arXiv:2304.01119v2 [math.OC] 4 Apr 2023

重尾情形下梯度裁剪方法的高概率改进的收敛性

Ta Duy Nguyen* Alina Ene† Huy L. Nguyen‡

摘要

在本文中，我们研究噪声分布具有重尾特性的梯度裁剪方法的高概率 收敛性，即其 p阶矩有界，其中某个 $1 < p \leq 2$成立。在这种设置下，先前的工作采用相同的做法：使用集中不等式和结合并合界的归纳论证，对所有迭代中的迭代点进行界定。这种方法会使失败概率增加 T倍，其中 T 是迭代次数。相较之下，我们提出了一种新的分析方法，其基础是对精心选择的超鞅序列的矩母函数进行界定。对于广泛的采用裁剪梯度的算法，我们改进了收敛保证中对 T的依赖性，包括适用于凸目标函数的随机（加速）镜像下降，以及适用于非凸目标函数的随机梯度下降。当时间范围未知时，该方法自然允许算法使用时变步长和裁剪参数，而这在先前的工作中似乎是不可能的。我们证明，在裁剪随机镜像下降的情形下，设置步长和裁剪参数时无需使用问题常数，包括到最优解的初始距离。

1 引言

随机优化是一个研究充分的领域，在机器学习、运筹学、数值线性代数等众多领域都有广泛应用。与确定性算法不同，随机算法可能会失败，因此一个相关的问题是了解这种情况发生的频率，以及如何通过改进来提高成功率。这个问题在不允许失败的关键应用中尤为重要；对于许多机器学习应用而言也至关重要，因为每次运行都非常昂贵且耗时。幸运的是，研究表明，在高斯分布等轻尾噪声分布下，标准随机梯度下降（SGD）算法能够以高概率收敛 [23, 12, 28, 14, 11, 10]。然而，深度学习应用中的近期观察揭示了更具挑战性的优化景观，因此需要对算法进行新的改进[31, 30, 34, 9]。

在这项工作中，我们考虑 [34] 提出的重尾噪声一般模型。在该模型中，梯度噪声是无偏的，并且其 p 阶矩有界，其中 $p \in (1, 2]$。尽管 SGD 在这种情形下可能无法收敛，但 [34] 表明，采用适当裁剪的 SGD（裁剪随机梯度下降）在期望意义下能够收敛。近期的后续研究 [24, 19, 29] 表明，该算法实际上能够以高概率收敛。这一结果令人欣喜，将 [7] 针对 $p = 2$ 的早期工作向前推进了一步。然而，与轻尾情形下相应的界相比，这些结果仍存在若干不足。首先，裁剪算法采用固定的步长和固定的裁剪参数，且二者都取决于迭代次数，因此无法处理未知时间范围的情形。其次，其收敛保证比轻尾情形下的界差一个 $\log T$ 因子。这些问题引出了一个定性问题：

重尾噪声是否天生比轻尾噪声更难处理？

在本研究中，我们通过为裁剪随机梯度下降提供一种改进且通用的分析框架，回答上述问题。我们的分析允许步长和裁剪参数随时间变化，并能得到更紧的

*波士顿大学计算机科学系，taduy@bu.edu。
†波士顿大学计算机科学系，aene@bu.edu。
‡东北大学Khoury计算机与信息科学学院，hu.nguyen@northeastern.edu。

界，与轻尾噪声情形下的界相匹配。该分析框架也适用于不同场景：既可利用镜像下降在任意大的定义域上寻找凸函数的极小化点，也可利用梯度下降寻找非凸函数的一阶驻点。

1.1 贡献

我们的工作贡献如下：

首先，我们提出了一种分析梯度裁剪方法高概率收敛的新方法。该方法具有通用性，适用于多种算法，包括凸情形下的裁剪随机镜像下降和随机加速镜像下降裁剪方法，以及非凸情形下的裁剪随机梯度下降。我们的方法基于 [18] 提出的“白盒”集中不等式思想。尽管此前的工作仅适用于轻尾噪声，但我们提出了新的要素，使其能够分析重尾噪声，即具有有界 p 阶矩的噪声。

其次，借助这一新的分析框架，我们为凸和非凸目标函数都给出了时间最优的收敛保证。在凸情形下，对于任意定义域以及用于度量距离的 Bregman 散度，我们的收敛速率为 $O\left(T^{\frac{1-p}{p}}\right)$。在非凸情形下，收敛速率为 $O\left(T^{\frac{2-2p}{3p-2}}\right)$。这两种速率都与 [27, 32, 34] 中证明的下界相匹配。

最后，我们表明，梯度裁剪方法现有的收敛速率可以进一步收紧并得到显著改善。具体而言，我们减少了以往分析不够紧致所造成的 $\log T$ 因子损失。我们的框架可直接扩展至时间范围未知的情形，并允许使用时变步长和裁剪参数，而这在先前的工作中似乎无法实现。此外，仅需非常温和的假设，我们就能为随机镜像下降给出不依赖于问题参数的裁剪参数和步长选择方案；这些问题参数包括噪声方差的上界，以及初始解与最优解之间的初始距离，而它们在实践中通常是未知的。

