 Hessian-Lipschitz假设，而这一假设很可能在（L_0，L_1）–光滑设定下失效。我们将这一方向留作未来工作。

致谢

本研究部分得到了 MIT-IBM Watson人工智能实验室和美国海军研究办公室（ONR）资助项目 N00014-20-1-2394 的支持。

参考文献

Zeyuan Allen-Zhu 和 Elad Hazan。通过方差缩减加速非凸优化。载于国际机器学习会议，第 699–707 页。PMLR，2016年。

Yossi Arjevani、Yair Carmon、John C Duchi、Dylan J Foster、Nathan Srebro 和 Blake Woodworth。非凸随机优化的下界。《数学规划》，第 1–50 页，2022年。Michael Crawshaw、Mingrui Liu、Francesco Orabona、Wei Zhang 和 Zhenxun Zhuang。广义 signsgd 对无界光滑性的鲁棒性。*arXiv* 预印本 *arXiv:2208.11195*，2022年。Ashok Cutkosky 和 Francesco Orabona。非凸随机梯度下降中的基于动量的方差缩减。神经信息处理系统进展，第 32 卷，2019年。Aaron Defazio 和 Léon Bottou。方差缩减优化对深度学习无效。神经信息处理系统进展，第 32 卷，2019年。Aaron Defazio、Francis Bach 和 Simon Lacoste-Julien。SAGA：一种支持非强凸复合目标函数的快速增量梯度方法。神经信息处理系统进展，第 27 卷，2014年。Cong Fang、Chris Junchi Li、Zhouchen Lin 和 Tong Z hang。Spider：通过随机路径积分微分估计器实现近乎最优的非凸优化。神经信息处理系统进展，第 31 卷，2018年。

Matthew Faw、Litu Rout、Constantine Caramanis 和 Sanjay Shakkottai。《超越一致光滑性：对自适应随机梯度下降的停止分析》。*arXiv*预印本 *arXiv:2302.06570*，2023。

Jonas Gehring、Michael Auli、David Grangier、Denis Yarats 和 Yann N Dauphin。《卷积序列到序列学习》。载于 国际机器学习会议，第 1243–1252 页。PMLR，2017。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。《用于图像识别的深度残差学习》。载于*IEEE*计算机视觉与模式识别会议论文集，第 770–778 页，2016。Samuel Horváth、Lihua Lei、Peter Richtárik 和 Michael I Jordan。《随机梯度方法在非凸优化中的自适应性》。 *arXiv*预印本 *arXiv:2002.05359*，2020。Ali Kavis、Stratis Skoulakis、Kimon Antonakopoulos、Leello Tadesse Dadi 和 Volkan Cevher。《用于非凸有限和最小化的自适应随机方差缩减》。 *arXiv*预印本 *arXiv:2211.01851*，2022。Yann LeCun、Corinna Cortes 和 Christopher J. C. Burges。《手写数字 MNIST 数据库》。1998年。网址 http://yann.lecun.com/exdb/mnist/。Lihua Lei、Cheng Ju、Jianbo Chen 和 Michael I Jordan。《通过 SCSG 方法进行非凸有限和优化》。 神经信息处理系统进展，第 30 卷，2017。Zhize Li、Hongyan Bao、Xiangliang Zhang 和 Peter Richtárik。《Page：一种用于非凸优化的简单且最优的概率梯度估计器》。载于 国际机器学习会议，第 6286–6295 页。PMLR，2021a。Zhize Li、Slavomír Hanzely 和 Peter Richtárik。《Zerosarah：一种无需计算全梯度的高效非凸有限和优化方法》。 *arXiv* 预印本 *arXiv:2103.01447*，2021b。Mingrui Liu、Zhenxun Zhuang、Yunwei Lei 和 Chunyang Liao。《一种用于训练深度神经网络的通信高效的分布式梯度裁剪算法》。 *arXiv*预印本 *arXiv:2205.05040*，2022。Stephen Merity、Nitish Shirish Keskar 和 Richard Socher。《LSTM语言模型的正则化与优化》。*arXiv* 预印本 *arXiv:1708.02182*，2017。Yurii Nesterov。 凸优化导论讲义：基础课程，第 87 卷。Springer Science & Business Media，2003年。Lam M Nguyen、Jie Liu、Katya Scheinberg 和 Martin Takáč。《用于非凸优化的随机递归梯度算法》。 *arXiv*预印本 *arXiv:1705.07261*，2017。ME Peters、M Neumann、M Iyyer、M Gardner、C Clark、K Lee 和 L Zettlemoyer。《深度上下文化词表示》。arXiv，2018年。 *arXiv*预印本 *arXiv:1802.05365*，第 12 卷，2018。Nhan H Pham、Lam M Nguyen、Dzung T Phan 和 Quoc Tran-Dinh。《Proxsarah：用于随机复合非凸优化的高效算法框架》。机器学习研究杂志，21(1):4455–4502，2020。Jiang Qian 、 Yuren Wu 、 Bojin Zhuang 、 Shaojun Wang 和 Jing Xiao。《理解增量梯度方法中的梯度裁剪》。载于国际人工智能与统计会议，第 1504–1512 页。PMLR，2021年。Sashank J Reddi、Ahmed Hefny、Suvrit Sra、Barnabás Póczos 和 Alex Smola。《用于非凸优化的随机方差缩减》。载于 国际机器学习会议，第 314–323 页。PMLR，2016年。Nicolas Roux、Mark Schmidt 和 Francis Bach。《一种具有指数收敛_速率的有限训练集随机梯度方法》。神经信息处理系统进展，第 25 卷，2012 年。Shai Shalev-Shwartz 和 Tong Zhang。《用于正则化损失最小化的随机对偶坐标上升方法》。 *arXiv*预印本 *arXiv:1209.1873*，2012年。

