arXiv:2305.01588v2 [cs.LG] 9 Nov 2023

重新审视梯度裁剪：随机偏差与紧确的收敛保证

Anastasia Koloskova [*1] **Hadrien Hendrikx** [*2] **Sebastian U. Stich** [3]

摘要

梯度裁剪是对标准（随机）梯度下降的一种常见改进：每次迭代都将梯度范数限制在某个值 $c > 0$。例如，它广泛用于稳定深度学习模型的训练（Goodfellow 等，2016年），或用于实现差分隐私（Abadi 等人，2016年）。尽管裁剪机制广受欢迎且简单易用，其收敛保证往往要求对 c 取特定值，并对噪声作出严格假设。

本文给出了收敛保证，明确刻画了其对任意裁剪阈值 c 的依赖，并证明这些保证对于确定性梯度和随机梯度都达到了紧确界。具体而言，我们证明：（i）对于确定性梯度下降，裁剪阈值只会影响收敛的高阶项；（ii）在随机设定下，即使步长任意小，在标准噪声假设下也无法保证收敛到真正的最优点。我们给出了裁剪随机梯度下降中梯度范数收敛的匹配上界和下界，并通过实验展示了这些结果。

1. 引言

本文重点研究求解以下形式的一般最小化问题

$$\min_{\mathbf{x}\in\mathbb{R}^d} \{f(\mathbf{x}) := \mathbb{E}_{\xi\sim\mathcal{D}}[f_\xi(\mathbf{x})]\}, \tag{1}$$

其中，f 是一个可能为非凸函数且可能具有随机性的函数。这一设定涵盖了许多应用场景，例如，当 $f_\xi \equiv f\ \forall \xi$ 时，它也涵盖了对确定性函数的优化。此外，它还涵盖机器学习应用中经验损失的最小化，其中 $\mathcal{D}$ 表示训练数据点上的均匀分布，而 $f_\xi(\mathbf{x})$ 是模型 $\mathbf{x}$在数据点 ξ上的损失。

我们研究采用梯度裁剪来求解（1）的梯度下降方法。给定裁剪半径 $c > 0$、步长$\eta > 0$，并从点$\mathbf{x}_0 \in \mathbb{R}^d$ 开始，梯度裁剪算法执行以下迭代：

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{g}_t, \quad \text{with} \quad \mathbf{g}_t = \text{clip}_c(\nabla f_\xi(\mathbf{x}_t)), \tag{2}$$

其中，$\mathbf{g}_t$ 是裁剪后的随机梯度，裁剪算子定义为

$$\text{clip}_c(\mathbf{u}) = \min\left(1, \frac{c}{\|\mathbf{u}\|}\right)\mathbf{u}, \qquad \text{for } \mathbf{u}\in\mathbb{R}^d. \tag{3}$$

梯度裁剪被广泛用于稳定神经网络的训练，防止偶发的大梯度值对训练造成损害（Goodfellow 等，2016年）。这对于减轻训练数据中的离群值影响以及训练循环模型（Pascanu 等人，2012年；2013年）尤其有用，因为其中的噪声可能导致非常大的梯度。梯度裁剪也是保护隐私的机器学习不可或缺的一部分。广泛使用的高斯机制（Dwork 和 Roth，2014年）会向各个梯度添加噪声，以增加其真实值的不确定性。然而，要使隐私保证成立，它要求梯度具有有界范数。在实践中，有界梯度通过裁剪来实现（Abadi 等人，2016年）。

梯度裁剪已经得到了广泛研究，下一节将对此进行详细介绍。然而，许多研究会选择特定的裁剪阈值 c，以保证收敛。这表明实践中需要仔细调节 c，但这非常不理想，尤其是因为裁剪阈值可能由其他因素（例如隐私）决定。此外，大多数研究都通过大批量（从而降低随机梯度噪声）、角度条件或范数的一致有界性，对随机梯度噪声施加了较强的假设，而这些假设在实践中可能并不成立。

在这项工作中，我们精确刻画了裁剪阈值 c 如何影响裁剪随机梯度下降在任意裁剪阈值 c下的收敛性质。

[*]共同贡献 [1]洛桑联邦理工学院，瑞士 [2]法国国家信息与自动化研究所格勒诺布尔中心，法国（部分工作在洛桑联邦理工学院期间完成） [3]德国 CISPA 亥姆霍兹信息安全中心。通信作者：Anastasia Koloskova <anastasia.koloskova@epfl.ch>、Hadrien Hendrikx <hadrien.hendrikx@inria.fr>。

《40th 国际机器学习大会》，美国夏威夷州檀香山。PMLR 202，2023 年。版权所有 © 2023，作者所有。

我们分别考虑确定性函数和随机函数，因为裁剪对这两种情形的影响方式不同。

在确定性情形下，裁剪只会改变所应用梯度的大小，而不会改变其方向。这意味着，裁剪梯度下降可以到达 f 的临界点，但速度较慢。直观而言，随着算法逐渐收敛，梯度的大小会变小，最终不再被裁剪。这意味着，裁剪只会在梯度较大时的初始阶段影响收敛速度。主要挑战在于精确刻画这一额外开销。

在随机情形下，情况则有所不同：即使期望梯度很小，单个随机梯度仍可能很大。即使在 f 的临界点处，期望（全量）梯度为零，单个随机梯度仍有一定概率被裁剪。此外，由于我们不假设随机梯度具有任何对称性，即使在 f 的临界点处，裁剪后的期望梯度也可能非零，从而迫使算法偏离这些临界点。直接结果是，裁剪随机梯度下降通常不会收敛到 f 的临界点(Chen 等人，2020年），而只会收敛到某个邻域。本文研究裁剪引入的偏差，并表明该偏差取决于噪声方差 σ^2 和裁剪参数 c，我们在第1.1节中对其进行了精确定义。正如我们将在第 1.2节中进一步详述的那样，现有工作通过采用较大的裁剪阈值或较大的小批量，或对噪声作出较强假设（例如一致有界性、随机梯度之间的角度受限等）来规避这一困难，而且通常还要求 c 取特定值。与此不同，我们对裁剪随机梯度下降的收敛性进行精确分析，并在不作任何额外假设的情况下，精确刻画裁剪所引入的偏差。

更具体地说，我们的贡献如下：

- 在确定性设置下，我们分析了裁剪梯度下降在非凸、凸和强凸函数上的收敛行为。我们的分析表明，在所有情形中，经过一段暂态阶段后，裁剪不会影响收敛率。在凸和非凸情形下，这一初始阶段不会影响收敛的主导项。然而，在强凸情形下，这个不可避免的初始阶段无法确保初始条件呈线性遗忘，从而导致收敛速度显著减慢。

- 对于随机梯度，我们证明，在“重尾”假设下，裁剪随机梯度下降会收敛到一个邻域，其大小为 $\min\{\sigma, \sigma^2/c\}$，以梯度范数衡量。

- 我们证明，这一邻域大小是紧的：只要步长足够小，裁剪随机梯度下降确实能够将梯度范数降低至 $\min\{\sigma, \sigma^2/c\}$ 。

- 我们使用$(L_0,\ L_1)$-光滑性假设（Zhang 等人，2019年）来表述我们的结果。这是对光滑性的一种标准放宽，非常适合分析裁剪算法。

通过这些结果，我们旨在全面、准确地描绘裁剪的收敛保证：在与标准随机梯度下降相同的假设下，并适用于任意裁剪阈值 c。我们的目标是通过这些更精确的界限，收紧所有下游应用（例如隐私）的保证；这些应用将裁剪随机梯度下降的收敛结果视为黑盒。事实上，在这些情形中，裁剪阈值通常被视为问题的外部参数，而我们灵活的保证使得能够针对 c 优化界限，并在收敛速度（或精度）与特定应用的要求之间进行权衡。

1.1. 主要假设

在讨论相关工作之前，我们将首先陈述本文采用的假设。

关于光滑性的假设。 优化文献中广泛使用的光滑性假设（例如 Nesterov 等人，2018）如下：

假设 1.1 （光滑性）。 函数 f 满足

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L \|\mathbf{x} - \mathbf{y}\|\,, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d\,.$$

尽管这一假设应用广泛，但它可能有所局限，因为常数 L 必须涵盖最坏情况下的光滑性。 Zhang 等人(2019年)通过实验发现，对于各种深度学习任务，局部光滑性常数 L在训练过程中会不断减小，并且与梯度范数成正比。他们报告称，在训练的最后阶段，局部曲率（光滑性）可能比初始化点处的曲率小 1000 倍（在 PTB 数据集上训练长短期记忆网络时）。$(L_0,$ $L_1)$-光滑性（Zhang 等人，2019年；2020a）被提出，作为对经典光滑性假设的一种自然放宽。

假设 1.2 ((L_0, L_1)-光滑性)。 如果可微函数 $f\colon \mathbb{R}^d \to \mathbb{R}$ 对于所有 $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ 满足 $\|\mathbf{x} - \mathbf{y}\| \le \frac{1}{L_1}$，则称其为 (L_0, L_1)-光滑函数：

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le (L_0 + \|\nabla f(\mathbf{x})\| L_1) \|\mathbf{x} - \mathbf{y}\|\,. \quad (4)$$

我们将此作为本文的主要假设。通过取1.1可恢复标准的光滑性假设 $L_1 = 0$。然而，取 $L_1 > 0$ 则可以使原本不光滑的函数也具有类似光滑函数的性质，例如$\mathbf{x} \mapsto \|\mathbf{x}\|^3$。此外， L-光滑函数可能是(L_0, L_1)-光滑的，且这两个常数 L_0, L_1 都显著小于 L，例如指数函数 $x \mapsto e^x$。

表 1。 关键假 设比较及非凸确定性情形下复杂度估计的示意（方差 $\sigma = 0$ ）.

参考文献	光滑性	方差上界	裁剪阈值	进一步的假设	收敛速率（非凸， $\sigma = 0$）	
Zhang 等人 (2019年)	二阶（L_0， L_1）	一致有界	$c = \Theta(\text{最小}\{L_0, \frac{L_0}{L_1}\})$		$\mathcal{O}\left(\frac{1}{\sqrt{\eta T}c}\right)$	$\eta \leq \text{最小}\left\{\frac{1}{10L_0}, \frac{1}{10cL_1}\right\}$
Zhang 等人 (2020a)	(L_0， L_1)	一致有界	$c = \Theta(\text{最大}\{\epsilon, \frac{L_0}{L_1}\})$		$\mathcal{O}\left(\frac{1}{\sqrt{\eta T}}\right)$	$\eta \leq \frac{1}{10L_0}$
Chen 等人(2020年)	L	期望	任意	正偏度	$\mathcal{O}\left(\frac{1}{\sqrt{\eta T}} + \frac{1}{\eta T c} + \sqrt{\eta L c^2}\right)$	
Qian 等人（2021年）	(L_0， L_1)	期望	任意	位置对齐	$\mathcal{O}\left(\frac{1}{\sqrt{\eta T}} + \frac{1}{\eta T c} + \sqrt{\eta L_0 c^2} + \sqrt{\eta L_1 c^3}\right)$	$\eta \leq \frac{1}{4cL_1}$
ours	(L_0， L_1)	期望	任意		$\mathcal{O}\left(\frac{1}{\sqrt{\eta T}} + \frac{1}{\eta T c}\right)$	$\eta \leq \frac{1}{9(L_0 + cL_1)}$

请注意，式（ $\|\mathbf{x} - \mathbf{y}\| \leq \frac{1}{L_1}$ 中的界限4）至关重要；否则，梯度的全局增长将与标准光滑函数受到类似限制（因此会排除前述$\mathbf{x} \mapsto \|\mathbf{x}\|^3$等函数）。

在关于裁剪算法的研究中，Zhang 等人(2019年)采用了稍强的光滑性条件，要求具备二阶可微性。对于二阶可微函数 f，他们将 (L_0, L_1)-光滑性定义为

$$\left\|\nabla^2 f(\mathbf{x})\right\| \leq L_0 + L_1 \|\nabla f(\mathbf{x})\|, \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{5}$$

后来，Zhang 等人(2020a)发现，较弱的假设 1.2 足以用于研究裁剪算法。我们在本文中采用他们的定义。

关于随机方差的假设。 裁剪领域的许多研究（Zhang 等人，2019年；2020a；Yang 等人，2022年）都假设

假设 1.3 （一致有界性）。如果对于所有$\mathbf{x} \in \mathbb{R}^d$， f_ξ 的随机噪声都由 σ^2 一致有界，则称其满足一致有界性，

$$\Pr\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \sigma^2\right] = 1. \tag{6}$$

尽管这一假设有助于简化裁剪算法的分析，但它是一个非常强的假设。

通过作出假设1.3 并使用足够大的裁剪半径（$c > \sigma$），我们可以保证在 f 的临界点处，随机梯度不会被裁剪，其中$\nabla f(\mathbf{x}) = 0$。这确保算法能够收敛到精确的临界点，从而简化（Zhang 等人，2019年；2020a；Yang 等人，2022年）中的理论分析。