1.2 相关工作

轻尾噪声下的高概率收敛 已有多项先前的工作证明了次高斯噪声下随机梯度算法的高概率收敛，其中包括 [23, 12, 28, 14, 11, 10] 针对具有有界域（或有界 Bregman 直径）或具有强凸性的凸问题所开展的研究。 [17, 20, 16] 的其他工作研究了随机梯度下降的若干变体在非凸目标函数下的收敛性。这些工作考虑的分布比仅具有有界方差的分布尾部更轻，涵盖了从次高斯噪声到次 Weibull 噪声的各种情形。轻尾噪声不需要梯度裁剪，且通常比重尾噪声更易于分析；对于后者，裁剪梯度所引入的偏差会在算法中传播，需要谨慎处理。这一研究方向中与我们工作最相关的是 [18]的工作。该工作提出了一种白盒方法，用于分析凸目标函数下的随机（加速）镜像下降和非凸函数下的随机梯度下降。在凸情形下，该工作将函数值差以及迭代点与最优解之间的距离一并纳入一个适当定义的鞅中，从而能够利用对矩母函数进行界定的机制，得到收敛速率的紧界。我们的工作基于类似的思想，但为应对尾部更重的噪声而作出了重大调整。特别地，我们提出了一种新颖的鞅序列系数选择方式，使系数依赖于历史信息，而非像 [18]中的系数那样独立于过去的信息。这一思想具有通用性，既适用于随机（加速）镜像下降（针对凸函数），也适用于随机梯度下降（针对非凸函数）。

有界方差和重尾噪声下的高概率收敛 在重尾噪声存在的情况下设计新的梯度算法并对其进行分析，近年来引起了广泛关注。从 [26] 提出裁剪随机梯度下降（Clipped-SGD）的工作开始， [31, 30, 34, 9] 的近期研究根据梯度噪声在实践中的表现提出了新的研究动机，并指出只有裁剪随机梯度下降（Clipped-SGD）可能有效。重尾噪声被定义为具有有界 p 阶矩的噪声，其中 $1 < p \leq 2$。尤其是在各种深度学习任务中，梯度呈现无限方差（$p < 2$）。正如 [34], 所示，这是个问题，因为在这种情况下，普通随机梯度下降可能无法收敛。

尽管普通随机梯度下降的期望收敛已得到广泛研究，例如 [5, 23, 13],，但直到最近，研究者才开始仔细考察重尾噪声下裁剪随机梯度下降（Clipped-SGD）的收敛性。 [34] 展示了非凸函数下裁剪随机梯度下降的期望收敛性，并证明了收敛速率的下界。 [22, 25, 7] 对有界方差噪声采用不同的裁剪方法，证明了有界定义域（第一种和第二种情况）或无约束情形（第三种情况）下，凸光滑问题的高概率收敛性。随后，[8] 使用类似方法将分析扩展到非光滑函数情形。 [2] 提出了一种融入动量的裁剪随机梯度下降变体，并证明了在具有有界 p 阶矩的噪声下，该方法能够高概率收敛。在这项工作中，动量对于保证算法达到最优收敛至关重要。此外，分析还相当严格地要求随机梯度有界。 [24, 19, 29, 35] 最近的研究也聚焦于这类噪声，并为若干裁剪随机梯度下降变体给出了（近乎）最优的收敛速率。

我们的论文沿着研究重尾噪声下高概率收敛的这条路线，在几个关键方面推进了当前最先进的成果。与 [1], 相比，我们的分析不需要使用动量；我们证明，只要恰当选择裁剪参数和步长，简单的裁剪策略就足以使算法达到最优收敛。有限梯度假设也并非必要。 [35] 关注的是旨在改善对初始距离依赖性的无参数算法，这与我们的研究有显著区别，问题也更加复杂。与 [7] 和 [24, 19, 29], 的工作相比，我们的分析得到的界显著更紧。这些工作都遵循相同的做法：使用集中不等式，具体而言，将 Freedman 型不等式 [4, 3] 作为黑箱工具。其关键技术是对所有迭代点进行归纳界定。这个过程会导致最终收敛速率中出现额外的 $\log T$ 项；换言之，成功概率会从 $1-\delta$ 降至 $1-T\delta$。相比之下，我们采用白盒方法处理该问题，这与针对次高斯噪声的工作 [18] 类似。这样，我们就能一次性界定所有迭代中的迭代点，同时保持相同的成功概率。值得注意的是，这种方法还使我们能够处理时间范围未知的情况，或问题参数（例如噪声上界 σ、失败概率 δ，以及或许更重要的到最优解的初始距离）未知的情况；在先前的这些工作中，这似乎是不可能的。最后，我们的工作还推广到了具有任意范数和定义域的随机（加速）镜像下降。

在相关研究中， [33] 表明，在某些特殊假设下，普通随机梯度下降可以在重尾噪声下实现收敛； [32] 则表明，对于定义在有界域上的强凸目标函数，在镜像映射的特殊选择下，随机镜像下降可以实现期望收敛。 [27, 32] （针对凸情形）和 [34] （针对非凸情形）给出了最优收敛速率的下界。在这两种情形下，我们的方法都能够给出最优的收敛保证。

2 预备知识

我们研究如下问题：$\min_{x\in\mathcal{X}} f(x)$，其中 $f:\ \mathbb{R}^d\to\mathbb{R}$，而 $\mathcal{X}$ 是该问题的定义域。在凸情形下，我们假设 $\mathcal{X}$ 是一个凸集，但不一定是紧集。令 $\|\cdot\|$ 为任意范数，并令$\|\cdot\|_*$ 为其对偶范数。在非凸情形下，我们取 $\mathcal{X}$ 为 $\mathbb{R}^d$，并且仅考虑 ℓ_2 范数。