Lukang Sun、Avetik Karagulyan 和 Peter Richtarik。较弱光滑性条件下 Stein变分梯度下降的收敛性。*arXiv* 预印本 *arXiv:2206.00508*，2022年。Bohan Wang、Yushun Zhang、Huishuai Zhang、Qi Meng、Zhi-Ming Ma、Tie-Yan Liu 和 Wei Chen。Adam 中可证明的自适应性。*arXiv* 预印本 *arXiv:2208.09900*，2022年。Zhe Wang、Kaiyi Ji、Yi Zhou、Yingbin Liang 和 Vahid Tarokh。SpiderBoost 与动量：更快的方差缩减算法。神经信息处理系统进展，32，2019年。Tianyu Xia、Shuheng Shen、Su Yao、Xinyi Fu、Ke Xu、Xiaolong Xu、Xing Fu 和 Weiqiang Wang。差分隐私学习中的逐样本自适应裁剪。*arXiv* 预印本 *arXiv:2212.00328*，2022年。Xiaodong Yang、Huishuai Zhang、Wei Chen 和 Tie-Yan Liu。带扰动的归一化/裁剪随机梯度下降，用于差分隐私非凸优化。*arXiv* 预印本 *arXiv:2206.13033*，2022年。Bohang Zhang、Jikai Jin、Cong Fang 和 Liwei Wang。改进非凸优化裁剪算法的分析。神经信息处理系统进展，33：15511–15521，2020年。Jingzhao Zhang、Tianxing He、Suvrit Sra 和 Ali Jadbabaie。梯度裁剪为何能加速训练：自适应性的理论论证。*arXiv* 预印本 *arXiv:1905.11881*，2019年。Shen-Yi Zhao、Yin-Peng Xie 和 Wu-Jun Li。关于随机归一化梯度下降的收敛性与改进。中国科学：信息科学，64(3)：1–13，2021年。Dongruo Zhou、Pan Xu 和 Quanquan Gu。非凸优化中的随机嵌套方差缩减。机器学习研究杂志，21(1)：4130–4192，2020年。

附录

A 实验设置

在这里，我们将进一步介绍实验的详细信息，并对主论文图2中展示的结果进行定量说明。

如正文所述，我们在多个数据集上训练三种不同的神经网络，分别是三层全连接网络（全卷积网络）、ResNet-20 和 ResNet-56，所使用的数据集包括 MNIST、CIFAR10 和 CI-FAR100。在每项实验中，我们比较 （L_0、 L_1）–SPIDER 与以下基准方法的性能，包括随机梯度下降、随机方差缩减梯度、SARAH 和 SPIDER。对于每种方法，我们在一定范围内调整其超参数（如学习率），以获得尽可能高的测试准确率。除上述数据集外，我们还在其带噪版本上开展实验。具体而言，我们在带噪的 CIFAR10 数据集上训练 ResNet-20。在此过程中，我们向 CIFAR10 数据集中的图像添加方差为 1 的 ℓ_2 范数高斯噪声，并以概率 0。1 将标签均匀地更改为所有可能的类别。此外，我们将 CI-FAR10 上的噪声尺度加倍，并训练 ResNet-56 模型。正文中的图2 展示了上述实验的结果：对于每种设置，我们进行三次运行，并报告均值和标准差。此外，我们在下表中给出了相同实验的定量结果。为获得表中的每个条目，我们沿三条轨迹分别选取最佳测试准确率，并报告其平均值，如表 3所示。在补充材料中，我们提供了实验代码，该代码是在（Horváth 等人，2020年）的代码基础上修改而成。

表3：测试准确率（%），对应图 2 中的实验。 *噪声数据。

	FCN	ResNet-20			ResNet-56	
方法	MNIST	CIFAR10	CIFAR10*	CIFAR100	CIFAR10	CIFAR10*
SGD	97.83	71.31	70.52	49.02	69.36	75.26
SVRG	97.98	81.72	77.00	52.24	82.76	73.49
萨拉	97.75	79.28	77.94	53.45	79.45	74.62
Spider	98.04	81.03	77.95	53.33	81.2	73.53
(L_0, L_1)–Spider	97.91	79.94	77.82	53.39	79.69	73.83

接下来，我们给出图2中所有曲线对应的详细超参数。首先，所有方法和数据集的小批量大小均固定为 1024。对于随机梯度下降、随机方差缩减梯度和 SARAH，唯一需要调整的超参数是学习率 η_0。然而，对于 Spider，迭代 k 时的步长为 $\eta_k = \eta_0 \min\{1, c_1/\|\mathbf{v}_k\|\}$，它由学习率 η_0 和裁剪参数 c_1决定。对于 (L_0, L_1)–Spider，步长为 $\eta_k = \eta_0 \min\{1, c_1/\|\mathbf{v}_k\|, c_2/\|\mathbf{v}_k\|^2\}$，它由学习率 η_0 和两个裁剪参数 c_1 及 c_2决定。所有超参数都经过调整，以获得最佳测试准确率。我们在表4中展示了所有实验调优后的学习率。此外，表5给出了 Spider 和 (c_1, c_2)–Spider 方法对应的调优后裁剪参数 c_1 和 (L_0, L_1)。请注意，除在 CIFAR100 上训练 ResNet-20 外，我们的所有实验均未使用动量或权重衰减等技术；在该训练中，我们使用 0.9 的动量参数和 10^{-4} 的权重衰减参数，以获得更好的测试准确率。

	FCN	ResNet-20			ResNet-56	
方法	MNIST	CIFAR10	CIFAR10*	CIFAR100	CIFAR10	CIFAR10*
SGD	0.1	0.2	0.025	0.0125	0.2	1.6
SVRG	0.05	0.2	0.05	0.8	0.2	0.025
Sarah	0.025	0.05	0.025	0.025	0.0125	0.0125
Spider	0.0125	0.05	0.05	0.0125	0.025	0.0125
$(L_0,\ L_1)$–Spider	0.0125	0.025	0.025	0.05	0.0125	0.00625