一致有界性假设 1.3是一个较强的假设，可能并不总能反映现实世界中的情况。例如，如果梯度受到高斯噪声的扰动，则一致有界性假设不成立。此外，在机器学习应用中，其中$\nabla f_\xi(\mathbf{x})$ 表示模型 $\mathbf{x}$ 在数据集 ξ 中的不同数据点 $\xi \in \mathcal{D}$处的梯度，则如果数据集 σ 中哪怕只有一个离群点，一致有界性 $\mathcal{D}$ 的上界也可能很大。

在本研究中，我们改用较弱且更为标准的方差定义（Lan，2012年；Dekel 等人，2012年），有时也称为重尾噪声（Gorbunov 等人，2020年）。

假设1.4 （有界方差）。 我们称 f_ξ 的方差由 σ^2 界定，如果对于所有$\mathbf{x} \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x})\|^2\right] \leq \sigma^2. \tag{7}$$

请注意，一致有界性意味着有界方差（且常数相同），但反之不成立。

1.2. 相关工作

关于梯度裁剪的文献已十分丰富，且仍在不断发展。下文介绍与我们工作最相关的贡献，并在表1中列出部分研究。

裁剪能够稳定学习。 梯度裁剪最初由（Mikolov，2012年）提出，旨在解决训练循环神经网络时的梯度爆炸问题。 Zhang 等人(2019年) 提出了一种理论解释，说明为何裁剪随机梯度下降能够提升（随机）一阶方法的稳定性：他们采用较宽松的二阶 (L_0, L_1)-光滑性假设（见公式 (5)），并证明了裁剪随机梯度下降相比未裁剪随机梯度下降具有收敛优势。然而，他们依赖关于随机方差的强假设 1.3 ，并将裁剪阈值设为一个特定的足够大值。随后，Zhang 等人 (2020a) 改进了这些有利的收敛保证，但仍依赖关于随机噪声的假设 1.3 ，并为裁剪阈值选择特定值。 Mai 和 Johansson (2021年) 表明，在非光滑情形下也存在同样的问题。

噪声假设。 梯度裁剪通常在对梯度的随机噪声满足一致有界性假设 1.3 的条件下进行分析，同时选择足够大的裁剪阈值 $c > \sigma$ （Zhang 等人，2019年；2020a；Yang 等人，2022年）。选择足够大的 c 可以简化理论分析。然而，在某些应用中，裁剪阈值 c 可能会受到其他约束（例如隐私约束）的限制。尤其是由于许多实际应用中的随机噪声具有重尾特性（Zhang 等人，2020b年），这将导致 σ取较大值，因此 c。

为避免一致有界噪声假设，一些研究对随机梯度的分布施加了其他强假设。例如，Qian 等人(2021年) 重新-

限制随机梯度与真实梯度之间的夹角，而Chen 等人(2020年) 对随机梯度的分布施加了对称性假设。Gorbunov 等人(2020年) 分析了有界方差条件下的裁剪（见假设1.4），但他们对小批量的大小施加了强假设，要求其随 T线性增长，从而使有效随机方差随迭代次数 T 而 $\mathcal{O}\left(\frac{\sigma^2}{T}\right)$递减。

本文采取了不同于上述所有工作的研究路线，在更弱得多的有界方差 假设下分析裁剪随机梯度下降。然而，我们并不试图收敛到 f 的精确临界点，而是量化裁剪导致的漂移有多大，从而为任意批大小和 c 提供保证。

无噪声情形。 由于大部分假设都与随机噪声有关，我们在确定性设置下采用的设定与上述论文相同。不过，在这种情形下，我们给出了更精确的保证，基本证明了裁剪不会影响主要的收敛项（见表 1）。

裁剪联邦平均。 Zhang 等人(2022a) 通过裁剪发送至服务器的模型差异，研究了 FedAvg（McMahan 等人，2016年）算法的裁剪版本。然而，这需要有界梯度，而且当裁剪阈值为 时，其收敛率无法恢复 FedAvg 的收敛率。此外，即使使用确定性梯度，裁剪联邦平均仍存在偏差。 Liu 等人(2022年) 也研究了一种类似裁剪联邦平均的算法，并通过假设噪声分布关于其均值对称，消除了偏差问题。

满足差分隐私的随机梯度下降。 差分隐私已成为保护隐私的黄金标准，因此引起了随机优化领域的关注（Chaud- huri 等人，2011年；Song 等人，2013年；Duchi 等人，2014年）。然而，为了满足差分隐私，随机梯度必须满足有界性（Wang 等人，2017；Bassily 等人，2019年；Das 等人，2022年）（或满足相关条件，例如目标函数的利普希茨连续性）。这一条件在实践中很少成立，通常需要通过裁剪来强制满足，例如在差分隐私随机梯度下降算法中（Abadi 等人，2016年）。实际上，利普希茨连续性要求梯度有界，而光滑性只要求海森矩阵有界。尽管在有界域上，光滑性蕴含利普希茨连续性，但这一界通常过于保守，因而导致较差的保证。

Bagdasaryan 等人（2019年）通过实验测量了差分隐私随机梯度下降（裁剪和额外噪声）对模型准确率的影响。他们观察到，梯度范数并未收敛至零，因此，证明精确收敛所依赖的假设实际上往往并不成立。此外，代表性不足的类别具有更高的梯度范数（因此，差分隐私随机梯度下降会影响公平性）。

与自适应方法的联系。值得注意的是，裁剪随机梯度下降与自适应方法有关，例如 Adam 算法（Kingma 和 Ba，2015年）或归一化随机梯度下降（Hazan 等人，2015年；Levy，2016年），这些方法也会对梯度进行缩放。然而，这两种算法都不等同于裁剪随机梯度下降，而且 Adam 算法（Reddi 等人，2018；Zhang 等人，2020b；2022b）和归一化随机梯度下降（Zhao 等人，2021年）的收敛结果无法直接推广到裁剪随机梯度下降。

2. 确定性设置

本节考虑采用完整（确定性）梯度的梯度裁剪算法（2），即

$$\nabla f_\xi(\mathbf{x}) \equiv \nabla f(\mathbf{x}), \qquad \forall \xi \in \mathcal{D}, \forall \mathbf{x} \in \mathbb{R}^d. \tag{8}$$

在这种设置下，裁剪算子（3）只改变所应用梯度的大小，而不改变其方向（这与对裁剪随机梯度取期望不同）。因此，我们可以预期算法收敛到函数 f的精确极小值（或临界点）。然而，梯度大小的这种变化会在多大程度上影响算法的收敛速度，仍不清楚。

我们的理论结果表明，只有当函数 f 是强凸时，才会出现显著的减速；在这种情况下，应用裁剪后，初始条件（到最优点的距离）不再以线性方式被遗忘。不过，误差 ϵ 中的主导项不受影响。如果函数f 是凸或非凸，裁剪阈值 c不会影响收敛的主导项，只会影响高阶项。

2.1. 非凸函数

定理2.1 （非凸）。 如果 f 满足假设*1.2*，那么采用确定性梯度（8）的裁剪梯度下降（2）并使用步长 $\eta \leq [9\left(L_0 + cL_1\right)]^{-1}$ 可保证误差：

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\| \leq \mathcal{O}\left(\sqrt{\frac{F_0}{\eta T}} + \frac{F_0}{\eta T c}\right), \tag{9}$$

其中 T是迭代次数， $F_0 = f\left(\mathbf{x}_0\right) - f^\star$。

该定理是定理 3.3 的推论，适用于 $\sigma = 0$。

与未裁剪梯度下降的比较。假设满足标准 L-光滑性假设时，梯度下降（不进行裁剪）的收敛率1.1 等于（Ghadimi & Lan，2013年）：

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|^2 \leq \mathcal{O}\left(\frac{F_0}{\eta T}\right), \tag{10}$$

其中，步长必须小于 $\eta \leq \frac{1}{L}$。在接下来的讨论中，我们假设 $L_0 + cL_1 \leq L$，因为我们总可以令 L_1 为零。在许多情况下，L_0 和 L_1 都显著小于 L（如第1.1节所述）。因此，与未裁剪梯度下降相比，裁剪梯度下降（2）：

（i）允许采用更大的步长 η（步长约束中的常数最高可达 9）。这一结果源于更精细的（L_0，L_1）光滑性假设；步长的这种改进与Zhang 等人（2019年）对（L_0，L_1）二阶光滑性假设（5）所得的发现思路相同，不过他们的步长上界有所不同。

（ii）还包含一项 $\frac{F_0}{\eta T c}$，该项取决于裁剪半径 c。这一项的阶为 $\frac{1}{T}$，而主导项（衰减最慢、渐近上占主导的项）的阶为 $\frac{1}{\sqrt{T}}$。

如果 c 很小，这一项会显著减慢算法。不过，当 c 大于最终目标精度 ϵ 时，裁剪只会使收敛速度受到一个常数因子的影响。直观地说，这是因为发生裁剪的步数仅占收敛所需总迭代次数的一个固定比例。[1] 由于我们通常知道最终目标精度，我们的结果表明，可以设置裁剪阈值来避免裁剪带来的不利影响。不过在实践中，裁剪阈值可能还需由其他需求决定。

(iii) 采用了不同的收敛度量 $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|$，而不是 $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|^2$这一更常用的度量（例如，未裁剪梯度下降的式（10））。

与先前研究的比较。我们在表 1 中总结了与先前研究的差异。Zhang 等人（2019）和 Zhang 等人（2020a）分析了确定性梯度裁剪，但将裁剪阈值 c设为某些特定的、足够大的值。Qian 等人（2021）和 Chen 等人（2020）研究了在任意选择裁剪阈值 c的情况下的裁剪随机梯度下降。特别地，在假设梯度为确定性梯度（$\sigma = 0$）时，Qian 等人（2021）得到的收敛速率为 $\mathcal{O}\left(\sqrt{\frac{F_0}{\eta T}} + \frac{F_0}{\eta T c} + c\sqrt{\eta L_0} + c^{3/2}\sqrt{\eta L_1}\right)$，$\eta < 1/4cL_1$；在恒定步长 η下，最后两项 $c\sqrt{\eta}L_0 + c^{3/2}\sqrt{\eta}L_1$ 不会趋于零。这严格劣于我们在定理2.1中得到的结果。Chen 等人（2020）证明了收敛速率 $\mathcal{O}\left(\sqrt{\frac{F_0}{\eta T}} + \frac{F_0}{\eta T c} + c\sqrt{\eta L}\right)$，且不对步长施加任何限制；然而，由于项 $c\sqrt{\eta L}$不会递减，他们必须采用较小的步长 $\eta = 1/\sqrt{T}$，

[1]形式上，由于最终精度为 $\epsilon = \sqrt{F_0/\eta T + F_0/\eta T c} \geq \sqrt{F_0/\eta T}$，因此如果裁剪阈值大于该值，$c \geq \sqrt{F_0/\eta T}$ $\sqrt{F_0/\eta T + F_0/\eta T c} \leq$，那么收敛速 $2\sqrt{F_0/\eta T}$ 度只会受到一个常数因子的影响。

T。值得注意的是，这一项会阻止误差在恒定步长下收敛到 0；而正如我们上文所示，在确定性设置下可以实现这一点。

2.2. 凸函数

现在我们证明当 f 为凸函数时的一个等价定理，即进一步假设：

假设 2.2（凸性）。函数 f 满足

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

我们还假设，f 的下确界在 $\mathbb{R}^d$处取得。

定理 2.3 (凸)。 若 f 是 L-光滑[2] *(假设 1.1)*，(L_0, L_1) 光滑（假设 *1.2*）且凸（假设 *2.2*），则采用确定性梯度 (8) 和步长 $\eta \leq (L_0 + cL_1)^{-1}$的裁剪梯度下降法（(2) 可保证误差满足：

$$f(\mathbf{x}_T) - f^\star \leq \mathcal{O}\left(\frac{R_0^2}{\eta T} + \frac{R_0^4 L}{\eta^2 T^2 c^2}\right), \tag{11}$$

其中，$R_0^2 = \|\mathbf{x}_0 - \mathbf{x}_\star\|^2$、$f^\star = f(\mathbf{x}_\star)$，以及$\mathbf{x}^\star = \arg\min_{\mathbf{x}} f(\mathbf{x})$。

相比之下，在与定理 2.3 相同的假设下，未裁剪梯度下降以该收敛速率收敛：

$$f(\mathbf{x}_T) - f^\star \leq \mathcal{O}\left(\frac{R_0^2}{\eta T}\right)$$

在步长小于 $\eta \leq L^{-1}$ (Nesterov 等人，2018）的条件下。与非凸情形类似，裁剪梯度下降的收敛率会因高阶项 $R_0^4 L/\eta^2 T^2 c^2$ 而减慢。然而，同样只需将裁剪阈值 c 设为大于最终目标精度 ϵ（这次乘以 $\sqrt{L}$），即可避免该项导致的减速，因为在追求较高准确率时，实际有更多时间使用未裁剪梯度。