2.1 假设

我们采用以下假设：

（1）极小化点的存在性：在凸情形下，我们假设 $x^*\in\arg\min_{x\in\mathcal{X}} f(x)$。令$f^*=f(x^*)$。

（1’）有限下界的存在性：在非凸情形下，我们假设 f 有有限下界，即 $f^*:=\inf_{x\in\mathbb{R}^d} f(x) > -\infty$。

（2）无偏估计量：我们假设算法可以通过随机一阶$\widehat{}$阶预言机查询 f，该预言机针对任意 $x\in\mathcal{X}$ 返回 $\nabla f(x)$ 的与历史信息无关的无偏梯度估计量 $\widehat{\nabla} f(x)$。也就是说，在给定历史信息和查询点 x 的条件下，有 $\mathbb{E}\left[\widehat{\nabla} f(x)\mid x\right]=\nabla f(x)$。

(3) 有界 p阶矩噪声：我们假设存在 $\sigma>0$，使得对于某个 $1<p\le 2$，且对于任意 $x\in\mathcal{X}$， $\widehat{\nabla} f(x)$ 满足 $\mathbb{E}\left[\left\|\widehat{\nabla} f(x)-\nabla f(x)\right\|_*^p\mid x\right]\le\sigma^p$.

(4) L-光滑性：我们考虑 L-光滑函数类：对于所有 x， $y \in \mathbb{R}^d$， $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x-y\|$.

2.2 梯度裁剪算子及其记号

我们介绍梯度裁剪算子及其一般性质，以便用于裁剪随机镜像下降（算法 1）和裁剪随机梯度下降（算法3）。令 x_t 为所关注算法在第 t 次迭代得到的解。记 $\nabla f(x_t)$ 为通过查询梯度预言机$\widetilde{}$得到的随机梯度。裁剪梯度估计 $\nabla f(x_t)$ 定义为

$$\widetilde{\nabla} f(x_t) = \min\left\{1, \frac{\lambda_t}{\left\|\widehat{\nabla} f(x_t)\right\|_*}\right\} \widehat{\nabla} f(x_t). \tag{1}$$

其中， λ_t是第 t次迭代使用的裁剪参数。在后续各节中，我们令 $\Delta_t := f(x_t) - f^*$表示在 x_t处的函数值差。令 $\mathcal{F}_t = \sigma\left(\widehat{\nabla} f(x_1) \ldots, \widehat{\nabla} f(x_t)\right)$为自然滤过，并定义以下记号：

$$\theta_t = \widetilde{\nabla} f(x_t) - \nabla f(x_t); \quad \theta_t^u = \widetilde{\nabla} f(x_t) - \mathbb{E}\left[\widetilde{\nabla} f(x_t) \mid \mathcal{F}_{t-1}\right]; \quad \theta_t^b = \mathbb{E}\left[\widetilde{\nabla} f(x_t) \mid \mathcal{F}_{t-1}\right] - \nabla f(x_t).$$

注意， $\theta_t^u + \theta_t^b = \theta_t$. 无论函数 f是否具有凸性，以下引理都给出了这些量的上界。对于特殊的 ℓ_2 范数，这些界可见于先前的工作 [7, 34, 24, 19, 29] 。推广到一般范数的过程类似，本文从略。

引理 2.1。我们有

$$\|\theta_t^u\|_* = \left\|\widetilde{\nabla} f(x_t) - \mathbb{E}\left[\widetilde{\nabla} f(x_t) \mid \mathcal{F}_{t-1}\right]\right\|_* \leq 2\lambda_t \tag{2}$$

此外，如果 $\|\nabla f(x_t)\|_* \leq \frac{\lambda_t}{2}$则

$$\left\|\theta_t^b\right\|_* = \left\|\mathbb{E}\left[\widetilde{\nabla} f(x_t) \mid \mathcal{F}_{t-1}\right] - \nabla f(x_t)\right\|_* \leq 4\sigma^p \lambda_t^{1-p}; \tag{3}$$

$$\mathbb{E}\left[\|\theta_t^u\|_*^2\right] = \mathbb{E}\left[\left\|\widetilde{\nabla} f(x_t) - \mathbb{E}_t\left[\widetilde{\nabla} f(x_t)\right]\right\|_*^2 \mid \mathcal{F}_{t-1}\right] \leq 40\sigma^p \lambda_t^{2-p}. \tag{4}$$

最后，我们陈述一个简单但重要的引理，它给出了零均值有界随机变量的矩母函数的界。证明可见例如 [17]。

引理 2.2。 设 X 是一个满足 $\mathbb{E}[X] = 0$ 且 $|X| \leq R$ 几乎处处成立的随机变量。则对于$0 \leq \lambda \leq \frac{1}{R}$

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(\frac{3}{4}\lambda^2 \mathbb{E}\left[X^2\right]\right).$$

3 裁剪随机镜像下降

本节将介绍并分析裁剪随机镜像下降算法（算法 1）。我们定义 Bregman散度 $D_\psi(x,y) = \psi(x) - \psi(y) - \langle \nabla\psi(y),\, x-y\rangle$，其中 ψ： $\mathbb{R}^d \to \mathbb{R}$ 是关于范数 $\|\cdot\|$ 在 $\mathcal{X}$上的 1-强凸可微函数。为方便起见，我们假设 $\mathrm{dom}\psi = \mathbb{R}^d$，或者更一般地， $\mathcal{X} \subseteq \mathrm{int}\,(\mathrm{dom}\psi)$。算法 1 是将适用于凸函数的裁剪随机梯度下降推广到任意范数的结果。它与标准随机镜像下降算法$\widetilde{}$ $\widehat{}$的唯一差别在于：计算新的迭代点 x_{t+1}时，使用裁剪梯度 $\nabla f(x_t)$代替真实的随机梯度 $\nabla f(x_t)$。