表4：图 2 中实验对应的学习率。 *噪声数据。

	FCN	ResNet-20			ResNet-56	
方法	MNIST	CIFAR10	CIFAR10*	CIFAR100	CIFAR10	CIFAR10*
Spider	0.5	1	0.5	16	8	16
$(L_0,\ L_1)$–Spider	(0.5, 0.5)	(2, 2)	(2, 2)	(16, 128)	(8, 128)	(16, 128)

表5：图 2 中实验所用的裁剪参数。 *噪声数据。

B 定理3的证明

我们首先陈述并证明一个关键引理，即下降引理；它是证明非凸优化算法收敛时常用的步骤。本节中，我们用记号 $\mathbb{E}_k[\cdot]$表示在给定历史 $\mathcal{F}_k$ （其中包含 $\{\mathbf{x}_{0:k},\ \mathbf{v}_{0:k-1}\}$）条件下的期望。

引理 2 (下降引理)。 假设 F 根据定义 $(L_0,\ L_1)$ 是光滑的（参见定义 *1* ），并考虑更新 $\mathbf{x}_{k+1}=\mathbf{x}_k-\eta_k\mathbf{v}_k$。那么，对于 $\epsilon\leq\frac{L_0}{2L_1}$ 和步长

$$\eta_k\leq\min\left\{\frac{1}{2L_0},\frac{\epsilon}{L_0\|\mathbf{v}_k\|}\right\},$$

对于任意迭代 $k=0,\ 1,\cdots,$ 我们有

$$F(\mathbf{x}_{k+1})\leq F(\mathbf{x}_k)-\frac{1}{8}\eta_k\|\mathbf{v}_k\|^2+\frac{5}{8}\eta_k\left\|\mathbf{v}_k-\nabla F(\mathbf{x}_k)\right\|^2.$$

证明。证明详见第 F.2 节。 □

展示方差缩减方法收敛性的另一个重要步骤，是控制梯度估计器的方差 $\mathbb{E}\|\mathbf{v}_k-\nabla F(\mathbf{x}_k)\|^2$，其中该估计器见算法 2[2]。在下面的引理中，我们证明，通过将步长分别按 $\|\mathbf{v}_k\|$ 和 $\|\mathbf{v}_k\|^2$ 的倒数进行缩放，可以控制方差 $\mathbb{E}\|\mathbf{v}_k-\nabla F(\mathbf{x}_k)\|^2$。

引理 3. 设假设条件 *1 (i)* 和 *2* 成立，并假设 $\epsilon\leq\frac{L_0}{2L_1}$。那么，对于如下选取的步长和参数

$$\eta_k\leq\min\left\{\frac{1}{L_0}\frac{\epsilon}{\|\mathbf{v}_k\|},\frac{1}{L_1}\frac{\epsilon}{\|\mathbf{v}_k\|^2}\right\},$$

[2]此处我们不严格地将 $\mathbb{E}\|\mathbf{v}_k-\nabla F(\mathbf{x}_k)\|^2$ 称为“梯度估计器的方差”，因为$\mathbf{v}_k$ 并非 $\nabla F(\mathbf{x}_k)$ 的无偏估计量，即 $\mathbb{E}[\mathbf{v}_k|\mathcal{F}_k]\neq\nabla F(\mathbf{x}_k)$。然而，有 $\mathbb{E}[\mathbf{v}_k]=\mathbb{E}[\nabla F(\mathbf{x}_k)]$。

$$S_1 = \frac{4\sigma^2}{\epsilon^2}, \quad S_2 = 48\frac{L_0}{L_1}\frac{\sigma}{\epsilon}, \quad q = 2\frac{L_0}{L_1}\frac{\sigma}{\epsilon},$$

我们有

$$\mathbb{E}_{k_0}\left[\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2\right] \leq \epsilon^2,$$

其中， $k_0 \leq k$ 表示满足 q 整除 k_0的、最接近 k 的迭代点，即 $k_0 = \lfloor k/q \rfloor \cdot q$。

证明。证明详见第F.3节。 □

定理3的证明：建立这两个主要辅助引理后，我们来证明定理3。首先，注意到定理 $\epsilon < \frac{L_0}{20L_1}$ 中的步长和精度条件3 满足引理2所要求的条件。因此，利用该引理以及条件 $\eta_k \leq 1/(2L_0)$，可得

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \frac{1}{8}\eta_k\|\mathbf{v}_k\|^2 + \frac{5}{8}\eta_k\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2 \leq F(\mathbf{x}_k) - \frac{1}{8}\eta_k\|\mathbf{v}_k\|^2 + \frac{5}{16L_0}\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2. \quad (7)$$

此外，对于指定的步长 η_k，我们有

$$\begin{aligned}
\eta_k\|\mathbf{v}_k\|^2 &= \min\left\{\frac{\|\mathbf{v}_k\|^2}{2L_0}, \frac{\epsilon\|\mathbf{v}_k\|}{L_0}, \frac{\epsilon}{L_1}\right\} \\
&= \min\left\{\frac{\epsilon^2}{L_0}\min\left\{\frac{1}{2}\left\|\frac{\mathbf{v}_k}{\epsilon}\right\|^2, \left\|\frac{\mathbf{v}_k}{\epsilon}\right\|\right\}, \frac{\epsilon}{L_1}\right\} \\
&\overset{(a)}{\geq} \min\left\{\frac{\epsilon}{L_0}\|\mathbf{v}_k\| - \frac{2\epsilon^2}{L_0}, \frac{\epsilon}{L_1}\right\} \\
&\geq \frac{\epsilon}{L_0}\min\left\{\|\mathbf{v}_k\|, \frac{L_0}{L_1}\right\} - \frac{2\epsilon^2}{L_0}, \quad (8)
\end{aligned}$$