此外，与非凸情形类似，裁剪 GD 允许采用更大的步长 η，从而实现更快的收敛。

2.3. 强凸函数

本节考虑强凸函数 f。

假设 2.4（强凸性）。存在一个常数$\mu > 0$，使得函数 f 对所有 $\mathbf{x}$，$\mathbf{y} \in \mathbb{R}^d$均满足，

$$f(\mathbf{x}) - f(\mathbf{y}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle.$$

与凸情形和非凸情形类似，裁剪不会影响收敛的主导项（如 $\epsilon \to 0$），正如下述定理所示。这是因为对于任意固定的 $c > 0$，梯度最终都不会再被裁剪。

[2]我们可以放宽这一假设。方程（11）中以 L 替换为 L_T，定义为 L_T：＝最大值$_{t \leq T}\{L_0 + L_1\|\nabla f(\mathbf{x}_t)\|\}$。

定理 2.5 (强凸情形)。 若 f 是 μ-强凸（假设2.4），L-光滑[3]（假设 1.1），且$(L_0,\ L_1)$ 光滑（假设 1.2），则采用确定性梯度(2) 的裁剪梯度下降(8) 并使用步长 $\eta \leq (L_0 + cL_1)^{-1}$ 至多需要

$$T = \mathcal{O}\left(\frac{1}{\mu\eta}\log\left(\frac{R_0^2}{\epsilon}\right) + \frac{R_0}{c\eta}\min\left(\sqrt{\frac{L}{\mu}}, \frac{LR_0}{c}\right)\right) \quad (12)$$

达到准确率 $R_T^2 \leq \epsilon$所需的迭代次数，其中 $R_t^2 = \|\mathbf{x}_t - \mathbf{x}_\star\|^2$ 且 $\mathbf{x}^\star = \arg\min_{\mathbf{x}} f(\mathbf{x})$。

与未裁剪情形相比，强凸情形中多出了一项，该项不会随 ϵ 减小，并对应于裁剪开销。这意味着，在收敛初始阶段，当梯度被裁剪时，收敛速度是次线性的；若要使裁剪不影响收敛速度，就必须设置 $c = \mathcal{O}(1/\log(\frac{1}{\epsilon}))$，这一要求远高于非凸和凸情形。

直观地说，由于裁剪梯度范数固定，迭代实际上每一步最多只能朝最优点移动 ηc。如果到最优点的初始距离 R_0 恰好很大，那么在最理想的情况下，至少需要 $\frac{R_0}{\eta c}$ 步才能到达最优点。最小值中第一项对c的依赖是紧的。[4]

请注意，经过常数（与 ϵ 无关）次迭代后，该算法将线性收敛，其收敛速率仅取决于 (L_0+cL_1)，而这对 L 的依赖可能显著弱于梯度下降收敛速率中的依赖。

3. 随机函数

在确定性设置下，梯度裁剪分别能够使算法收敛到精确最小值点或驻点，而且裁剪只会影响初始收敛速度。然而，在随机设定下，情况并非如此，因为裁剪会引入不可避免的偏差。其主要原因是，裁剪后的随机梯度的期望与裁剪真实梯度不同（无论是范数还是方向）。

尽管此前已有研究指出，在有界方差假设 1.4 (Chen 等人，2020年 下，裁剪随机梯度下降不会收敛，但在本节中，我们将精确刻画裁剪随机梯度下降所能达到的误差下界，然后，我们给出与这些下界相匹配的上界。

3.1. 裁剪引入的不可避免的偏差

如果梯度裁剪算法收敛，它必然收敛到其不动点，即满足 $\mathbf{x}^\star$ 使得$\mathbb{E}\left[\text{裁剪}_c(\nabla f_\xi(\mathbf{x}^\star))\right] = 0$的点。通过采用足够小的步长（以抵消随机噪声），即可在极限情况下实现这一点。

现在，我们正式证明一个下界，该下界表明，裁剪随机梯度下降的不动点不一定是目标函数 f 的最优点或临界点。事实上，存在某些随机梯度噪声分布，使得裁剪随机梯度下降的临界点与 f 的临界点相距 σ 甚远。

定理 *3.1*（小裁剪半径）。我们固定一类方差至多为 σ^2 （定义 *1.4*）、光滑性参数为 $L_0 = 1, L_1 = 0$ （假设 *1.2*）的函数。于是，对于任意裁剪阈值 $c \leq 2\sigma$，我们都能在这一固定函数类中找到一个函数 f，使得裁剪随机梯度下降的不动点存在（即满足 $x^\star$，且$\mathbb{E}\left[\text{裁剪}_c(\nabla f_\xi(x^\star))\right] = 0$的点）；并且对于所有此类裁剪随机梯度下降的不动点 $x^\star$，都有 $\|\nabla f(x^\star)\| \geq \sigma/12$。

证明概要。 我们定义随机函数

$$f_\xi(x) = \frac{1}{2}\begin{cases}(x+a)^2 & \text{w. p. } p\\ x^2 & \text{w. p. } (1-p)\end{cases}$$

其中 $a > 0$ 和 $p < 1/2$。因此，期望函数为$f(x) = \frac{1}{2}[p(x+a)^2 + (1-p)x^2]$。
然后，通过选择 $a = 4\sigma$ 和 $p = (2-\sqrt{3})/4 < 1/4$ 使得 $p(1-p) = 1/16$，即可得到该结果。利用标准代数运算，结论随之得出，详见附录C.5。 □

当裁剪半径较小时（$c < 2\sigma$），前述不可能性结果成立。我们进一步证明，增大裁剪半径 c 可以将裁剪随机梯度下降收敛到的邻域大小从 σ 减小至 σ^2/c，但无法将其完全消除。

定理 3.2 (大裁剪半径)。 我们固定一个函数类，其中函数的方差至多为 σ^2 (定义 *1.4*)，光滑参数为 $L_0 = 1$, $L_1 = 0$ (假设 *1.2*)。那么，对于任意裁剪阈值 $c \leq 2\sigma$，我们都能在该固定函数类中找到一个函数 f，使得裁剪随机梯度下降的$\mathbf{x}^\star$ 不动点存在（$\mathbb{E}\left[\text{裁剪}_c(\nabla f_\xi(\mathbf{x}^\star))\right] = 0$），且$\|\nabla f(\mathbf{x}^\star)\| \geq \sigma^2/6c$。

证明概要。 我们使用与定理 3.1相同的函数，这次取 $a = 2c$ 和 $p(1-p) = \sigma^2/a^2$。 □

定理中的下界3.1 和3.2 表明，仅假设定义1.4 和假设1.2，就不可能

[3]我们可以通过改为定义 $L := \max_{t\leq T}\{L_0 + L_1\|\nabla f(\mathbf{x}_t)\|\}$.[4]来放宽这一假设。从形式上说，考虑函数 $x \mapsto \frac{1}{2}x^2$ 和初始迭代点 $x_0 = 1$。假设我们的目标是达到目标精度 $\epsilon < \frac{1}{4}$，裁剪阈值为 $c < \frac{1}{2}$。我们可以看到，除非 $|x| < c$，否则梯度$f'(x) = x$ 将被裁剪为 c，因此经过 $\frac{1}{2c}$ 次迭代后，无法达到范数小于1_2 且平方范数小于1_4 的点。

存在依赖于问题的 c 取值，使其对任意函数都能实现精确收敛。我们指出，当噪声仅在期望意义下有界时，裁剪可能引入偏差，这并非新发现（Chen 等人，2020年）。定理3.1 和3.2 的意义在于，定理3.3中的上界与之相匹配，这意味着我们精确刻画了这种情况下引入的偏差大小。

一致有界噪声。 请注意，这一下界的成立关键依赖于噪声在 σ 期望意义下 （假设 1.4）有界。事实上，这些结果取决于随机梯度以概率p被裁剪这一事实，从而引入了偏差。如果我们保持该伯努利噪声不变（因此可以确保一致有界噪声假设1.3成立），并增大 c，那么对于 $c \approx a$，这一偏差就会完全消失，因为此时裁剪半径将大于方差的一致上界。

3.2. 收敛结果

现在我们介绍本文的核心结果：裁剪随机梯度下降的收敛性，与上述下界相匹配。

定理3.3。 如果 f 是（L_0, L_1）-光滑（但不一定是凸的），并且我们以步长$\eta \leq 1/[9(L_0+cL_1)]$,运行裁剪随机梯度下降 T 步，那么$\min_{t\in[0,T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|$的上界为

$$\mathcal{O}\left(\min\left\{\sigma, \frac{\sigma^2}{c}\right\} + \sqrt{\eta(L_0+cL_1)}\sigma + \sqrt{\frac{F_0}{\eta T}} + \frac{F_0}{\eta T c}\right),$$

其中 $F_0 = f(\mathbf{x}_0) - f^\star$。

收敛率包含四项：第一项不会随着步长 η 或迭代次数 T 的增加而减小，这是由不可避免的偏差导致的，正如上一节所述。根据定理 3.1、3.2，这一项是紧确的，无法改进。第二项是随机噪声项，会随步长增大而减小；最后两项是优化项，描述了随机噪声为零时裁剪对收敛的影响（$\sigma = 0$），与定理 2.1中的收敛结果一致。值得注意的是，我们在一般的有界方差假设下精确量化了裁剪随机梯度下降的偏差。

与未裁剪随机梯度下降的比较。 在标准光滑性假设1.1下，未裁剪随机梯度下降要求步长小于 $\eta \leq L^{-1}$，并以如下速率收敛（Bottou等人，2018）

$$E_T \leq \mathcal{O}\left(\sqrt{\eta L}\sigma + \sqrt{\frac{F_0}{\eta T}}\right).$$

其中[5] $E_T := \left(\frac{1}{T}\sum_{t=0}^{T}\|\nabla f(\mathbf{x}_t)\|^2\right)^2$。与未裁剪随机梯度下降相比，裁剪随机梯度下降（2），

- 具有一个不可避免的偏差项 最小值$\left\{\sigma, \frac{\sigma^2}{c}\right\}$ 我们已在上一节中对此进行了详细讨论 3.1。
- 在[6] $L_0 + cL_1 \leq L$的假设下，随机噪声项更小。
- 与确定性情形类似（定理2.1），还多出一个高阶项 $F_0/\eta T c$。

我们想强调，收敛率中的偏差项$\min\left\{\sigma, \frac{\sigma^2}{c}\right\}$是紧的。这尤其意味着，在一般的期望有界噪声假设1.4下，裁剪随机梯度下降无法收敛到f的精确临界点，但收敛邻域大小会随着 c的增大而减小。

与Zhang 等人（2019年）类似，裁剪随机梯度下降将光滑性参数的依赖关系从 L 改善为 L_0。与Zhang 等人（2019年）不同，我们的工作不假设 c具有特定取值，并采用较弱的期望有界噪声假设1.4。

定理2.5 的完整证明见附录C。现在，我们给出定理3.3的直观证明概要。

证明概要（定理 *3.3*）。证明分为两种不同的情况，具体取决于 c 相对于 σ 的大小。

情况 $c \leq 4\sigma$。在这种情况下，根据定理 3.1 中的下界，我们只能证明梯度范数收敛到 $\Theta(\sigma)$。为此，只需考虑梯度满足 $\|\nabla f(\mathbf{x}_t)\| \geq 6\sigma$ 的情况，因为否则收敛到 $\Theta(\sigma)$ 已经实现。我们可以证明，在梯度较大的情况下（即 $\|\nabla f(\mathbf{x}_t)\| \geq 6\sigma$），在噪声一致有界的条件下，标准的收敛结果仍然成立，因为梯度范数足够大，能够抵消（固定的）方差。

情况 $c \geq 4\sigma$。在这种情况下，我们将裁剪随机梯度下降分析为某种带偏差的梯度下降。注意，在一致有界性（假设 1.3）条件下，对于足够大的裁剪阈值，偏差最终会消失。这里我们对剩余偏差项 B_t 进行精确量化。经过一些推导，我们得到附录 C中的方程（32）所示的下降项，以及一个形式如下的偏差项：

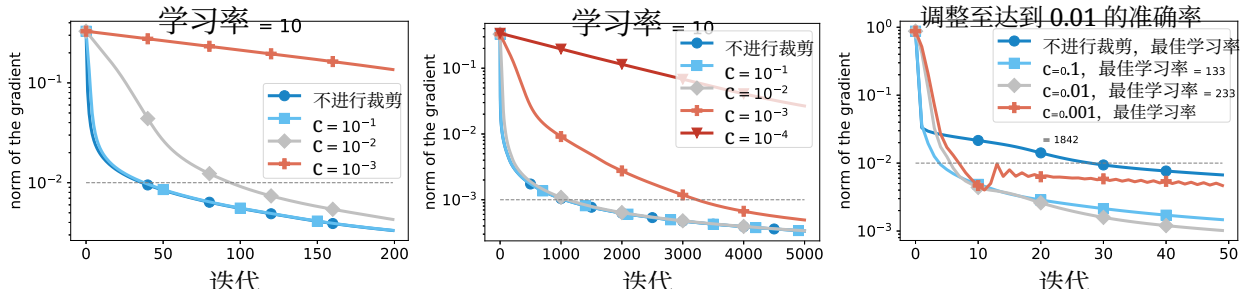
$$B_t = \|\mathbb{E}\left[\mathrm{clip}_c(\nabla f_\xi(\mathbf{x}_t))\right] - \mathrm{clip}_c(\nabla f(\mathbf{x}_t))\|^2. \quad (13)$$