先前的工作（如 [7] ）仅考虑全局最小化器位于 $\mathcal{X}$ 中的情形。本节中的算法不要求这一限制，而只采用来自 [22] 的以下温和假设：

算法 1裁剪随机镜像下降

参数：in　　初始点 x_1、步长 $\{\eta_t\}$、裁剪参数 $\{\lambda_t\}$, ψ 关于 $\|\cdot\|$ 是 1-强凸的

for $t=1$ 到 T do

$\widetilde{\nabla} f(x_t)=\min\left\{1, \frac{\lambda_t}{\left\|\widehat{\nabla} f(x_t)\right\|_*}\right\} \widehat{\nabla} f(x_t)$

$x_{t+1}=\arg\min_{x\in\mathcal{X}}\left\{\eta_t\left\langle\widetilde{\nabla} f(x_t), x\right\rangle+\mathbf{D}_\psi\left(x, x_t\right)\right\}$

(5) 良好梯度估计的存在性：我们假设可以获取一个向量 $g_0\in\mathbb{R}^d$ 和一个常数 $\mu\geq 0$，使得在点 $x_0\in\mathcal{X}$ 处有 $\|g_0-\nabla f(x_0)\|_*\leq\mu\sigma$。 [22] 的研究中也采用了这一假设，该研究表明，对于噪声具有有界方差的情形，裁剪随机镜像下降的一个变体在有界域上高概率收敛。该假设的合理性如下。对于无约束问题，以及我们知道最优解 x^* 也是全局最小化器的约束问题，只需取 $x_0=x^*$、$g_0=0$ 和 $\mu=0$。否则，可以选择某个 $x_0\in\mathcal{X}$ 和一个成功概率 ϵ，并向梯度预言机查询 $O\left(\ln\frac{1}{\epsilon}\right)$ 次，从而获得如下保证（来自 [21]）：$\Pr\left[\|g_0-\nabla f(x_0)\|_*>\mu\sigma\right]\leq\epsilon$，其中 g_0 取为随机梯度的几何平均值。We will 首先在下述定理中给出该算法最终的收敛保证。

定理 3.1. 假设 f 满足假设 *(1)*、*(2)*、*(3)*、*(4)* 和 *(5)*。令 $\gamma=\max\left\{\log\frac{1}{\delta}; 1\right\}$;$R_0=\sqrt{2\mathbf{D}\psi\left(x^*, x_0\right)}$ 以及 $R_1=\sqrt{2\mathbf{D}\psi\left(x^*, x_1\right)}$

1. 对于已知的 T，我们选择 λ_t 和 η_t，使得

$$\lambda_t=\lambda=\max\left\{\left(\frac{26T}{\gamma}\right)^{1/p}\sigma; 2\left(2LR_1+LR_0+\mu\sigma+\|g_0\|_*\right)\right\},\ and$$

$$\eta_t=\eta=\frac{R_1}{24\lambda_t\gamma}=\frac{R_1}{24\gamma}\min\left\{\left(\frac{26T}{\gamma}\right)^{-1/p}\sigma^{-1}; \frac{1}{2}\left(2LR_1+LR_0+\mu\sigma+\|g_0\|_*\right)^{-1}\right\},$$

则以至少 $1-\delta$ 的概率

$$\frac{1}{T}\sum_{t=2}^{T+1}\Delta_t\leq 48R_1\max\left\{26^{\frac{1}{p}}T^{\frac{1-p}{p}}\sigma\gamma^{\frac{p-1}{p}}; 2\left(2LR_1+LR_0+\mu\sigma+\|g_0\|_*\right)T^{-1}\gamma\right\}=O\left(T^{\frac{1-p}{p}}\right).$$

2. 对于未知的 T，我们选择

$$\lambda_t=\max\left\{\left(\frac{52t\left(1+\log t\right)^2}{\gamma}\right)^{1/p}\sigma; 2\left(2LR_1+LR_0+\mu\sigma+\|g_0\|_*\right)\right\},\ and$$

$$\eta_t=\frac{R_1}{24\lambda_t\gamma}=\frac{R_1}{24\gamma}\min\left\{\left(\frac{52t\left(1+\log t\right)^2}{\gamma}\right)^{-1/p}\sigma^{-1}; \frac{1}{2}\left(2LR_1+LR_0+\mu\sigma+\|g_0\|_*\right)^{-1}\right\},$$

则以至少 $1-\delta$ 的概率

$$\frac{1}{T}\sum_{t=2}^{T+1}\Delta_t\leq 48R_1\max\left\{52^{\frac{1}{p}}T^{\frac{1-p}{p}}\left(1+\log T\right)^{\frac{2}{p}}\sigma\gamma^{\frac{p-1}{p}}; 2\left(2LR_1+LR_0+\mu\sigma+\|g_0\|_*\right)T^{-1}\gamma\right\}=\widetilde{O}\left(T^{\frac{1-p}{p}}\right).$$