其中，在(a)中，我们使用了不等式$\min\{x^2/2, |x|\} \geq |x| - 2$ 对所有 x成立。将式(7)与式(8)结合并重新整理各项，可得

$$\frac{\epsilon}{8L_0}\min\left\{\|\mathbf{v}_k\|, \frac{L_0}{L_1}\right\} - \frac{\epsilon^2}{4L_0} \leq F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}) + \frac{5}{16L_0}\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2.$$

接下来，对上述不等式两边取期望，并使用引理3 中的界，可得

$$\frac{\epsilon}{8L_0}\mathbb{E}\left[\min\left\{\|\mathbf{v}_k\|, \frac{L_0}{L_1}\right\}\right] \leq \mathbb{E}[F(\mathbf{x}_k)] - \mathbb{E}[F(\mathbf{x}_{k+1})] + \frac{9}{16L_0}\epsilon^2.$$

现在，我们对等式两边关于 $k = 0, \cdots, K-1$ 取平均值，这意味着

$$\frac{\epsilon}{8L_0} \cdot \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\min\left\{\|\mathbf{v}_k\|, \frac{L_0}{L_1}\right\}\right] \leq \frac{F(\mathbf{x}_0) - \mathbb{E}[F(\mathbf{x}_K)]}{K} + \frac{9}{16L_0}\epsilon^2.$$

将两边相乘 两边乘以$\frac{8L_0}{\epsilon}$，并利用 $F(\mathbf{x}_0) - \mathbb{E}[F(\mathbf{x}_K)] \leq F(\mathbf{x}_0) - F^* = \Delta$ 这一事实可得

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\min\left\{\|\mathbf{v}_k\|, \frac{L_0}{L_1}\right\}\right] \leq \frac{8\Delta L_0}{\epsilon K} + \frac{9}{2}\epsilon \leq 5\epsilon, \quad (9)$$

其中，我们采用了如下迭代次数的选取方式

$$K = \left\lceil \frac{16\Delta L_0}{\epsilon^2} \right\rceil.$$

˜现在，考虑从 k 中均匀随机选取的索引 $\{0, \cdots, K-1\}$。(9) 中的平均值论证表明

$$\mathbb{E}\left[\min\left\{\left\|\mathbf{v}_{\tilde{k}}\right\|, \frac{L_0}{L_1}\right\}\right] \leq 5\epsilon,$$

˜其中，期望是针对 k 和算法的随机性而言的。接下来，我们使用马尔可夫不等式，得出以至少 $3/4$ 的概率，有

$$\min\left\{\left\|\mathbf{v}_{\tilde{k}}\right\|, \frac{L_0}{L_1}\right\} \leq 20\epsilon.$$

注意，对于 $\epsilon < \frac{L_0}{20L_1}$，有 $L_0/L_1 < 20\epsilon$。因此，上述界可简化为 $\left\|\mathbf{v}_{\tilde{k}}\right\| \leq 20\epsilon$。

接下来，根据引理 3 可得 $\mathbb{E}[\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\|^2] \leq \epsilon^2$ 对每个 k均成立。对于均匀选取的 $\tilde{k} \in \{0, \cdots, K-1\}$，有

$$\mathbb{E}\left[\left\|\mathbf{v}_{\tilde{k}} - \nabla F(\mathbf{x}_{\tilde{k}})\right\|^2\right] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2\right] \leq \epsilon^2,$$

结合詹森不等式可得 $\mathbb{E}\|\mathbf{v}_{\tilde{k}} - \nabla F(\mathbf{x}\tilde{k})\| \leq \epsilon$。使用马尔可夫不等式，我们以概率 $3/4$ 有 $\|\mathbf{v}_{\tilde{k}} - \nabla F(\mathbf{x}\tilde{k})\| \leq 4\epsilon$。最后，我们对上述两个良好事件使用并集界˜，从而得出，对于随机且均匀选取的索引 $k \in \{0, \cdots, K-1\}$，

$$\left\|\nabla F(\mathbf{x}_{\tilde{k}})\right\| \leq \left\|\mathbf{v}_{\tilde{k}}\right\| + \left\|\mathbf{v}_{\tilde{k}} - \nabla F(\mathbf{x}_{\tilde{k}})\right\| \leq 20\epsilon + 4\epsilon = 24\epsilon,$$

概率至少为 $1 - 1/4 - 1/4 = 1/2$。

总迭代复杂度：算法 SPIDER 中的梯度总复杂度2 可界定如下

$$\begin{aligned}
\left\lceil K \cdot \frac{1}{q} \right\rceil S_1 + KS_2 &\leq K \cdot \frac{1}{q} \cdot S_1 + S_1 + KS_2 \\
&\leq \left(\frac{16\Delta L_0}{\epsilon^2} + 1\right)\frac{L_1\epsilon}{2L_0\sigma} \cdot \frac{4\sigma^2}{\epsilon^2} + \frac{4\sigma^2}{\epsilon^2} + \left(\frac{16\Delta L_0}{\epsilon^2} + 1\right)\frac{48L_0\sigma}{L_1\epsilon} \\
&= 32\Delta\sigma\left(L_1 + 24\frac{L_0^2}{L_1}\right)\frac{1}{\epsilon^3} + \frac{4\sigma^2}{\epsilon^2} + 2\sigma\left(\frac{L_1}{L_0} + 24\frac{L_0}{L_1}\right)\frac{1}{\epsilon}.
\end{aligned}$$

C 定理4的证明

在证明定理 4（对应有限和设定）之前，我们先给出两个辅助引理。首先，引理2 可直接应用于有限和设定以及随机设定。其次，我们界定有限和设定下的方差 $\mathbb{E}\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\|^2$，这与随机设定下的引理 3 类似。

引理 4。 设假设 *1 (ii)* 成立，并假设 $\epsilon \leq \frac{L_0}{2L_1}$。然后，按照以下方式选取步长和参数

$$\eta_k \leq \min\left\{\frac{1}{L_0}\frac{\epsilon}{\|\mathbf{v}_k\|}, \frac{1}{L_1}\frac{\epsilon}{\|\mathbf{v}_k\|^2}\right\}$$

$$S_1 = n, \quad S_2 = 12\sqrt{n}, \quad q = \sqrt{n},$$

我们有

$$\mathbb{E}_{k_0}\left[\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2\right] \leq \epsilon^2,$$