利用裁剪操作是凸集（半径为 c 的球）上的投影这一事实，我们可以直接将该项界定为 $B_t \leq \sigma^2$。特别是，当 $\|\nabla f(\mathbf{x}_t)\|$ 足够大时，我们可以用下降项将其抵消。

当 $\|\nabla f(\mathbf{x})\| \leq c/2$ 时，我们可以给出更精确的结果。特别地，我们可以证明，随机梯度被裁剪的概率小于 σ^2/c^2。利用这一点，我们

[5] 注意，$E_T \geq \min_{t\in[0,T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|$。

[6] 我们总是可以选择 $L_0 = L$ 和 $L_1 = 0$ 来满足该方程。通常，L_0 和 L_1 都远小于 L（参见第 1.1 节的讨论）



(a) 恒定步长，目标值 $\epsilon = 10^{-2}$ (b) 恒定步长，目标值 $\epsilon = 10^{-3}$ (c) 针对目标值 $\epsilon = 10^{-2}$调优步长

图 *1.* 在 w1a 数据集上进行确定性裁剪梯度下降。我们研究收敛率对裁剪参数的依赖关系 c。在图 (a) 和 (b) 中可以看到，只要裁剪阈值小于或等于目标梯度范数ϵ，收敛速度就只受到常数因子的影响。在图 (c) 中可以看到，随着裁剪阈值 c 降低，为达到 $\epsilon = 10^{-2}$ 最快速度而调优的步长也会减小。这些观察结果与定理2.3一致。

可以将偏差的估计细化为

$$B_t \leq 8\frac{\sigma^4}{c^2} + 32\frac{\sigma^4}{c^4}\left\|\nabla f(\mathbf{x})\right\|^2. \tag{14}$$

The σ^4/c^2 是我们在收敛率中发现的偏差项，而 $\|\nabla f$（$\mathbf{x}$）$\|^2$ 项可以与下降项（它们也与此成正比）相互抵消，前提是 σ^4/c^4 足够小。

□

与先前研究的比较。所有先前研究都采用了更强的假设，从而简化了分析，并有助于减轻裁剪引入的偏差。

例如，（Zhang 等人，2019年）、（Zhang 等人，2020a）、（Yang 等人，2022年）考虑了较大的裁剪阈值（$c \geq \sigma$）和更强的一致有界性假设（假设1.3），从而确保随着我们接近最优点，偏差趋于消失。另一项先前研究（Gorbunov 等人，2020年）采用了特定的裁剪阈值 c 以及特定的、足够大的批大小，因此也只需考虑其中一种情况（$c \geq 4\sigma$）。他们要求批大小随迭代次数 T线性增长，从而减轻梯度中的随机性。（Qian 等人，2021年）和（Chen 等人，2020年）对随机梯度的分布施加了一些对称性假设，使其能够减轻裁剪算子引入的偏差。在完全对称分布这一极限情况下，裁剪算子不会改变任意点处期望梯度的方向，因此可以采用与确定性梯度类似的收敛分析。

3.3. 扩展至满足差分隐私的随机梯度下降

在差分隐私随机梯度下降（Abadi 等人，2016年）中，批次中的每个随机梯度都会先单独裁剪，然后再对批次中的梯度求平均，

也就是说，该算法

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\left(\frac{1}{B}\sum_{i\in\mathcal{B}_t}\mathrm{clip}_c(\nabla f_{\xi_i}(\mathbf{x}_t)) + \mathbf{z}_t\right), \tag{15}$$

其中 $\mathbf{z}_t \sim \mathcal{N}\left(0, \frac{\sigma^2}{d}_{\text{DP}}\mathbf{I}\right)$ 是满足差分隐私所引入的额外噪声。

如附录 C.4中所述，我们可以以直接的方式将分析扩展到该算法，并证明T 次迭代（15）即可使梯度范数小于：

$$\mathcal{O}\left(\frac{L\eta}{c}\sigma_{\text{DP}}^2 + \sqrt{L\eta\sigma_{\text{DP}}} + \min\left(\sigma^2, \frac{\sigma^4}{c^2}\right) + \eta L\frac{\sigma^2}{B} + \frac{F_0}{\eta T} + \frac{F_0^2}{\eta^2 T^2 c^2}\right).$$

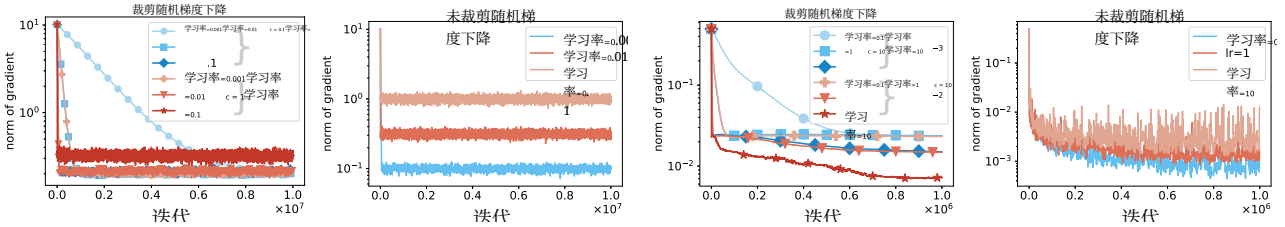
其中 B 是小批量大小，$L = L_0 + \max_t \|\nabla f(\mathbf{x}_t)\| L_1$对应于根据标准的 L-光滑性假设得到的光滑性常数。

与前文考虑的裁剪随机梯度下降算法类似，差分隐私随机梯度下降也存在偏差项$\min\left(\sigma^2, \sigma^4/c^2\right)$。定理 3.1、3.2 中的下界同样适用于差分隐私随机梯度下降，因此该偏差也是紧的且不可避免。

与裁剪随机梯度下降（2）相比，差分隐私随机梯度下降还包含与注入的隐私噪声相关的额外项 σ_{DP}，并且由于小批量处理，随机噪声（第四项）减少了 B。

为了获得形式化的隐私保证，必须恰当设置额外差分隐私噪声的方差，Abadi 等人 (2016年)证明，对于 $\sigma_{\text{DP}} \geq \Omega\left(cd\frac{\sqrt{T\log\frac{1}{\delta}}}{\epsilon}\right)$差分隐私随机梯度下降，它满足 (ϵ, δ)-满足差分隐私。

与先前在收敛分析中纳入裁剪的差分隐私随机梯度下降（DP-SGD）研究相关（Chen 等人，2020年；Yang 等人，



(a) 含有随机噪声的二次函数　　(b) 在 w1a 数据集上进行逻辑回归（批量大小 = 1）。

图 2. 在含有 χ^2 随机噪声的二次函数上进行随机梯度下降（左），以及在 w1a 数据集上进行随机梯度下降（右）。不进行裁剪时，减小步长可以达到更小的梯度范数。然而，进行裁剪时，减小步长并不能带来更好的性能。这是因为定理 3.3中存在不可避免的偏差项。

2022年），我们的收敛率仅在假设期望意义下的有界方差（定义 1.4）且不对噪声作额外假设的情况下得到证明。它们展示了裁剪阈值对差分隐私随机梯度下降（DP-SGD）收敛的影响。

4. 实验

本节中，我们研究在 w1a 数据集（Platt，1998年）上进行逻辑回归时梯度裁剪的性能，以及在人工二次函数 $f(\mathbf{x}) = \mathbb{E}_{\xi\sim\chi^2(1)}\left[f(\mathbf{x}\ \xi) := \frac{L}{2}\|\mathbf{x}\|^2 + \langle \mathbf{x}\ \xi\rangle\right]$上的性能，其中 $\mathbf{x} \in \mathbb{R}^{100}$，我们选择 $L = 0.1$，而 $\chi^2($1) 是自由度为 1 的（逐坐标）卡方分布。目标是突出展示我们的理论结果。

确定性设置。 在确定性设置中，我们的一项发现是，只要裁剪阈值大于最终目标精度，裁剪就不会过度降低性能。我们通过图1(a)、1(b) 中的逻辑回归实验对此进行测试，使用`w1a` 数据集，绘制不同目标精度下采用不同 c值的裁剪梯度下降曲线。我们看到，要达到准确率 10^{-3}，除了 c （除了 $c = 10^{-4}$之外）的所有取值表现都相对不错。然而，如果我们只想达到误差 $\epsilon = 10^{-2}$，那么不建议选择$c = 10^{-3}$，如图 1(b) 所示（请注意两幅图中 x轴的刻度不同）。

在图 1(c)中，我们考察裁剪阈值 c 与步长之间的关系。我们针对每个裁剪参数 c，在 10^{-1} 到 10^4, 之间的对数网格上分别调整步长，以确保最优值不在网格边缘。选择最快达到目标梯度范数 $\epsilon = 10^{-2}$的步长。在图 1(c) 中可以看到，选择较小的裁剪半径可以采用更大的步长，从而整体加快收敛。

特别是，我们已经验证，在确定性设定中，只要裁剪阈值相对于目标精度而言不太小，裁剪就不会损害学习。此外，裁剪能够稳定学习，从而允许采用更大的步长——规模，因此收敛更快。

随机设定。 我们研究了在带有 $\chi^2(1)$ 随机噪声的二次函数上使用裁剪随机梯度下降，以及在 `w1a`数据集上进行逻辑回归。结果绘制在图2中。这些结果也验证了理论，因为使用裁剪时，学习率越大总是越好（与未裁剪相比）。另一个有趣的现象是，曲线由乘积 $c\eta$决定；这个乘积表示发生裁剪时步长有多大。不过，我们看到，在这种情况下， c取值如此之小，裁剪随机梯度下降的表现仍略逊于原始随机梯度下降。

5. 结论

在本文中，我们对梯度裁剪进行了严格分析，既涵盖确定性设置，也涵盖标准噪声假设下的情形。尽管以往工作侧重于强假设下的精确收敛（尤其通常针对固定的裁剪阈值），但我们针对任意裁剪阈值，对裁剪随机梯度下降所引入的偏差进行了精确刻画（给出了上界和下界）。

我们的工作为进一步理解裁剪在其他算法中的应用铺平了道路，例如加速方法、动量方法或 FedAvg。具体而言，这有助于改进对隐私保证的分析，适用于依赖裁剪随机梯度下降这一底层黑盒的应用；既包括现有应用，也包括未来应用。

致谢

作者谨此感谢 Ryan McKenna 和 Martin Jaggi 富有成效的讨论。我们还感谢匿名审稿人提出的宝贵意见。SS 感谢 Meta 隐私增强技术研究奖 2022年 提供的部分资助。AK 感谢 Google 博士奖学金 提供的资助。

参考文献

Abadi, M., Chu, A., Goodfellow, I., Mcmahan, H. B., Mironov, I., Talwar, K., 和 Zhang, L. 差分隐私下的深度学习。载于 *2016年ACM SIGSAC计算机与通信安全会议*，CCS '16，第308–318页，美国纽约州纽约市，2016年。计算机协会。ISBN 9781450341394。doi: 10.1145/2976749.2978318。URL https://doi.org/10.1145/2976749.2978318。

Bagdasaryan, E., Poursaeed, O., 和 Shmatikov, V. 差分隐私对模型准确率的影响各不相同。第*33*届国际神经信息处理系统大会论文集，2019年。

Bassily, R., Feldman, V., Talwar, K., 以及 Guha Thakurta, A. 具有最优速率的私有随机凸优化。神经信息处理系统进展，32，2019年。

Bottou, L., Curtis, F. E., 以及 Nocedal, J. 大规模机器学习的优化方法。*SIAM*评论，60(2):223–311，2018。doi: 10.1137/16M1080173。URL https://doi.org/10.1137/16M1080173。

Chaudhuri, K., Monteleoni, C., 以及 Sarwate, A. D. 差分隐私经验风险最小化。机器学习研究期刊，12(3)，2011年。