注释 3.2。该定理表明，第一种情况下的收敛速率为 $O\left(T^{\frac{1-p}{p}}\right)$，第二种情况下为$\widetilde{O}\left(T^{\frac{1-p}{p}}\right)$。如 [27, 32]所示，该速率已知是最优的。上述保证也能适应 σ，即当 $\sigma\to 0$ 时，我们可获得确定性镜像下降的标准 $O\left(T^{-1}\right)$ 收敛速率。

备注 3.3。</style}上述保证中的项 $LR_0+\mu\sigma+\|g_0\|_*$ 源于对 $\|\nabla f(x_0)\|_*$的非精确估计 $\|g_0\|_*$。如果假设全局最优解位于定义域 $\mathcal{X}$内，我们只需选择 $x_0=x^*$，该项便会消失。否则，如果假设已知 $\|\nabla f(x_0)\|_*$的精确值，该项将变为$LR_0+\|\nabla f(x_0)\|_*$。

在深入分析之前，我们先将上述定理与 [7]（针对 $p=2$）和 [24, 29]（针对一般的 $1<p\leq 2$）中关于裁剪随机梯度下降的收敛保证进行比较。在已知时间范围 T 的第一种情况下，与先前的工作相比，定理 3.1 中的收敛结果没有额外的 $\log T$ 项。这一改进源于对鞅差序列更好的集中分析。先前这些工作还有一项限制：它们严格要求已知时间范围，才能设置合适的步长和裁剪参数。这意味着在分析中没有直接的方法来消除这一要求。相比之下，定理 3.1 可以自然地推广到未知 T 的情况，但会多出 $\log T$ 项，这是由于未知时间范围所付出的代价。事实上，我们还可以更进一步，在设置步长和裁剪参数时消除对常数 σ、δ 和 R_1 的要求。我们在定理 3.10 中给出明确表述。

我们将从下面的基本引理开始分析。

引理 3.4。 假设 f 满足假设 *(1)*、*(2)*、*(3)*、*(4)* 和 $\eta_t \leq \frac{1}{4L}$，由算法 *1* 输出的迭代序列 $(x_t)_{t\geq 1}$满足以下条件：

$$\begin{aligned}\eta_t\Delta_{t+1} \leq \mathbf{D}_\psi\left(x^*, x_t\right) - \mathbf{D}_\psi\left(x^*, x_{t+1}\right) + \eta_t\left\langle x^* - x_t, \theta_t^u\right\rangle + \eta_t\left\langle x^* - x_t, \theta_t^b\right\rangle \\ + 2\eta_t^2\left(\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right) + 2\eta_t^2\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right] + 2\eta_t^2\left\|\theta_t^b\right\|_*^2.\end{aligned}$$

Remark 3.5。 在附录中，我们针对 f 满足的情形给出了更一般的表述。 s

$$f(y) - f(x) \leq \langle \nabla f(x), y - x\rangle + G\|y - x\| + \frac{L}{2}\|y - x\|^2, \quad \forall y, x \in \mathcal{X}.$$

利普希茨函数（当 $L=0$ 时）和光滑函数（当 $G=0$ 时）都满足这一条件。该证明由 [15], 推出，详见附录。下面我们仅考虑光滑函数的情形（$G=0$），但利用附录中的一般表述，可以将分析自然地扩展到一般情形。

在引理 3.4 中，我们已经将右端分解为适当的项，从而可以定义一个鞅。这一分解的思想可见于 [7]。在同一项工作中，作者以归纳方式分别分析了两个不同的鞅差序列：$\left(\eta_t\left\langle x^* - x_t, \theta_t^u\right\rangle\right)_{t\geq 1}$和 $\left(2\eta_t^2\left(\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\right)_{t\geq 1}$。这里，他们利用并合界将距离 $\|x^* - x_t\|$ 在所有迭代中进行界定。这使得鞅差序列满足应用 Freedman 不等式所需的有界性条件。然而，由于并合界，成功概率从 $1-\delta$ 降至 $1-T\delta$，这是次优的。我们将深入研究集中不等式背后的机制，以收紧分析。我们首先为 $t\geq 1$定义以下各项：

$$\begin{aligned}Z_t = z_t\left(\eta_t\Delta_{t+1} + \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x^*, x_t\right) - \eta_t\left\langle x^* - x_t, \theta_t^b\right\rangle - 2\eta_t^2\left\|\theta_t^b\right\|_*^2 - 2\eta_t^2\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right) \\ - \left(\frac{3}{8\lambda_t^2} + 24 z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|^2 \mid \mathcal{F}_{t-1}\right]\end{aligned}$$

$$\text{where } z_t = \frac{1}{2\eta_t\lambda_t \max_{i\leq t}\sqrt{2\mathbf{D}_\psi\left(x^*, x_i\right)} + 16Q\eta_t^2\lambda_t^2}$$

对于常数 $Q\geq 1$ 以及

$$S_t = \sum_{i=1}^{t} Z_i$$

我们引入以下引理 3.6，其证明将帮助读者理解本文的主要技术。证明该引理的技术与证明集中不等式（如 Freedman 不等式 [4, 3]）时界定矩母函数的标准方法类似。该引理的主要难点在于找到利用裁剪随机镜像下降结构的方法。在这种情况下，我们必须选择合适的系数 z_t。