其中， $k_0 \leq k$ 表示截至 k 满足 k_0可被 q 整除的最近一个迭代点，即 $k_0 = \lfloor k/q \rfloor \cdot q$。

证明。证明推迟到第F.4节。 □

定理*4*的证明：利用引理2 中的下降引理以及引理4中给出的方差界，其余证明可由定理3的证明推出。特别地，对于随机均匀˜选取的索引 $k \in \{0, \cdots, K-1\}$，我们有 $\|\nabla F(\mathbf{x}\tilde{k})\| \leq 24\epsilon$，其成立概率至少为 $1/2$。找到这样一个驻点的总迭代复杂度有界于

$$\begin{aligned}\left\lceil K \cdot \frac{1}{q}\right\rceil S_1 + KS_2 &\leq K \cdot \frac{1}{q} \cdot S_1 + S_1 + KS_2 \\ &\leq \left(\frac{16\Delta L_0}{\epsilon^2} + 1\right)\frac{1}{\sqrt{n}} \cdot n + n + \left(\frac{16\Delta L_0}{\epsilon^2} + 1\right)12\sqrt{n} \\ &= 208\Delta L_0\sqrt{n}\frac{1}{\epsilon^2} + n + 13\sqrt{n}.\end{aligned}$$

D 定理1的证明

我们首先采用 D　　下降引理（引理2 并将 $\mathbf{v}_k$ 视为随机梯度 $\mathbf{g}_k$），这意味着　　at

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \frac{1}{8}\eta_k\|\mathbf{g}_k\|^2 + \frac{5}{16L_0}\left\|\mathbf{g}_k - \nabla F(\mathbf{x}_k)\right\|^2. \tag{10}$$

接下来，对于指定的步长 η_k，我们可以写出

$$\begin{aligned}\eta_k\|\mathbf{g}_k\|^2 &= \min\left\{\frac{\|\mathbf{g}_k\|^2}{2L_0}, \frac{\epsilon\|\mathbf{g}_k\|}{L_0}\right\} \\ &= \frac{\epsilon^2}{L_0}\min\left\{\frac{1}{2}\left\|\frac{\mathbf{g}_k}{\epsilon}\right\|^2, \left\|\frac{\mathbf{g}_k}{\epsilon}\right\|\right\} \\ &\overset{(a)}{\geq} \frac{\epsilon}{L_0}\|\mathbf{g}_k\| - \frac{2\epsilon^2}{L_0},\end{aligned} \tag{11}$$

其中 (a) 是由以下事实推出的： 最小值$\{x^2/2, |x|\} \geq |x| - 2$ 对所有 x均成立。将其代回（13）可得

$$\frac{\epsilon}{8L_0}\|\mathbf{g}_k\| - \frac{\epsilon^2}{4L_0} \leq F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}) + \frac{5}{16L_0}\left\|\mathbf{g}_k - \nabla F(\mathbf{x}_k)\right\|^2.$$

我们取期望 对上述两边进行……，再结合 $\mathbb{E}\|\mathbf{g}_k - \nabla F(\mathbf{x}_k)\|^2 \leq \frac{\sigma^2}{S}$ 可得 — t

$$\mathbb{E}\|\mathbf{g}_k\| \leq \frac{8L_0}{\epsilon}\left(\mathbb{E}[F(\mathbf{x}_k)] - \mathbb{E}[F(\mathbf{x}_{k+1})]\right) + \frac{5}{2\epsilon}\cdot\frac{\sigma^2}{S} + 2\epsilon.$$

接下来，将上述不等式对 $k = 0, \cdots, K-1$ 求和，并得出

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{g}_k\| \leq \frac{8\Delta L_0}{\epsilon K} + \frac{5}{2\epsilon}\cdot\frac{\sigma^2}{S} + 2\epsilon \leq \frac{\epsilon}{2} + \frac{5}{2}\epsilon + 2\epsilon = 5\epsilon, \tag{12}$$

其中我们使用了以下参数取值

$$S = \frac{\sigma^2}{\epsilon^2}, \quad K = \left\lceil \frac{16\Delta L_0}{\epsilon^2} \right\rceil.$$

˜现在，考虑从 k 中均匀随机选取的索引 $\{0, \cdots, K-1\}$。(12) 中的平均值论证表明˜， $\mathbb{E}\|\mathbf{g}_{\tilde{k}}\| \leq 5\epsilon$ 其中期望是相对于 k 和算法中的随机性而言的。此外，随机梯度 $\mathbf{g}_k$ 的方差有界于 $\mathbb{E}\|\mathbf{g}_k - \nabla F(\mathbf{x}_k)\|^2 \leq \frac{\sigma^2}{S} = \epsilon^2$ 对每个 k均成立，结合詹森不等式可得 $\mathbb{E}\|\mathbf{g}_{\tilde{k}} - \nabla F(\mathbf{x}_{\tilde{k}})\| \leq \epsilon$。因此，我们可以写成

$$\mathbb{E}\left\|\nabla F(\mathbf{x}_{\tilde{k}})\right\| \leq \mathbb{E}\left\|\mathbf{g}_{\tilde{k}}\right\| + \mathbb{E}\left\|\mathbf{g}_{\tilde{k}} - \nabla F(\mathbf{x}_{\tilde{k}})\right\| \leq 6\epsilon.$$

˜最后，我们使用马尔可夫不等式得出，对于随机且均匀选取的索引 $k \in \{0, \cdots, K-1\}$，以至少 $1/2$ 的概率有 $\|\nabla F(\mathbf{x}_{\tilde{k}})\| \leq 12\epsilon$。