Chen, X., Wu, S. Z., and Hong, M. 理解隐私保护随机梯度下降中的梯度裁剪：几何视角。载于 Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H.（编），神经信息处理系统进展，第 33 卷，第 13773–13782 页。Curran Associates, Inc.，2020年。网址 https://proceedings.neurips.cc/paper/2020/file/9ecff5455677b38d19f49ce658ef0608-Paper.pdf。Das, R., Kale, S., Xu, Z., Zhang, T., and Sanghavi, S. 超越一致的 Lipschitz条件：差分隐私优化。 *arXiv* 预印本 *arXiv:2206.10713*，2022年。Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. 使用小批量进行最优分布式在线预测。机器学习研究期刊，13(null)：165–202，2012 年 1 月。ISSN 1532-4435。Duchi, J. C., Jordan, M. I., and Wainwright, M. J. 隐私感知学习。 美国计算机学会会刊（*JACM*），61(6)：1–57，2014年。Dwork, C. and Roth, A. 差分隐私的算法基础。 理论计算机科学基金会与趋势，9(3–4)：211–407，2014年 8 月。ISSN 1551-305X。doi：10.1561/0400000042。网址 https://doi.org/10.1561/0400000042。Ghadimi, S. and Lan, G. 用于非凸随机规划的随机一阶和零阶方法。 *SIAM*优化期刊，23(4)：2341–2368，2013年。Goodfellow, I., Bengio, Y., and Courville, A. 深度学习。麻省理工学院出版社，2016年。 http://www.deeplearningbook.org。Gorbunov, E., Danilova, M., and Gasnikov, A. 通过加速梯度裁剪实现带重尾噪声的随机优化。载于 第 *34* 届国际神经信息处理系统会议论文集，NIPS'20，美国纽约州雷德胡克，2020年。Curran Associates Inc. 国际标准书号 9781713829546。

Hazan, E., Levy, K. Y., and Shalev-Shwartz, S. 超越凸性：随机拟凸优化。载于 第 *28* 届神经信息处理系统国际会议论文集——第 *1* 卷，NIPS'15，第 1594–1602 页，美国马萨诸塞州剑桥，2015年。麻省理工学院出版社。

金马，Adam：以及随机优化方法。载于 Bengio, Y. 和 LeCun, Y.（编），第 *3* 届国际学习表征会议，*ICLR 2015*，美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议专题论文集，2015 年。网址：http://arxiv.org/abs/1412.6980。

Lan, G. 一种用于随机复合优化的最优方法。《数学规划》，133:1–33，2012年 06。doi: 10.1007/s10107-010-0434-y。

Levy, K 归一化的力量：更快地逃离鞍点。11 2016年。

Liu, M., Zhuang, Z., Lei, Y., 和 Liao, C. 一种用于训练深度神经网络的通信高效分布式梯度裁剪算法。发表于神经信息处理系统进展，2022年。

Mai, V. 随机梯度裁剪的稳定性与收敛性：超越利普希茨连续性与光滑性。在国际机器学习大会，2021年。

McMahan, H. 以及 By Ar cas, B. A. 从去中心化数据中进行深度网络的通信高效学习。在人工智能与统计国际会议上，2016年。

Mikolov, 基于神经网络的统计语言模型。博士论文，布尔诺理工大学，2012年。

Nesterov, Y. 等。凸优化讲义，第 137 卷。施普林格，2018。Pascanu, R.、Mikolov, T. 和 Bengio, Y. 理解梯度爆炸问题。*arXiv*，abs/1211.5063，2012。Pascanu, R.、Mikolov, T. 和 Bengio, Y. 论循环神经网络训练的困难。载于 第 *30* 届机器学习国际会议论文集——第 *28* 卷，ICML'13，第 III–1310–III–1318 页。JMLR.org，2013。Platt, J. 使用序列最小优化快速训练支持向量机。核方法进展——支持向量学习，麻省理工学院出版社，1998。Qian, J.、Wu, Y.、Zhuang, B.、Wang, S. 和 Xiao, J. 理解增量梯度方法中的梯度裁剪。载于 Banerjee, A. 和 Fukumizu, K.（编），第 *24* 届人工智能与统计国际会议论文集，第 130 卷，收录于机器学习研究论文集，第 1504–1512 页。PMLR，2021 年 4 月 13–15 日。网址：https://proceedings.mlr.press/v130/qian21a.html。Reddi, S. J.、Kale, S. 和 Kumar, S. 论 Adam 及其后续方法的收敛性。载于国际学习表征会议，2018。网址：https://openreview.net/forum?id=ryQu7f-RZ。Song, S.、Chaudhuri, K. 和 Sarwate, A. D. 使用差分隐私更新的随机梯度下降。载于*2013*年电气电子工程师学会全球信号与信息处理会议，第 245–248 页。电气电子工程师学会，2013年。

Wang, D.、Ye, M. 和 Xu, J. 重新审视差分隐私经验风险最小化：更快且更通用。神经信息处理系统进展，30，2017年。Yang, X.、Zhang, H.、Chen, W. 和 Liu, T.-Y. 用扰动实现满足差分隐私的非凸优化归一化/裁剪随机梯度下降，2022年。网址https://arxiv.org/abs/2206.13033。Zhang, B.、Jin, J.、Fang, C. 和 Wang, L. 非凸优化裁剪算法的改进分析。神经信息处理系统进展，33 15521，2020 15511：–a。Zhang, J.、He, T.、Sra, S. 和 Jadbabaie, A. 梯度裁剪为何能加速训练：自适应性的理论依据，2019年。网址https://arxiv.org/abs/1905.11881。Zhang, J.、Karimireddy, S. P.、Veit, A.、Kim, S.、Reddi, S.、Kumar, S. 和 Sra, S. 为什么自适应方法适用于注意力模型？载于第*34*届神经信息处理系统国际会议论文集，NIPS'20，美国纽约州雷德胡克，2020b年。Curran Associates Inc. 国际标准书号 9781713829546。Zhang, X.、Chen, X.、Hong, M.、Wu, S. 和 Yi, J. 理解联邦学习中的裁剪：收敛与客户端级差分隐私。载于 Chaudhuri, K.、Jegelka, S.、Song, L.、Szepesvari, C.、Niu, G. 和 Sabato, S.（编），第*39*届国际机器学习会议论文集，第 162 卷，机器学习研究论文集，第 26048–26067 页。PMLR，2022年7月17日至23日。网址https://proceedings.mlr.press/v162/zhang22b.html。Zhang, Y.、Chen, C.、Shi, N.、Sun, R. 和 Luo, Z. Adam 无需修改更新规则即可收敛。*arXiv*，abs/2208.09632，2022b。Zhao, S.-Y.、Xie, Y.-P. 和 Li, W.-J. 关于随机归一化梯度下降的收敛与改进。中国科学：信息科学，64，2021年。

A. (L_0, L_1) 光滑性的含义

引理 A.1. 如果假设*1.2* 成立，则还成立：

$$f(\mathbf{y}) - f(\mathbf{x}) \leq \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{(L_0 + \|\nabla f(\mathbf{x})\| L_1)}{2} \|\mathbf{x} - \mathbf{y}\|^2 , \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \text{ with } \|\mathbf{x} - \mathbf{y}\| \leq \frac{1}{L_1} . \tag{16}$$

证明见（Zhang 等人，2020a），附录 A.1。

引理 A.2。 如果假设*1.2* 成立，则还可得到

$$\|\nabla f(\mathbf{x})\|^2 \leq 2(L_0 + L_1 \|\nabla f(\mathbf{x})\|) (f(\mathbf{x}) - f^\star) \qquad \forall \mathbf{x} \in \mathbb{R}^d,$$

其中 $f^\star = \text{下确界}_{\mathbf{x}} f(\mathbf{x})$。

引理 *A.2* 的证明。 我们首先应用前面的引理 A.1 来证明 $\mathbf{y} = \mathbf{x} - \frac{1}{L_0 + \|\nabla f(\mathbf{x})\| L_1} \nabla f(\mathbf{x})$。注意到 $\|\mathbf{x} - \mathbf{y}\| = \frac{\|\nabla f(\mathbf{x})\|}{L_0 + \|\nabla f(\mathbf{x})\| L_1} \leq \frac{1}{L_1}$，我们可以应用以下不等式：

$$f^\star \leq f\left(\mathbf{x} - \frac{1}{L_0 + \|\nabla f(\mathbf{x})\| L_1} \nabla f(\mathbf{x})\right) \overset{(16)}{\leq} f(\mathbf{x}) - \frac{1}{2(L_0 + \|\nabla f(\mathbf{x})\| L_1)} \|\nabla f(\mathbf{x})\|^2 ,$$

重新整理后即可得到所需的性质。 □

B. 确定性证明

本节包含论文中的主要证明。我们跳过非凸证明，因为它将直接由随机结果推出。

B.1. 凸情形（定理2.3）

定义 $\alpha_t = \min\left\{1, \frac{c}{\|\nabla f(\mathbf{x}_t)\|}\right\}$，我们有：

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 &\leq \|\mathbf{x}_t - \mathbf{x}^\star - \eta \alpha_t \nabla f(\mathbf{x}_t)\|^2 = \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \eta^2 \alpha_t^2 \|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t \eta \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle \\ &\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \eta^2 \alpha_t^2 \|\nabla f(\mathbf{x}_t)\|^2 - 2\eta \alpha_t (f(\mathbf{x}_t) - f^\star) .\end{aligned}$$

我们考虑两种情况：发生裁剪时，以及未发生裁剪时。

情况 1： $\alpha_t = 1$，即 $\|\nabla f(\mathbf{x}_t)\| \leq c$。于是

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \eta^2 \|\nabla f(\mathbf{x}_t)\|^2 - 2\eta (f(\mathbf{x}_t) - f^\star) .$$

利用 (L_0, L_1) 引理 A.2 中光滑性与凸性的推论，

$$\|\nabla f(\mathbf{x}_t)\|^2 \leq 2(L_0 + L_1 \|\nabla f(\mathbf{x}_t)\|) (f(\mathbf{x}_t) - f^\star) \leq 2(L_0 + L_1 c) (f(\mathbf{x}_t) - f^\star) .$$

进一步地，

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + 2(L_0 + L_1 c)\eta^2 (f(\mathbf{x}_t) - f^\star) - 2\eta (f(\mathbf{x}_t) - f^\star) ,$$

并令 $\eta \leq \frac{1}{2(L_0 + L_1 c)}$，我们得到

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \eta (f(\mathbf{x}_t) - f^\star) .$$

情况 2： $\alpha_t = \frac{c}{\|\nabla f(\mathbf{x}_t)\|}$，这意味着 $\|\nabla f(\mathbf{x}_t)\| > c$。然后，

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \eta^2 c^2 - 2\eta \frac{c}{\|\nabla f(\mathbf{x}_t)\|} (f(\mathbf{x}_t) - f^\star) .$$

如果满足 $\eta^2 c^2 \leq \eta \frac{c}{\|\nabla f(\mathbf{x}_t)\|}(f(\mathbf{x}_t) - f^\star)$，那么我们将得到

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \eta \frac{c}{\sqrt{2L}} \sqrt{(f(\mathbf{x}_t) - f^\star)}\,. \tag{17}$$

现在来看看满足什么步长条件时 $\eta \leq \frac{1}{c\|\nabla f(\mathbf{x}_t)\|}(f(\mathbf{x}_t) - f^\star)$ 成立：对右侧进行上界估计。根据 (L_0, L_1)光滑性（以及引理 A.2) 我们知道 $(f(\mathbf{x}_t) - f^\star) \geq \frac{\|\nabla f(\mathbf{x}_t)\|^2}{2(L_0 + L_1\|\nabla f(\mathbf{x}_t)\|)}$ 因此

$$\frac{1}{c\|\nabla f(\mathbf{x}_t)\|}(f(\mathbf{x}_t) - f^\star) \geq \frac{1}{2(L_0 \frac{c}{\|\nabla f(\mathbf{x}_t)\|} + L_1 c)} \geq \frac{1}{2(L_0 + L_1 c)}\,,$$

其中最后一个不等式成立，是因为 $\frac{c}{\|\nabla f(\mathbf{x}_t)\|} \leq 1$ 根据我们对 α_t 在此情形下的假设。这意味着，采用步长 $\eta \leq \frac{1}{2(L_0 + L_1 c)}$ 时，将有 $\eta \leq \frac{1}{c\|\nabla f(\mathbf{x}_t)\|}(f(\mathbf{x}_t) - f^\star)$ 因此（17）成立。

对两种情况求和。 我们将 $\mathcal{T}_1$ 定义为未发生裁剪的迭代次数集合，将 $\mathcal{T}_2$ 定义为发生裁剪的迭代次数集合。对 $T+1$ 次迭代取平均值

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_1}(f(\mathbf{x}_t) - f^\star) + \frac{1}{T+1}\sum_{t\in\mathcal{T}_2}\frac{c}{\sqrt{2L}}\sqrt{f(\mathbf{x}_t) - f^\star} \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\eta(T+1)}\,.$$

这意味着 (i)

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_1}(f(\mathbf{x}_t) - f^\star) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\eta(T+1)}\,,$$

和 (ii)

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_2}\sqrt{f(\mathbf{x}_t) - f^\star} \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2\sqrt{2L}}{\eta c(T+1)}\,.$$

对于第一个不等式 (i)，利用 $x^2 \geq 2\epsilon x - \epsilon^2$ 对任意 $\epsilon,\ x > 0$成立，并为简洁起见定义 $A := \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\eta(T+1)}$，可得

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_1}\left(2\epsilon\sqrt{f(\mathbf{x}_t) - f^\star} - \epsilon^2\right) \leq A\,,$$