我们将这一技术与 [18] 中用于分析带次高斯噪声的随机镜像下降的方法进行比较。两者都基于以“白盒”方式分析鞅差序列这一思路。在

在先前的工作中，由于噪声是轻尾的，系数 z_t 的选择只需取决于问题参数，而不必依赖算法历史。这项工作利用距离 $\mathrm{D}_\psi(x^*, x_t)$ 来吸收分析过程中产生的误差。在我们的情形下，这种方法行不通。要使用引理 2.2 对矩母函数进行界定，我们必须确保 $z_t \le \frac{1}{R}$ 对于作为鞅元素上界的 R 成立。这里关键的新思路是，我们可以选择依赖于过去迭代点的 z_t。这一选择确保满足引理 2.2 的条件。

先前工作 [18] 与引理 3.6 的另一个区别在于，在前者中，矩母函数的界会立即给出式 (5) 右端的一个常数界。这里并非如此。不过，我们建立了一个关系，该关系以概率 $1-\delta$ 对所有时间步都成立。与采用 [7, 24, 29] 中的归纳论证所得到的成功概率 $1-T\delta$ 相比，这是一个改进。现在我们指定 η_t 和 λ_t 的取值。以下引理给出了 η_t 和 λ_t 的一般取值条件，使得时间 T 上具有正确的收敛速率。

引理 3.6. 对于任意 $\delta > 0$，令 $E(\delta)$ 表示如下事件：对所有 $1 \le k \le T$

$$\sum_{t=1}^{k} z_t\eta_t\Delta_{t+1} + z_k\mathbf{D}_\psi(x^*, x_{k+1}) \le z_1\mathbf{D}_\psi(x^*, x_1) + \log\frac{1}{\delta} + \sum_{t=1}^{k} z_t\eta_t\left\langle x^* - x_t, \theta_t^b\right\rangle + 2\sum_{t=1}^{k} z_t\eta_t^2\left\|\theta_t^b\right\|_*^2 + \sum_{t=1}^{k}\left(\left(2z_t\eta_t^2 + \frac{3}{8\lambda_t^2} + 24z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right) \tag{5}$$

则 概率 $[E(\delta)] \ge 1-\delta$。

证明。我们有

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \times \exp\left(\left(\frac{3}{8\lambda_t^2} + 24z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\\
&\overset{(a)}{\le} \mathbb{E}\left[\exp\left(z_t\left(\eta_t\left\langle x^* - x_t, \theta_t^u\right\rangle + 2\eta_t^2\left(\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\right)\right) \mid \mathcal{F}_{t-1}\right]\\
&\overset{(b)}{\le} \exp\left(\mathbb{E}\left[\frac{3}{4}\left(z_t\left(\eta_t\left\langle x^* - x_t, \theta_t^u\right\rangle + 2\eta_t^2\left(\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\right)\right)^2 \mid \mathcal{F}_{t-1}\right]\right)\\
&\overset{(c)}{\le} \exp\left(\left(\frac{3}{2}z_t^2\eta_t^2\left\|x^* - x_t\right\|^2\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right] + 6z_t^2\eta_t^4\mathbb{E}\left[\left\|\theta_t^u\right\|_*^4 \mid \mathcal{F}_{t-1}\right]\right)\right)\\
&\overset{(d)}{\le} \exp\left(\left(\frac{3}{2}z_t^2\eta_t^2\left\|x^* - x_t\right\|^2 + 24z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\\
&\overset{(e)}{\le} \exp\left(\left(\frac{3}{8\lambda_t^2} + 24z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)
\end{aligned}$$

对于 (a)，我们使用引理 3.4。对于 (b)，我们使用引理 2.2。注意，

$$\mathbb{E}\left[\left\langle x^* - x_t, \theta_t^u\right\rangle\right] = \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right] = 0$$

并且由于 $\left\|\theta_t^u\right\|_* \le 2\lambda_t$，我们有

$$\begin{aligned}
&\left|\eta_t\left\langle x^* - x_t, \theta_t^u\right\rangle + 2\eta_t^2\left(\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\right|\\
&\le \eta_t\left\|x^* - x_t\right\|\left\|\theta_t^u\right\|_* + 2\eta_t^2\left(\left\|\theta_t^u\right\|_*^2 + \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)\\
&\le 2\eta_t\lambda_t\left\|x^* - x_t\right\| + 16\eta_t^2\lambda_t^2\\
&\le 2\eta_t\lambda_t\sqrt{2\mathbf{D}_\psi(x^*, x_t)} + 16\eta_t^2\lambda_t^2
\end{aligned}$$

因此 $z_t \leq \frac{1}{2\eta_t\lambda_t\sqrt{2\mathbf{D}_\psi\left(x^*,x_t\right)}+16\eta_t^2\lambda_t^2}$。对于 (c)，我们使用不等式 $(a+b)^2 \leq 2a^2+2b^2$ 和 $\mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]\right)^2\right] \leq \mathbb{E}\left[X^2\right]$。对于 (e)，我们利用这一事实：$\left\|\theta_t^u\right\|_* \leq 2\lambda_t$ 并且

$$z_t\eta_t\left\|x^*-x_t\right\| \leq \frac{\eta_t\left\|x^*-x_t\right\|}{2\eta_t\lambda_t\sqrt{2\mathbf{D}_\psi\left(x^*,x_t\right)}} \leq \frac{1}{2\lambda_t}.$$

我们得到 $\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \leq 1$。因此

$$\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_{t-1}\right] = \exp\left(S_{t-1}\right)\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \leq \exp\left(S_{t-1}\right)$$