总梯度复杂度可按如下方式界定

$$KS \leq \left(\frac{16\Delta L_0}{\epsilon^2} + 1\right)\frac{\sigma^2}{\epsilon^2} = 16\Delta L_0\sigma^2\frac{1}{\epsilon^4} + \frac{\sigma^2}{\epsilon^2}.$$

E 定理2的证明

首先，注意到使用全批次（$S = n$）时，有 $\mathbf{g}_k = \nabla F(\mathbf{x}_k)$。根据下降引理（引理 2 并将 $\mathbf{v}_k = \nabla F(\mathbf{x}_k)$视为），可得

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \frac{1}{8}\eta_k\left\|\nabla F(\mathbf{x}_k)\right\|^2. \tag{13}$$

接下来，按照（11）中的相同论证，可得

$$\frac{\epsilon}{8L_0}\left\|\nabla F(\mathbf{x}_k)\right\| - \frac{\epsilon^2}{4L_0} \leq F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}).$$

对 $k = 0, \cdots, K-1$ 求和可得

$$\frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla F(\mathbf{x}_k)\right\| \leq \frac{8\Delta L_0}{\epsilon K} + 2\epsilon \leq \frac{\epsilon}{2} + 2\epsilon = \frac{5}{2}\epsilon.$$

˜设迭代点 k 从 $\{0, \cdots, K-1\}$ 中均匀随机选取。上述平均值论证表明$\mathbb{E}\|\nabla F(\mathbf{x}_{\tilde{k}})\| \leq \frac{5}{2}\epsilon$；结合马尔可夫不等式可得， $\|\nabla F(\mathbf{x}_{\tilde{k}})\| \leq 5\epsilon$的概率至少为 $1/2$。

总梯度复杂度可界定如下

$$KS \leq \left(\frac{16\Delta L_0}{\epsilon^2} + 1\right)n = 16\Delta L_0 n\frac{1}{\epsilon^2} + n.$$

F 延后引理的证明

F.1 命题1 的证明

设条件（4）成立。那么，对于任意单位范数向量 $\mathbf{u}$，有

$$\left\|\nabla^2 F(\mathbf{x})\mathbf{u}\right\| = \left\| \lim_{t\to 0} \frac{1}{t}\left(\nabla F(\mathbf{x}+t\mathbf{u}) - \nabla F(\mathbf{x})\right) \right\| = \lim_{t\to 0}\frac{1}{t}\left\|\nabla F(\mathbf{x}+t\mathbf{u}) - \nabla F(\mathbf{x})\right\| \le L_0 + L_1\|\nabla F(\mathbf{x})\|.$$

因此，我们有 $\|\nabla^2 F(\mathbf{x})\| = \sup_{\|\mathbf{u}\|=1}\|\nabla^2 F(\mathbf{x})\mathbf{u}\| \le L_0 + L_1\|\nabla F(\mathbf{x})\|$，这与条件（3）相同。反方向是以下结果的一个特例，该结果在（Zhang 等人， 2020年）中得到了证明。

引理 5（Zhang 等人 （2020年）中的推论 A.4）。 假设 F 满足 （3），即 $\|\nabla^2 F(\mathbf{x})\| \le L_0 + L_1\|\nabla F(\mathbf{x})\|$ 对所有 $\mathbf{x}$均成立。对于任意 $c > 0$，若 $\|\mathbf{x} - \mathbf{y}\| \le c/L_1$，则

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \le (AL_0 + BL_1\|\nabla F(\mathbf{x})\|)\|\mathbf{x} - \mathbf{y}\|,$$

其中 $A = 1 + e^c - \frac{e^c - 1}{c}$且 $B = \frac{e^c - 1}{c}$。

将 $c = 1$ 代入上述引理可得 $A = 2$ 以及 $B = e - 1 \le 2$。因此，条件（4）成立，其中$2L_0$ 且 $2L_1$。

F.2 引理2的证明

考虑任意一次迭代 k，并定义 $\mathbf{x}(t) = t(\mathbf{x}_{k+1} - \mathbf{x}_k) + \mathbf{x}_k$，其中任意 $t \in [0,1]$ 均位于 $\mathbf{x}_k$ 与 $\mathbf{x}_{k+1}$ 之间。根据泰勒定理，我们有

$$\begin{aligned}
F(\mathbf{x}_{k+1}) &= F(\mathbf{x}_k) + \int_0^1 \langle \nabla F(\mathbf{x}(t)), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle dt \\
&= F(\mathbf{x}_k) + \langle \nabla F(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + \int_0^1 \langle \nabla F(\mathbf{x}(t)) - \nabla F(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle \, \mathrm{d}t \\
&\overset{(a)}{\le} F(\mathbf{x}_k) + \langle \nabla F(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + \int_0^1 \left(L_0 + L_1\|\nabla F(\mathbf{x}_k)\|\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 t \, \mathrm{d}t \\
&= F(\mathbf{x}_k) + \langle \nabla F(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + \frac{1}{2}\left(L_0 + L_1\|\nabla F(\mathbf{x}_k)\|\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2, \qquad (14)
\end{aligned}$$

其中，在（a）中，我们使用了定义1，因为 $\|\mathbf{x}(t) - \mathbf{x}_k\| \le \|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \eta_k\|\mathbf{v}_k\| \le \epsilon/L_0 \le 1/(2L_1) \le 1/L_1$。我们可以通过将（14）中的 $\mathbf{x}_{k+1} - \mathbf{x}_k = -\eta_k\mathbf{v}_k$ 替换为如下形式，继续对 $F(\mathbf{x}_{k+1})$ 进行界定