因此,

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_1}\sqrt{f(\mathbf{x}_t) - f^\star} \leq \frac{A}{2\epsilon} + \frac{\epsilon}{2}\,.$$

取 $\epsilon = \sqrt{A}$，可得

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_1}\sqrt{f(\mathbf{x}_t) - f^\star} \leq \sqrt{A} \leq \sqrt{\frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\eta(T+1)}}\,.$$

这意味着

$$\frac{1}{T+1}\sum_{t=0}^{T}\sqrt{f(\mathbf{x}_t) - f^\star} \leq \sqrt{\frac{R_0^2}{\eta(T+1)}} + \frac{R_0^2\sqrt{2L}}{\eta c(T+1)}\,.$$

我们进一步利用 $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$，并得到最后一次迭代的收敛速率

$$\sqrt{f(\mathbf{x}_T) - f^\star} \leq \sqrt{\frac{R_0^2}{\eta(T+1)}} + \frac{R_0^2\sqrt{2L}}{\eta c(T+1)}\,.$$

将两边平方，并利用 $(a+b)^2 \leq 2a^2 + 2b^2\ \forall a,\ b$，我们得到

$$f(\mathbf{x}_T) - f^\star \leq \frac{2R_0^2}{\eta(T+1)} + \frac{4LR_0^4}{\eta^2 c^2 (T+1)^2}\,.$$

B.2. 强凸情形（定理2.5）

递归论证。首先，由于强凸函数也是凸函数，我们可以在此应用前一个定理的结果，得到

$$f(\mathbf{x}_T) - f^\star \leq \frac{2R_0^2}{\eta T} + \frac{4LR_0^4}{\eta^2 c^2 T^2}\,.$$

我们注意到 $R_0 = \|\mathbf{x}_0 - \mathbf{x}^\star\|$。利用强凸性，我们还知道

$$f(\mathbf{x}_T) - f^\star \geq \frac{\mu}{2} R_t^2\,,$$

因此，

$$R_t^2 \leq \frac{4R_0^2}{\mu\eta T} + \frac{8LR_0^4}{\mu\eta^2 c^2 T^2}\,.$$

因此，要得到 $R_t^2 \leq \frac{R_0^2}{2}$，只需取 $t \geq \max\{\frac{16}{\mu\eta}, \frac{6R_0\sqrt{L}}{\eta c\sqrt{\mu}}\}$（因为两项都会变得小于 $R_0^2/4$）。重复这一论证，我们可以看到迭代复杂度可以被界定为

$$T = \mathcal{O}\left(\frac{1}{\mu\eta}\log\left(\frac{R_0^2}{\epsilon}\right) + \frac{R_0\sqrt{L}}{\eta c\sqrt{\mu}}\right)\,.$$

小梯度。让我们从凸界（定理 2.3）重新开始。现在，我们将改为使用以下事实：

$$\|\nabla f(\mathbf{x}_t)\|^2 \leq 2L\left(f(\mathbf{x}_t) - f^\star\right) \leq 2L\frac{R_0^2}{\eta t}\left(1 + \frac{R_0^2 L}{c^2\eta t}\right)\,.$$

现在引入 t_0，使其满足：

$$t_0 = \frac{8LR_0^2}{\eta c^2}, \tag{18}$$

于是，对于所有 $t \geq t_0$，我们有：

$$\|\nabla f(\mathbf{x}_t)\| \leq c. \tag{19}$$

特别地，我们知道在 t_0之后不会发生裁剪，因此得到标准的线性收敛率，最终收敛时间为：

$$T = O\left(\frac{1}{\eta\mu}\log\left(\frac{R_0^2}{\epsilon}\right) + \frac{LR_0^2}{\eta c^2}\right)\,. \tag{20}$$

比较这两个收敛速率。请注意，两种收敛速率并无优劣之分，我们可以根据 c 与 $\sqrt{L\mu}R_0$之间的关系选择其中一种。

C. 随机证明。

现在开始证明定理3.3。证明分为两部分：首先证明收敛至 σ^2，然后针对 c的较大取值进一步完善这一结果。

C.1. 预备知识

现在我们陈述一个非常 简单的引理。它虽直接，却是我们分解的核心，因此我们在此特别强调。 e.

引理 C.1. 对于任意 $\alpha > 0$ 和$\mathbf{u} \in \mathbb{R}^d$，以下结论成立：

$$-\nabla f(\mathbf{x})^\top \mathbf{u} = -\frac{\alpha}{2}\|\nabla f(\mathbf{x})\|^2 - \frac{1}{2\alpha}\|\mathbf{u}\|^2 + \frac{1}{2\alpha}\|\mathbf{u} - \alpha\nabla f(\mathbf{x})\|^2\,. \tag{21}$$

C.2. 证明的第一部分：收敛至 σ （小 c）

为简化起见，本节假设 $c < 4\sigma$ 并证明梯度范数收敛至某个水平 σ。请注意，对 c 的这一假设并不构成限制，因为情况 $c > 4\sigma$ 已由证明的另一部分涵盖；在该部分中，我们证明了收敛至 $\mathcal{O}(\frac{\sigma^2}{c})$的效果更好。

大梯度。我们首先假设 $\|\nabla f(\mathbf{x}_t)\| \geq 6\sigma$。请注意，数值常数（相对而言）是任意的，本可以取得更紧，但我们将其选得较大，以使证明保持简洁明了。

我们首先利用 (L_0, L_1)引理 A.1中的光滑性。注意，对于任意步长 $\eta < \frac{1}{L_0 + cL_1}$ 都有 $\|\mathbf{x}_{t+1} - \mathbf{x}_t\| = \eta\|\mathbf{g}(\mathbf{x}_t)\| \leq \eta c \leq \frac{1}{L_1}$

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq -\eta \nabla f(\mathbf{x}_t)^\top \mathbf{g}(\mathbf{x}_t) + \frac{\eta^2 (L_0 + \|\nabla f(\mathbf{x}_t)\| L_1)}{2} \|\mathbf{g}(\mathbf{x}_t)\|^2 \\
&\leq -\eta \nabla f(\mathbf{x}_t)^\top \mathbf{g}(\mathbf{x}_t) + \frac{\eta^2 (L_0 + \|\nabla f(\mathbf{x}_t)\| L_1)}{2} c^2 \\
&\leq -\eta \nabla f(\mathbf{x}_t)^\top \mathbf{g}(\mathbf{x}_t) + \frac{\eta^2 (L_0 + cL_1)}{2} c \|\nabla f(\mathbf{x}_t)\| \,, \qquad (22)
\end{aligned}
$$

其中，最后一个不等式是因为我们假设了 $c \leq 4\sigma \leq \|\nabla f(\mathbf{x}_t)\|$。

一致有界方差，定义 1.3。在这种情况下，我们首先假设强方差以常数 3成立，即，即$\|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\| \leq 3\sigma$ 以概率 1 成立。在这种情况下，我们可以写成，其中 $\alpha_\xi = \min(1, c/\|\nabla f_\xi(\mathbf{x}_t)\|)$：

$$
\begin{aligned}
-\nabla f(\mathbf{x}_t)^\top \mathbf{g}(\mathbf{x}_t) &= -\alpha_\xi \|\nabla f(\mathbf{x}_t)\|^2 - \alpha_\xi \nabla f(\mathbf{x}_t)^\top (\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)) \\
&\leq -\alpha_\xi \|\nabla f(\mathbf{x}_t)\|^2 + \alpha_\xi \|\nabla f(\mathbf{x}_t)\| \|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\| \\
&\leq -\alpha_\xi \|\nabla f(\mathbf{x}_t)\|^2 + 3\alpha_\xi \|\nabla f(\mathbf{x}_t)\| \sigma \\
&\leq -\frac{\alpha_\xi}{2} \|\nabla f(\mathbf{x}_t)\|^2 ,
\end{aligned}
$$

其中，最后一行是由 $\sigma < \|\nabla f(\mathbf{x}_t)\|/6$ 这一事实推出的。特别地，根据强方差假设，我们知道 $\|\nabla f_\xi(\mathbf{x}_t)\| \leq 2\|\nabla f(\mathbf{x}_t)\|$，因此 $\alpha_\xi \geq \min(1, c/(2\|\nabla f(\mathbf{x}_t)\|)) \geq c/(2\|\nabla f(\mathbf{x}_t)\|)$。特别地：

$$
-\nabla f(\mathbf{x}_t)^\top \nabla f_\xi(\mathbf{x}_t) \leq -\frac{c}{4} \|\nabla f(\mathbf{x}_t)\| . \qquad (23)
$$

接下来，我们可以将其代入方程（22），从而得到：

$$
\mathbb{E}[f(\mathbf{x}_{t+1})] - f(\mathbf{x}_t) \leq -\frac{\eta c}{4} (1 - 2\eta(L_0 + cL_1)) \|\nabla f(\mathbf{x}_t)\| . \qquad (24)
$$

特别地，取 $\eta \leq (4[L_0 + cL_1])^{-1}$，可得：

$$
\frac{\eta c}{8} \|\nabla f(\mathbf{x}_t)\| \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}). \qquad (25)
$$

期望意义下的有界方差，定义 1.4。在这种情况下，我们无法像之前那样写出以概率 1 成立的相同不等式。不过，我们仍然可以以足够高的概率保证该界。我们定义 $\delta = \mathbb{1}\{\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x})\| > 3\sigma\}$。我们将使用条件期望来表示

$$
\mathbb{E}\left[-\alpha_\xi \nabla f(\mathbf{x})^\top \nabla f_\xi(\mathbf{x})\right] \leq p(\delta = 0) \underbrace{\mathbb{E}\left[-\alpha_\xi \nabla f(\mathbf{x})^\top \nabla f_\xi(\mathbf{x}) | \delta = 0\right]}_{:= T_1} + p(\delta = 1) \underbrace{\mathbb{E}\left[-\alpha_\xi \nabla f(\mathbf{x})^\top \nabla f_\xi(\mathbf{x}) | \delta = 1\right]}_{:= T_2} .
$$

我们对第一项 T_1 采用与前面一致有界噪声情形相同的方法进行界定。对于第二项，根据柯西-施瓦茨不等式，并定义 $\alpha = \min(1, c/\|\nabla f(\mathbf{x})\|)$，可得

$$
T_2 = \mathbb{E}\left[-\alpha_\xi \nabla f(\mathbf{x})^\top \nabla f_\xi(\mathbf{x}) | \delta = 1\right] \leq \|\nabla f(\mathbf{x})\| \, \mathbb{E}\left[\|\alpha_\xi \nabla f_\xi(\mathbf{x})\| \, | \delta = 1\right] \leq \alpha \|\nabla f(\mathbf{x})\|^2 ,
$$

其中，最后一个不等式是因为我们假设全梯度较大 $\|\nabla f(\mathbf{x})\| > 6\sigma$，但裁剪阈值较小 $c \leq 4\sigma$。因此，全梯度总会被裁剪，并且 $\|\alpha \nabla f(\mathbf{x})\| = c \geq \|\alpha_\xi \nabla f_\xi(\mathbf{x})\|$。提醒一下，$\alpha_\xi = \min(1, c/\|\nabla f_\xi(\mathbf{x})\|)$。

现在，只需对 $p(\delta=1)$进行界定。根据马尔可夫不等式，我们有：

$$p(\delta=1)=p(\|\nabla f_\xi(\mathbf{x})-\nabla f(\mathbf{x})\|^2>9\sigma^2)\leq 1/9. \tag{26}$$

同样，$p(\delta=0)=1-p(\delta=1)\geq 8/9$。最终，我们得到：

$$-\mathbb{E}\left[\nabla f(\mathbf{x})^\top \mathbf{g}(\mathbf{x})\right]\leq -c\left(\frac{1}{4}\times\frac{8}{9}-\frac{1}{9}\right)\|\nabla f(\mathbf{x})\|=-\frac{c}{9}\|\nabla f(\mathbf{x})\|\,. \tag{27}$$

我们进一步将结果代入（22），得到

$$\mathbb{E}\left[f(\mathbf{x}_{t+1})\right]-f(\mathbf{x}_t)\leq -\frac{\eta c}{9}\left(1-\frac{9\eta}{2}(L_0+cL_1)\right)\|\nabla f(\mathbf{x}_t)\|\,, \tag{28}$$

因此，利用 $\eta\leq(9[L_0+cL_1])^{-1}$，我们得到：

$$\mathbb{E}\left[f(\mathbf{x}_{t+1})\right]-f(\mathbf{x}_t)\leq -\frac{\eta c}{18}\|\nabla f(\mathbf{x}_t)\|\,. \tag{29}$$

最终收敛。 如果至少在一次迭代中 t 梯度范数较小 $\|\nabla f(\mathbf{x}_t)\|\leq 6\sigma$，那么显然有

$$\min_{t\in[1,T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\leq O\left(\sigma^2\right).$$