这意味着 $(S_t)_{t\geq 1}$ 是一个超鞅。根据 Ville 不等式，对于所有 $k \geq 1$，我们有

$$\Pr\left[S_k \geq \log\frac{1}{\delta}\right] \leq \delta\mathbb{E}\left[\exp\left(S_1\right)\right] \leq \delta$$

换言之，以至少 $1-\delta$ 的概率，对于所有 $k \geq 1$

$$\sum_{t=1}^{k} Z_t \leq \log\frac{1}{\delta}$$

代入 Z_t 的定义，我们有

$$\begin{aligned}
&\sum_{t=1}^{k} z_t\eta_t\Delta_{t+1} + \sum_{t=1}^{k}\left(z_t\mathbf{D}_\psi\left(x^*,x_{t+1}\right) - z_t\mathbf{D}_\psi\left(x^*,x_t\right)\right) \\
\leq &\log\frac{1}{\delta} + \sum_{t=1}^{k} z_t\eta_t\left\langle x^*-x_t,\theta_t^b\right\rangle + 2\sum_{t=1}^{k} z_t\eta_t^2\left\|\theta_t^b\right\|_*^2 \\
&+ \sum_{t=1}^{k}\left(\left(2z_t\eta_t^2 + \frac{3}{8\lambda_t^2} + 24z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right)
\end{aligned}$$

注意，我们有 e z_t 是一个递减序列，因此上述不等式左侧可以由 b 界定 y

$$\begin{aligned}
\text{LHS} &= \sum_{t=1}^{k} z_t\eta_t\Delta_{t+1} + z_k\mathbf{D}_\psi\left(x^*,x_{k+1}\right) - z_1\mathbf{D}_\psi\left(x^*,x_1\right) + \sum_{t=2}^{k}\left(z_{k-1}-z_k\right)\mathbf{D}_\psi\left(x^*,x_k\right) \\
&\geq \sum_{t=1}^{k} z_t\eta_t\Delta_{t+1} + z_k\mathbf{D}_\psi\left(x^*,x_{k+1}\right) - z_1\mathbf{D}_\psi\left(x^*,x_1\right)
\end{aligned}$$

由此我们得到所需不等式。 □

命题 3.7。 我们假设事件 $E(\delta)$ 发生。假设对于某个 $\ell \leq T$，存在常数C_1 和 C_2，使得对所有 $t \leq \ell$

1. $\lambda_t\eta_t = C_1$
2. $\sum_{t=1}^{\ell}\left(\frac{1}{\lambda_t}\right)^p \leq C_2$
3. $\left(\frac{1}{\lambda_t}\right)^{2p} \leq C_3\left(\frac{1}{\lambda_t}\right)^p$
4. $\left\|\nabla f(x_t)\right\|_* \leq \frac{\lambda_t}{2}$

Then for all $t \leq \ell+1$

$$\sum_{i=1}^{t}\eta_i\Delta_{i+1} + \mathbf{D}_\psi\left(x^*,x_{t+1}\right) \leq \frac{1}{2}\left(R_1+8AC_1\right)^2$$

对于 $A \geq \max\left\{\log\frac{1}{\delta} + 26\sigma^p C_2 + \frac{2\sigma^{2p}C_2C_3}{A};\right.$
$\left.1\right\}$。

证明。 我们将通过归纳法证明，在 k

$$\sum_{i=1}^{k}\eta_i\Delta_{i+1}+\mathbf{D}_\psi\left(x^*,x_{k+1}\right)\leq\frac{1}{2}\left(R_1+8AC_1\right)^2.$$

基例 $k=0$ 显然成立。我们有 $\mathrm{D}\psi\left(x^*,x_1\right)=\frac{R_1^2}{2}$。假设该命题对所有 $t\leq k\leq\ell$ 均成立。现在我们证明它对 $k+1$ 也成立。回顾

$$z_t=\frac{1}{2\eta_t\lambda_t\max_{i\leq t}\sqrt{2\mathbf{D}_\psi\left(x^*,x_i\right)}+16Q\eta_t^2\lambda_t^2}$$

我们选择 $Q=A>1$。根据归纳假设以及 $\lambda_t\eta_t=C_1$ 这一假设，我们有

$$\begin{aligned}z_t&\leq\frac{1}{2C_1\left(R_1+8AC_1\right)}\\z_k&\geq\frac{1}{2\eta_k\lambda_k\left(R_1+8AC_1\right)+16A\eta_k^2\lambda_k^2}\\&=\frac{1}{2C_1\left(R_1+16AC_1\right)}\end{aligned}$$

由于 z_k 是一个递减序列

$$\begin{aligned}z_k\sum_{t=1}^{k}\eta_t\Delta_{t+1}+z_k\mathbf{D}_\psi\left(x^*,x_{k+1}\right)\leq{}&z_1\mathbf{D}_\psi\left(x^*,x_1\right)+\log\frac{1}{\delta}+\sum_{t=1}^{k}z_t\eta_t\left\langle x^*-x_t,\theta_t^b\right\rangle+2\sum_{t=1}^{k}z_t\eta_t^2\left\|\theta_t^b\right\|_*^2\\&+\sum_{t=1}^{k}\left(\left(2z_t\eta_t^2+\frac{3}{8\lambda_t^2}+24z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2\mid\mathcal{F}_{t-1}\right]\right)\end{aligned}$$