$$\begin{aligned}
F(\mathbf{x}_{k+1}) &\le F(\mathbf{x}_k) - \eta_k\langle \nabla F(\mathbf{x}_k), \mathbf{v}_k\rangle + \frac{1}{2}L_0\eta_k^2\|\mathbf{v}_k\|^2 + \frac{1}{2}L_1\eta_k^2\|\mathbf{v}_k\|^2\|\nabla F(\mathbf{x}_k)\| \\
&\le F(\mathbf{x}_k) - \eta_k\langle \nabla F(\mathbf{x}_k), \mathbf{v}_k\rangle + \frac{1}{2}L_0\eta_k^2\|\mathbf{v}_k\|^2 + \frac{1}{4}L_1\eta_k^2\|\mathbf{v}_k\|^3 + \frac{1}{4}L_1\eta_k^2\|\mathbf{v}_k\|\cdot\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2 \\
&\le F(\mathbf{x}_k) - \frac{1}{2}\eta_k\left(1 - L_0\eta_k\right)\|\mathbf{v}_k\|^2 + \frac{1}{4}L_1\eta_k^2\|\mathbf{v}_k\|^3 + \frac{1}{2}\eta_k\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2 \\
&\quad + \frac{1}{4}L_1\eta_k^2\|\mathbf{v}_k\|\cdot\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2 \\
&\overset{(b)}{\le} F(\mathbf{x}_k) - \frac{1}{4}\eta_k\|\mathbf{v}_k\|^2 + \frac{1}{8}\eta_k\|\mathbf{v}_k\|^2 + \frac{1}{2}\eta_k\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2 + \frac{1}{8}\eta_k\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2 \\
&= F(\mathbf{x}_k) - \frac{1}{8}\eta_k\|\mathbf{v}_k\|^2 + \frac{5}{8}\eta_k\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2.
\end{aligned}$$

在推导上述 (b) 时，我们特别用到了步长满足条件 $L_0\eta_k \le 1/2$ 和 $\eta_k\|\mathbf{v}_k\| \le \epsilon/L_0 \le 1/(2L_1)$。

F.3 引理 3的证明

考虑迭代点 $k=0,\ 1,\ \cdots$并记 $k_0 \le k$为距离 k最近的迭代点，且满足 q整除 k_0，也就是说，$k_0 = \lfloor k/q \rfloor \cdot q$。这意味着 SPIDER 更新 $\mathbf{v}_{k_0} = \nabla f(\mathbf{x}^{k_0};\ \mathcal{S}_1)$（算法 2）。因此，对于$k=k_0$，有

$$\mathbb{E}_{k_0}\left[\left\|\mathbf{v}_{k_0} - \nabla F(\mathbf{x}_{k_0})\right\|^2\right] = \mathbb{E}_{k_0}\left[\left\|\nabla f(\mathbf{x}_{k_0};\mathcal{S}_1) - \nabla F(\mathbf{x}_{k_0})\right\|^2\right] \le \frac{\sigma^2}{S_1} = \frac{\epsilon^2}{4}.$$

对于任意 $k_0 \le k < k_0+q$，我们有

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{v}_{k+1} - \nabla F(\mathbf{x}_{k+1})\right\|^2 \mid \mathcal{F}_{k+1}\right] &= \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_{k+1};\mathcal{S}_2) - \nabla f(\mathbf{x}_k;\mathcal{S}_2) + \mathbf{v}_k - \nabla F(\mathbf{x}_{k+1})\right\|^2 \mid \mathcal{F}_{k+1}\right] \\ &= \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_{k+1};\mathcal{S}_2) - \nabla f(\mathbf{x}_k;\mathcal{S}_2) + \nabla F(\mathbf{x}_k) - \nabla F(\mathbf{x}_{k+1})\right\|^2 \mid \mathcal{F}_{k+1}\right] \\ &\quad + \left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2. \end{aligned} \tag{15}$$

注意到小批量 $\mathcal{S}_2$ 的大小为 $|\mathcal{S}_2| = S_2$，上述右端中的第一项可按如下方式界定

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_{k+1};\mathcal{S}_2) - \nabla f(\mathbf{x}_k;\mathcal{S}_2) + \nabla F(\mathbf{x}_k) - \nabla F(\mathbf{x}_{k+1})\right\|^2 \mid \mathcal{F}_{k+1}\right] \\ \le\ & \frac{1}{S_2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_{k+1};\xi) - \nabla f(\mathbf{x}_k;\xi)\right\|^2 \mid \mathcal{F}_{k+1}\right] \\ \le\ & \frac{1}{S_2}\left(L_0 + L_1\left\|\nabla F(\mathbf{x}_k)\right\|\right)^2 \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ \le\ & \frac{2}{S_2}L_0\eta_k^2\|\mathbf{v}_k\|^2 + \frac{2}{S_2}L_1^2\eta_k^2\|\mathbf{v}_k\|^2 \cdot \left\|\nabla F(\mathbf{x}_k)\right\|^2 \\ \le\ & \frac{2}{S_2}L_0^2\eta_k^2\|\mathbf{v}_k\|^2 + \frac{4}{S_2}L_1^2\eta_k^2\|\mathbf{v}_k\|^4 + \frac{4}{S_2}L_1^2\eta_k^2\|\mathbf{v}_k\|^2 \cdot \left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2 \\ \le\ & \frac{6}{S_2}\epsilon^2 + \frac{4}{S_2}\left(\frac{L_1}{L_0}\right)^2 \epsilon^2 \left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2. \end{aligned} \tag{16}$$

在推导上述最后一个不等式时，我们用到了以下事实： $L_0\eta_k\|\mathbf{v}_k\| \le \epsilon$ 且 $L_1\eta_k\|\mathbf{v}_k\|^2 \le \epsilon$。将 (15) 和 (16) 合并可得

$$\mathbb{E}\left[\left\|\mathbf{v}_{k+1} - \nabla F(\mathbf{x}_{k+1})\right\|^2 \mid \mathcal{F}_{k+1}\right] \le \left(1 + \frac{4}{S_2}\left(\frac{L_1}{L_0}\right)^2 \epsilon^2\right)\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2 + \frac{6}{S_2}\epsilon^2.$$