否则，对于所有 t 迭代，梯度范数都较大 $\|\nabla f(\mathbf{x}_t)\|>6\sigma$ 因此（29）对所有迭代均成立。对 $1\leq t\leq T+1$取平均，得到

$$\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla f(\mathbf{x}_t)\|\leq \mathcal{O}\left(\frac{f(\mathbf{x}_0)-f^\star}{\eta cT}\right), \tag{30}$$

结合这两种情况，我们得出

$$\min_{t\in[1,T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\leq \mathcal{O}\left(\sigma^2+\frac{f(\mathbf{x}_0)-f^\star}{\eta cT}\right),$$

C.3. 证明的第二部分：收敛至 σ^2/c（较大的 c）。

在第二部分中，我们假设裁剪半径较大，$c\geq 4\sigma$。尽管算法（2）会裁剪随机梯度 $\nabla f_\xi(\mathbf{x}_t)$，但在证明中，我们将根据全梯度 $\nabla f(\mathbf{x}_t)$ 考虑两种情况：全梯度$\nabla f(\mathbf{x}_t)$ 被裁剪，以及未被裁剪。

与前一种情况类似，我们首先使用（L_0，L_1）光滑性

$$f(\mathbf{x}_{t+1})-f(\mathbf{x}_t)\leq -\eta\nabla f(\mathbf{x}_t)^\top\mathbf{g}(\mathbf{x}_t)+\frac{\eta^2(L_0+\|\nabla f(\mathbf{x}_t)\|L_1)}{2}\|\mathbf{g}(\mathbf{x}_t)\|^2\,. \tag{31}$$

第一种情况，全梯度被裁剪 $\|\nabla f(\mathbf{x}_t)\|>c$。

在这种情况下，我们使用（21）和 $\alpha=\frac{c}{\|\nabla f(\mathbf{x}_t)\|}$ 以及 $\mathbf{u}=\mathbf{g}(\mathbf{x}_t)$。由于 $\alpha\nabla f(\mathbf{x}_t)=\text{裁剪}_c(\nabla f(\mathbf{x}_t))$，这导致

$$-\nabla f(\mathbf{x}_t)^\top\mathbf{g}(\mathbf{x}_t)=-\frac{c}{2}\|\nabla f(\mathbf{x}_t)\|-\frac{1}{2\alpha}\|\mathbf{g}(\mathbf{x}_t)\|^2+\frac{1}{2\alpha}\|\mathbf{g}(\mathbf{x}_t)-\text{clip}_c(\nabla f(\mathbf{x}_t))\|^2\,. \tag{32}$$

现在我们使用$\mathbf{g}(\mathbf{x}_t)=\text{裁剪}_c(\nabla f_\xi(\mathbf{x}_t))$，并利用裁剪是到一个凸集（半径为 c 的球）上的投影这一事实，因此它是 Lipschitz 常数为 1 的 Lipschitz 算子，我们写作

$$\begin{aligned}
-\nabla f(\mathbf{x}_t)^\top\mathbb{E}\mathbf{g}(\mathbf{x}_t)&\leq -\frac{c}{2}\|\nabla f(\mathbf{x}_t)\|-\frac{1}{2\alpha}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}_t)\|^2\right]+\frac{1}{2\alpha}\mathbb{E}\|\nabla f_\xi(\mathbf{x}_t)-\nabla f(\mathbf{x}_t)\|^2\\
&\leq -\frac{c}{2}\|\nabla f(\mathbf{x}_t)\|-\frac{1}{2\alpha}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}_t)\|^2\right]+\frac{\sigma^2}{2c}\|\nabla f(\mathbf{x}_t)\|\\
&=-\frac{1}{2\alpha}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}_t)\|^2\right]-\frac{c}{2}\|\nabla f(\mathbf{x}_t)\|\left(1-\frac{\sigma^2}{c^2}\right)\\
&\leq -\frac{\|\nabla f(\mathbf{x}_t)\|}{2c}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}_t)\|^2\right]-\frac{c}{4}\|\nabla f(\mathbf{x}_t)\|\,,
\end{aligned}$$

其中，最后一行用到了 $\sigma^2/c^2 \le 1/2$ 且 $\alpha \le 1$。将其代入（31），得到

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1})\right] - f(\mathbf{x}_t) &\le -\frac{\eta\left\|\nabla f(\mathbf{x}_t)\right\|}{2c}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] - \frac{\eta c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| + \frac{\eta^2(L_0 + \left\|\nabla f(\mathbf{x}_t)\right\| L_1)}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] \\ &= -\frac{\eta c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| - \frac{\eta\left\|\nabla f(\mathbf{x}_t)\right\|}{2c}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right](1 - \eta c L_1) + \frac{\eta^2 L_0}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] \\ &\le -\frac{\eta c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| - \frac{\eta}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right](1 - \eta c L_1) + \frac{\eta^2 L_0}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] \\ &= -\frac{\eta c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| - \frac{\eta}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right](1 - \eta[L_0 + c L_1])\,.\end{aligned}$$

特别地，取 $\eta \le (L_0 + cL_1)^{-1}$，可得：

$$\frac{c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| \le \frac{f(\mathbf{x}_t) - \mathbb{E}f(\mathbf{x}_{t+1})}{\eta}\,. \tag{33}$$

请注意，我们没有得到方差项；但与上一节类似，这是因为我们假设梯度范数大于 σ，因此噪声项可以并入梯度范数项中。

第二种情况， $c > \left\|\nabla f(\mathbf{x}_t)\right\| > c/2$。 证明与前一种情况非常相似，区别在于全梯度 $\nabla f(\mathbf{x}_t)$ 没有被裁剪。我们使用带有21的方程（ $\alpha = 1$。由此可得

$$\begin{aligned}-\nabla f(\mathbf{x}_t)^\top \mathbb{E}\mathbf{g}(\mathbf{x}_t) &= -\frac{1}{2}\left\|\nabla f(\mathbf{x}_t)\right\|^2 - \frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 + \frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\right\|^2 \\ &\le -\frac{1}{2}\left\|\nabla f(\mathbf{x}_t)\right\|^2 - \frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 + \frac{\sigma^2}{2}\,,\end{aligned}$$

其中，最后一行用到裁剪是常数为 1 的 Lipschitz 算子，因为它是到凸集上的投影。现在，对于第一项，我们使用 $-\left\|\nabla f(\mathbf{x}_t)\right\| \le -c/2$；对于最后一项，使用 $1 \le \left\|\nabla f(\mathbf{x}_t)\right\|/c$：

$$\begin{aligned}-\nabla f(\mathbf{x}_t)^\top \mathbb{E}\mathbf{g}(\mathbf{x}_t) &\le -\frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 - \frac{c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| + \frac{\sigma^2}{2c}\left\|\nabla f(\mathbf{x}_t)\right\| \\ &\le -\frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 - \frac{c}{4}\left\|\nabla f(\mathbf{x}_t)\right\|\left(1 - 2\frac{\sigma^2}{c^2}\right) \\ &\le -\frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 - \frac{c}{8}\left\|\nabla f(\mathbf{x}_t)\right\|,\end{aligned}$$

其中在最后一行中，我们使用了 $\sigma^2/c^2 \le 1/4$。与前一种情况类似，我们将其代入（31）并利用 $-1 \le -\frac{\left\|\nabla f(\mathbf{x}_t)\right\|}{c}$ 可得

$$\begin{aligned}\mathbb{E}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\le -\frac{\eta\left\|\nabla f(\mathbf{x}_t)\right\|}{2c}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] - \frac{c\eta}{8}\left\|\nabla f(\mathbf{x}_t)\right\| + \frac{\eta^2(L_0 + \left\|\nabla f(\mathbf{x}_t)\right\| L_1)}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] \\ &\le -\frac{c\eta}{8}\left\|\nabla f(\mathbf{x}_t)\right\| - \frac{\eta}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right]\left(\frac{\left\|\nabla f(\mathbf{x}_t)\right\|}{c}(1 - \eta c L_1) - \eta L_0\right) \\ &\le -\frac{c\eta}{8}\left\|\nabla f(\mathbf{x}_t)\right\| - \frac{\eta}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right]\left(\frac{1}{2} - \eta[L_0 + cL_1]\right)\end{aligned}$$

其中在最后一行中，我们使用了 $\left\|\nabla f(\mathbf{x}_t)\right\| > c/2$。利用 $\eta \le \frac{1}{2}(L_0 + cL_1)^{-1}$

$$\frac{c}{8}\left\|\nabla f(\mathbf{x}_t)\right\| \le \frac{f(\mathbf{x}_t) - \mathbb{E}f(\mathbf{x}_{t+1})}{\eta}\,. \tag{34}$$

第三种情况， $\left\|\nabla f(\mathbf{x}_t)\right\| < c/2$。在这种情况下，我们不会收敛到精确最优解 .

我们首先定义 $\delta_t = \mathbb{1}\{\left\|\nabla f_\xi(\mathbf{x}_t)\right\| > c\}$ 为指示函数，用于表示在时间步 t 随机梯度正在被裁剪。接下来我们将证明 $\mathbb{E}\delta_t \le \frac{4\sigma^2}{c^2}$。

$$\mathbb{E}\delta_t = \Pr[\delta_t = 1] = \Pr\left[\left\|\nabla f_\xi(\mathbf{x}_t)\right\| > c\right] \le \Pr\left[\left\|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\right\| > \frac{c}{2}\right] \le \frac{4\sigma^2}{c^2}\,,$$

其中，最后一个不等式由马尔可夫不等式得出。第一个不等式是因为 $\|\nabla f_\xi(\mathbf{x}_t)\| \le \|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\| + \|\nabla f(\mathbf{x}_t)\| \le \|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\| + \frac{c}{2}$。

现在我们有了 $\mathbb{E}\delta_t \le \frac{4\sigma^2}{c^2}$，可以用它来界定差值 $\|\nabla f(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2$。具体而言，由于 δ_t 取值为 0 或 1，我们有 $\mathbb{E}[\delta_t] = p(\delta_t = 1)$，因此对任意随机变量 X，都有 $\mathbb{E}[\delta_t X] = \mathbb{E}[\delta_t]\,\mathbb{E}[X|\delta_t]$。

$$\|\nabla f(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 = \left\|\mathbb{E}\left(1 - \frac{c}{\|\nabla f_\xi(\mathbf{x}_t)\|}\right)\nabla f_\xi(\mathbf{x}_t)\delta_t\right\|^2 = \mathbb{E}[\delta_t]^2\left\|\mathbb{E}\left[\left(1 - \frac{c}{\|\nabla f_\xi(\mathbf{x}_t)\|}\right)\nabla f_\xi(\mathbf{x}_t)|\delta_t = 1\right]\right\|^2.$$

此时，我们对条件期望应用詹森不等式（因为所有项均为正，且平方范数是一个凸函数），得到：

$$\begin{aligned}
\|\nabla f(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 &\le \mathbb{E}[\delta_t]^2\,\mathbb{E}\left[\left(1 - \frac{c}{\|\nabla f_\xi(\mathbf{x}_t)\|}\right)^2\|\nabla f_\xi(\mathbf{x}_t)\|^2|\delta_t = 1\right] \\
&\le \mathbb{E}[\delta_t]^2\,\mathbb{E}\left[\|\nabla f_\xi(\mathbf{x}_t)\|^2|\delta_t = 1\right] \\
&\le 2\mathbb{E}[\delta_t]^2\,\mathbb{E}\left[\|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2|\delta_t = 1\right] + 2\mathbb{E}[\delta_t]^2\,\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2|\delta_t = 1\right] \\
&\le 2\mathbb{E}[\delta_t]\,\mathbb{E}\left[\|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2\right] + 2\mathbb{E}[\delta_t]^2\|\nabla f(\mathbf{x}_t)\|^2 \\
&\le \frac{8\sigma^4}{c^2} + \frac{32\sigma^4}{c^4}\|\nabla f(\mathbf{x}_t)\|^2, \qquad (35)
\end{aligned}$$

其中，在第二行中，我们利用了 $\left(1 - \frac{c}{\|\nabla f_\xi(\mathbf{x}_t)\|}\right)^2 \le 1$ 当 $\delta_t = 1$时成立这一点，而在最后一行中，我们利用了 $\mathbb{E}[\delta_t] \le 4\sigma^2/c^2$。此外，结合（21）以及 $\alpha = 1$和 $\mathbf{g}(\mathbf{x}_t)$，我们得到

$$\begin{aligned}
-\nabla f(\mathbf{x})^\top\mathbb{E}\mathbf{g}(\mathbf{x}_t) &= -\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 - \frac{1}{2}\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 + \frac{1}{2}\|\mathbb{E}\mathbf{g}(\mathbf{x}_t) - \nabla f(\mathbf{x})\|^2 \\
&\le -\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 - \frac{1}{2}\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 + \frac{4\sigma^4}{c^2} + \frac{4\sigma^2}{c^2}\|\nabla f(\mathbf{x}_t)\|^2 \\
&\overset{\sigma\le\frac{c}{4}}{\le} -\frac{1}{4}\|\nabla f(\mathbf{x})\|^2 - \frac{1}{2}\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 + \frac{4\sigma^4}{c^2}.
\end{aligned}$$

将其代入（31）中，对于 $\eta \le \frac{1}{8(L_0 + cL_1)}$，去掉 $\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2$ 项，并利用 $\|\nabla f(\mathbf{x}_t)\| \le c$，可得：

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\le -\frac{\eta}{4}\|\nabla f(\mathbf{x})\|^2 - \frac{\eta}{2}\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 + \frac{4\eta\sigma^4}{c^2} + \frac{\eta^2(L_0 + \|\nabla f(\mathbf{x}_t)\|L_1)}{2}\mathbb{E}\|\mathbf{g}(\mathbf{x}_t)\|^2 \qquad (36) \\
&\le -\frac{\eta}{4}\|\nabla f(\mathbf{x})\|^2 + \frac{4\eta\sigma^4}{c^2} + \frac{\eta^2(L_0 + cL_1)}{2}\mathbb{E}\|\mathbf{g}(\mathbf{x}_t) - \nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)\|^2 \\
&\le -\frac{\eta}{4}\|\nabla f(\mathbf{x})\|^2 + \frac{4\eta\sigma^4}{c^2} + \eta^2(L_0 + cL_1)\mathbb{E}\|\mathbf{g}(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2 + \eta^2(L_0 + cL_1)\|\nabla f(\mathbf{x}_t)\|^2 \\
&\overset{\eta\le\frac{1}{8(L_0+cL_1)}}{\le} -\frac{\eta}{8}\|\nabla f(\mathbf{x})\|^2 + \frac{4\eta\sigma^4}{c^2} + \eta^2(L_0 + cL_1)\mathbb{E}\|\mathbf{g}(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2.
\end{aligned}$$

注意，裁剪是投影到半径为 c 的球上的正交投影，我们将其记为 proj_c 和 $\|\nabla f(x_t)\| \le c$，因此它不受该投影的影响。特别地：

$$\mathbb{E}\|g(x_t) - \nabla f(x_t)\|^2 = \mathbb{E}\|\mathrm{proj}_c(\nabla f_\xi(x_t)) - \mathrm{proj}_c(\nabla f(x_t))\|^2 \le \mathbb{E}\|\nabla f_\xi(x_t) - \nabla f(x_t)\|^2 \le \sigma^2, \qquad (37)$$

因此我们得到

$$\mathbb{E}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \le -\frac{\eta}{8}\|\nabla f(\mathbf{x})\|^2 + \frac{4\eta\sigma^4}{c^2} + \eta^2(L_0 + cL_1)\sigma^2, \qquad (38)$$

因此：

$$\frac{1}{8}\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{f(\mathbf{x}_t) - \mathbb{E}f(\mathbf{x}_{t+1})}{\eta} + \eta(L_0 + cL_1)\sigma^2 + \frac{4\sigma^4}{c^2}. \qquad (39)$$

具体而言，我们有：

- 一个会随步长减小而逐渐消失的方差项。
- 即使步长非常小，仍然存在的一个偏差项。

总结一下。 现在我们将上述三种情况结合起来。定义 $\mathcal{T}_1$ 为满足以下条件的指标集合 $\|\nabla f(\mathbf{x}_t)\| \geq \frac{c}{2}$(注意，这涵盖了上述第一种和第二种情况，但两者导出的最终不等式相同（34）），而 $\mathcal{T}_2$ 为满足以下条件的指标集合 $\|\nabla f(\mathbf{x}_t)\| < \frac{c}{2}$，该不等式（39）成立。对所有指标 $1 \leq t \leq T+1$求和，得到

$$\frac{1}{8(T+1)}\left(\sum_{t\in\mathcal{T}_1} c\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| + \sum_{t\in\mathcal{T}_2}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2\right) \leq \frac{f(\mathbf{x}_0)-f^\star}{\eta(T+1)} + \eta(L_0+cL_1)\sigma^2 + \frac{4\sigma^4}{c^2}\,.$$

这意味着 (i)

$$\frac{1}{8(T+1)}\sum_{t\in\mathcal{T}_1} c\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \frac{f(\mathbf{x}_0)-f^\star}{\eta(T+1)} + \eta(L_0+cL_1)\sigma^2 + \frac{4\sigma^4}{c^2}\,,$$

以及 (ii)

$$\frac{1}{8(T+1)}\sum_{t\in\mathcal{T}_2} \mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq \frac{f(\mathbf{x}_0)-f^\star}{\eta(T+1)} + \eta(L_0+cL_1)\sigma^2 + \frac{4\sigma^4}{c^2}\,,$$

最后一个不等式利用了这一点： $x^2 \geq 2\epsilon x - \epsilon^2$ 对于任意 ϵ， $x>0$，并为简化起见定义 $A := \frac{8f(\mathbf{x}_0)-f^\star}{\eta T} + 8\eta(L_0+cL_1)\sigma^2 + \frac{32\sigma^4}{c^2}$ 我们得到

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_2}\left(2\epsilon\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| - \epsilon^2\right) \leq A\,,$$

因此，

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_2}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \frac{A}{2\epsilon} + \frac{\epsilon}{2}\,.$$

选择$\epsilon = \sqrt{A}$，我们得到

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_2}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \sqrt{A} \leq \sqrt{8\frac{f(\mathbf{x}_0)-f^\star}{\eta(T+1)}} + \sqrt{8\eta(L_0+cL_1)\sigma^2} + \sqrt{\frac{32\sigma^4}{c^2}}\,.$$

再次总结这两种情况，并利用 $\frac{\sigma}{c} \leq \frac{1}{4}$ 我们得到

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \mathcal{O}\left(\sqrt{\frac{f(\mathbf{x}_0)-f^\star}{\eta T}} + \frac{f(\mathbf{x}_0)-f^\star}{\eta c T} + \sqrt{\eta(L_0+cL_1)}\sigma + \frac{\sigma^2}{c}\right)\,.$$

C.4. 差分隐私随机梯度下降

C.4.1. 纳入小批量的证明修改

使用 $\mathbf{g}(\mathbf{x}_t) = \frac{1}{B}\sum_{\xi\in\mathcal{B}_t}$裁剪$_c(\nabla f_\xi(\mathbf{x}_t))$，证明与前一种情况完全相同，唯一的区别在于 $c \geq 4\sigma$和小梯度（第三种情况） $\|\nabla f(\mathbf{x}_t)\| < \frac{c}{2}$。从方程（36）出发，我们得到：

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq -\frac{\eta}{4}\left\|\nabla f(\mathbf{x})\right\|^2 - \frac{\eta}{2}\left\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 + \frac{4\eta\sigma^4}{c^2} + \frac{\eta^2(L_0+\|\nabla f(\mathbf{x}_t)\|L_1)}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 \\
&\leq -\frac{\eta}{4}\left\|\nabla f(\mathbf{x})\right\|^2 - \frac{\eta}{2}\left\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 + \frac{4\eta\sigma^4}{c^2} + \frac{\eta^2(L_0+cL_1)}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 \\
&\quad + \frac{\eta^2(L_0+cL_1)}{2}\left\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2\,.
\end{aligned}$$

现在，我们对方差项 $\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2$ 进行更精确的估计，以获得批大小 B带来的方差降低。

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 &= \mathbb{E}\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\mathrm{clip}_c(\nabla f_{\xi_i}(\mathbf{x}_t)) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 = \frac{1}{B^2}\sum_{i\in\mathcal{B}_t}\mathbb{E}\left\|\mathrm{clip}_c(\nabla f_{\xi_i}(\mathbf{x}_t)) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 \\
&\leq \frac{1}{B^2}\sum_{i\in\mathcal{B}_t} 2\mathbb{E}\left\|\mathrm{clip}_c(\nabla f_{\xi_i}(\mathbf{x}_t)) - \nabla f(\mathbf{x}_t)\right\|^2 + \frac{2}{B}\left\|\nabla f(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 \\
&\overset{(35)}{\leq} \frac{2\sigma^2}{B} + \frac{2}{B}\left[\frac{8\sigma^4}{c^2} + \frac{8\sigma^2}{c^2}\left\|\nabla f(\mathbf{x}_t)\right\|^2\right] \\
&\leq \frac{2\sigma^2}{B} + \frac{2}{B}\left[\frac{\sigma^2}{2} + 2\sigma^2\right] \leq 6\frac{\sigma^2}{B},
\end{aligned}$$

其中我们用到了 $\|\nabla f(\mathbf{x}_t)\| \leq \frac{c}{2}$以及 $\sigma \leq \frac{c}{4}$。证明的其余部分与之前完全相同，只需将 σ^2 项替换为 $\frac{\sigma^2}{B}$，即可得到收敛率 —

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \mathcal{O}\left(\sqrt{\frac{f(\mathbf{x}_0) - f^\star}{\eta T}} + \frac{f(\mathbf{x}_0) - f^\star}{\eta c T} + \sqrt{\eta(L_0 + cL_1)}\frac{\sigma}{\sqrt{B}} + \frac{\sigma^2}{c}\right).$$

C.4.2. 纳入随机噪声的证明修改

差分隐私随机梯度下降（15）中使用的梯度形式为 $\mathbf{g}(\mathbf{x}_t) + \mathbf{z}_t$，其中 $\mathbf{z}_t$ 是方差为 σ_{DP} 的高斯噪声。为了加入这项额外的高斯噪声，我们需要修改证明的第一步，即使用 $(L_0,\ L_1)$ 光滑性

$$\mathbb{E}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\eta\nabla f(\mathbf{x}_t)^\top\mathbf{g}(\mathbf{x}_t) + \frac{\eta^2(L_0 + \|\nabla f(\mathbf{x}_t)\|\, L_1)}{2}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 + \frac{\eta^2(L_0 + \|\nabla f(\mathbf{x}_t)\|\, L_1)}{2}\sigma_{\mathrm{DP}}^2.$$

证明的其余部分保持不变，只是在收敛界中多出一项 σ_{DP}^2。因此，我们可以得到以下收敛率。为简便起见，定义 $L = L_0 + \max_t \|\nabla f(\mathbf{x}_t)\|\, L_1$

$$O\left(\frac{L\eta}{c}\sigma_{\mathrm{DP}}^2 + \sqrt{L\eta\sigma_{\mathrm{DP}}} + \min\left(\sigma, \frac{\sigma^2}{c}\right) + \sqrt{\eta L}\frac{\sigma}{\sqrt{B}} + \sqrt{\frac{F_0}{\eta T}} + \frac{F_0}{\eta T c}\right).$$

C.5. 下界

现在我们证明下界。

定理 *3.1*和*3.2*的证明。 考虑简单的噪声 $a\mathcal{B}(p)$，其中 $a > 0$ 且 $\mathcal{B}(p)$ 是均值为 $p \leq 1/2$的伯努利随机变量。考虑一个函数，使得随机梯度具有如下形式：

$$\nabla f_\xi(x) = x + a\mathcal{B}(p). \tag{40}$$

现在考虑 $x = -pc/(1-p)$。当伯努利随机变量为 0 时，x 处的随机梯度不会被裁剪，因为 $|x| = pc/(1-p) \leq c$。然而，当伯努利随机变量取正值时，随机梯度为：

$$\nabla f_a(x) = -pc/(1-p) + a \geq a - c \geq c. \tag{41}$$

特别地，我们有：

$$\mathbb{E}\left[\mathrm{clip}_c(\nabla f_\xi(x))\right] = (1-p)x + pc = (1-p)\times(-pc)/(1-p) + pc = 0. \tag{42}$$

现在让我们评估 $\nabla f(x)$。我们有：

$$\nabla f(x) = x + pa = p\left(a - \frac{c}{1-p}\right). \tag{43}$$

较小的 c。现在固定裁剪半径 c，使得 $c \leq 2\sigma$，并取 $a = 4\sigma$。我们选择 $p(1-p) = 1/16$，使得$p = (2-\sqrt{3})/4 \leq 1/4$。在这种情况下，

$$\nabla f(x) = p\left(a - \frac{c}{1-p}\right) \geq p\left(4\sigma - 2\sigma \times \frac{4}{3}\right) \geq \frac{(2-\sqrt{3})\sigma}{3} \geq \frac{\sigma}{12}. \tag{44}$$

较大 c。现在固定一个裁剪半径 c，使得 $c \geq \sigma$ 和 $c \leq a/2$。为确保噪声的方差为 σ^2，p 必须满足：

$$p(1-p) = \sigma^2/a^2 \leq 1/16\,. \tag{45}$$

因此，我们有 $p \leq 1/4$（因为我们选择了 $p < 1/2$）。特别地，再利用 $c \leq a/2$，可得：

$$\nabla f(x) = \frac{\sigma^2}{a^2(1-p)}\left(a - \frac{c}{1-p}\right) \geq \frac{\sigma^2}{3a(1-p)} \geq \frac{\sigma^2}{3a}\,. \tag{46}$$

现在只需选择 $a = 2c$（它满足此前的所有条件），即可得到：

$$\nabla f(x) \geq \frac{\sigma^2}{6c}\,. \tag{47}$$

□