我们对上述不等式两边关于$\{\mathbf{x}_{k_0+1:k+1}$、$\mathbf{v}_{k_0:k}\}$中包含的所有随机性来源取期望，并以 $e_k := \mathbb{E}[\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\|^2 \mid \mathcal{F}_{k_0}] = \mathbb{E}_{k_0}[\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\|^2]$表示。因此，我们证明了非负序列 $\{e_k\}$ 对于 $k_0 \le k < k_0+q$满足以下条件

$$e_{k+1} \le a e_k + b$$

其中

$$a = 1 + \frac{4}{S_2}\left(\frac{L_1}{L_0}\right)^2 \epsilon^2, \quad b = \frac{6}{S_2}\epsilon^2, \quad \text{and} \quad e_{k_0} \le \frac{\epsilon^2}{4}.$$

这意味着对于 $k_0 \le k < k_0+q$，我们有

$$e_k \le a^{k-k_0}e_{k_0} + b\sum_{i=0}^{k-k_0-1} a^i \le a^q e_{k_0} + b\sum_{i=0}^{q-1} a^i \le a^q e_{k_0} + bqa^q.$$

接下来，我们 将上述不等式中指定的 q 和 S_2 取值代入，如下所述 g,

$$q = 2\frac{L_0}{L_1}\frac{\sigma}{\epsilon}, \quad S_2 = 48\frac{L_0}{L_1}\frac{\sigma}{\epsilon}.$$

我们首先将 a^q 界定如下，

$$a^q = \left(1 + \frac{4}{S_2}\left(\frac{L_1}{L_0}\right)^2\epsilon^2\right)^q \leq \left(1 + \frac{1}{\sigma}\left(\frac{L_1}{L_0}\frac{\epsilon}{2}\right)^3\right)^q = \left(1 + \frac{\tilde{\epsilon}^3}{\sigma}\right)^{\sigma/\tilde{\epsilon}} \leq \left(1 + \frac{1}{16}\frac{\tilde{\epsilon}}{\sigma}\right)^{\sigma/\tilde{\epsilon}} \leq 2,$$

其中，我们记 $\tilde{\epsilon} = \frac{L_1}{L_0}\frac{\epsilon}{2} \leq \frac{1}{4}$，并利用了这样一个事实：对于任意 $x > 0$，都有 $(1 + x/16)^{1/x} \leq 2$。此外，另一项 bqa^q 可界定如下，

$$bqa^q \leq 2bq = 2 \cdot \frac{6}{S_2}\epsilon^2 \cdot q = 2 \cdot \frac{1}{8}\frac{L_1}{L_0}\frac{\epsilon}{\sigma} \cdot \epsilon^2 \cdot 2\frac{L_0}{L_1}\frac{\sigma}{\epsilon} = \frac{\epsilon^2}{2}.$$

综上所述，我们已证明

$$\mathbb{E}_{k_0}\left[\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2\right] = e_k \leq a^q e_{k_0} + bqa^q \leq 2\frac{\epsilon^2}{4} + \frac{\epsilon^2}{2} = \epsilon^2,$$

由此完成证明。

F.4 引理4

证明过程与引理 3中的步骤相同。从 $k = k_0$ 开始，SPIDER 计算全批次梯度。因此，$\mathbf{v}_{k_0} = \nabla F(\mathbf{x}^{k_0})$ 和

$$\mathbb{E}_{k_0}\left[\left\|v_{k_0} - \nabla F(\mathbf{x}_{k_0})\right\|^2\right] = 0.$$

接下来，对于任意 $k_0 \leq k \leq k_0 + q$，我们可以采用引理 3 证明中的论证（见第 F.3 节）。特别地，由（15）和（16），可得

$$\mathbb{E}\left[\left\|\mathbf{v}_{k+1} - \nabla F(\mathbf{x}_{k+1})\right\|^2 \mid \mathcal{F}_{k+1}\right] \leq \left(1 + \frac{4}{S_2}\left(\frac{L_1}{L_0}\right)^2\epsilon^2\right)\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2 + \frac{6}{S_2}\epsilon^2.$$

与引理 3 的证明类似，对上述不等式两边关于包含在 $\{\mathbf{x}_{k_0+1:k+1}$ 中的所有随机性来源以及 $\mathbf{v}_{k_0:k}\}$ 在给定 $\mathcal{F}_{k_0}$ 的条件下取期望，并记 $e_k := \mathbb{E}[\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\|^2 \mid \mathcal{F}_{k_0}]$。因此，我们证明了非负序列 $\{e_k\}$ 对于 $k_0 \leq k < k_0 + q$ 满足以下关系

$$e_{k+1} \leq ae_k + b$$

其中

$$a = 1 + \frac{4}{S_2}\left(\frac{L_1}{L_0}\right)^2\epsilon^2, \quad b = \frac{6}{S_2}\epsilon^2, \quad \text{and} \quad e_{k_0} = 0.$$

由此可得

$$e_k \leq b\sum_{i=0}^{k-k_0-1} a^i \leq b\sum_{i=0}^{q-1} a^i \leq bqa^q.$$

首先，对于我们选择的参数 $S_2 = 12\sqrt{n}$ 和 $q = \sqrt{n}$，可以如下界定 a^q —

$$a^q = \left(1 + \frac{4}{S_2}\left(\frac{L_1}{L_0}\right)^2 \epsilon^2\right)^q = \left(1 + \frac{1}{3q}\left(\frac{L_1}{L_0}\epsilon\right)^2\right)^q = \left(1 + \frac{1}{12\sqrt{n}}\right)^{\sqrt{n}} \leq 2,$$

其中我们用到了以下事实：$\frac{L_1}{L_0}\epsilon \leq \frac{1}{2}$。最后，对于每个 $k_0 \leq k < k_0 + q$，我们有

$$\mathbb{E}_{k_0}\left[\left\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\right\|^2\right] = e_k \leq bqa^q \leq \frac{6}{S_2}\epsilon^2 \cdot q \cdot 2 = \epsilon^2,$$

由此完成证明。