根据 λ_t 的选取，对于所有 $t\leq k$, $\left\|\nabla f(x_t)\right\|_*\leq\frac{\lambda_t}{2}$，我们可以应用引理 2.1，得到

$$\begin{aligned}\left\|\theta_t^b\right\|_*&\leq4\sigma^p\lambda_t^{1-p};\\\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2\mid\mathcal{F}_{t-1}\right]&\leq40\sigma^p\lambda_t^{2-p}.\end{aligned}$$

因此，我们有

$$\begin{aligned}&z_k\sum_{t=1}^{k}\eta_t\Delta_{t+1}+z_k\mathbf{D}_\psi\left(x^*,x_{k+1}\right)\\\leq{}&z_1\mathbf{D}_\psi\left(x^*,x_1\right)+\log\frac{1}{\delta}+4\sum_{t=1}^{k}z_t\eta_t\sigma^p\lambda_t^{1-p}\sqrt{2\mathbf{D}_\psi\left(x^*,x_t\right)}+32\sum_{t=1}^{k}z_t\eta_t^2\sigma^{2p}\lambda_t^{2-2p}\\&+40\sum_{t=1}^{k}\left(\left(2z_t\eta_t^2+\frac{3}{8\lambda_t^2}+24z_t^2\eta_t^4\lambda_t^2\right)\sigma^p\lambda_t^{2-p}\right)\\\leq{}&z_1\mathbf{D}_\psi\left(x^*,x_1\right)+\log\frac{1}{\delta}+\frac{2C_1\left(R_1+8AC_1\right)\sigma^p}{C_1\left(R_1+8AC_1\right)}\sum_{t=1}^{k}\left(\frac{1}{\lambda_t}\right)^p+\frac{16C_1^2\sigma^{2p}}{C_1\left(R_1+8AC_1\right)}\sum_{t=1}^{k}\left(\frac{1}{\lambda_t}\right)^{2p}\\&+40\left(\frac{C_1^2}{C_1\left(R_1+8AC_1\right)}+\frac{3}{8}+\frac{6C_1^4}{C_1^2\left(R_1+8AC_1\right)^2}\right)\sigma^p\sum_{t=1}^{k}\left(\frac{1}{\lambda_t}\right)^p\\\leq{}&\frac{R_1^2}{4\left(C_1R_1+8AC_1^2\right)}+\log\frac{1}{\delta}+2\sigma^pC_2+\frac{2\sigma^{2p}C_2C_3}{A}+24\sigma^pC_2\\\leq{}&\frac{R_1^2}{4\left(C_1R_1+8AC_1^2\right)}+A\end{aligned}$$

其中，最后一个不等式使用了 $\sum_{t=1}^{k}\left(\frac{1}{\lambda_t}\right)^p \le C_2$ 和 $\left(\frac{1}{\lambda_t}\right)^{2p} \le C_3\left(\frac{1}{\lambda_t}\right)^p$。我们得到

$$\begin{aligned}\sum_{t=1}^{k}\eta_t\Delta_{t+1}+\mathbf{D}_\psi\left(x^*,x_{k+1}\right) &\le 2C_1\left(R_1+16AC_1\right)\left(\frac{R_1^2}{4\left(C_1R_1+8AC_1^2\right)}+A\right)\\ &=\frac{1}{2}R_1^2+\frac{4AC_1^2R_1^2}{C_1R_1+8AC_1^2}+2A\left(C_1R_1+16AC_1^2\right)\\ &\le \frac{1}{2}R_1^2+6AC_1R_1+32A^2C_1^2\\ &\le \frac{1}{2}\left(R_1+8AC_1\right)^2.\end{aligned}$$

□

No下面给出定理 3.1 的证明，它可直接由命题 3.7 推出。

证明。 1. 注意到 $\eta \le \frac{R_1}{16}\frac{1}{4LR_1} \le \frac{1}{4L}$。我们有，事件 $E(\delta)$ 发生的概率至少为 $1-\delta$。在条件化于该事件的情况下，在 3.7 中我们选择

$$C_1=\frac{R_1}{24\gamma};\quad C_2=\frac{\gamma}{26\sigma^p};\quad C_3=\frac{\gamma}{26T\sigma^p};\quad A=3\gamma$$

我们有

$$\begin{aligned}\lambda_t\eta_t &= C_1\\ \sum_{t=1}^{T}\left(\frac{1}{\lambda_t}\right)^p &\le \sum_{t=1}^{T}\left(\frac{\gamma}{26T}\right)\frac{1}{\sigma^p}=C_2\\ \left(\frac{1}{\lambda_t}\right)^{2p} &\le \frac{1}{\sigma^p}\left(\frac{\gamma}{26T}\right)\left(\frac{1}{\lambda_t}\right)^p = C_3\left(\frac{1}{\lambda_t}\right)^p\\ \max\left\{\log\frac{1}{\delta}+26\sigma^pC_2+\frac{2\sigma^{2p}C_2C_3}{A};1\right\} &\le 3\gamma=A\end{aligned}$$

我们只需证明，对所有 t

$$\left\|\nabla f(x_t)\right\|_* \le \frac{\lambda_t}{2}$$

我们将用归纳法证明这一点。事实上，我们有

$$\begin{aligned}\left\|\nabla f(x_1)\right\|_* &= \left\|\nabla f(x_1)-\nabla f(x^*)\right\|_*+\left\|\nabla f(x_0)-\nabla f(x^*)\right\|_*\\ &\quad+\left\|\nabla f(x_0)-g_0\right\|_*+\left\|g_0\right\|_*\\ &\le L\left\|x_1-x^*\right\|+L\left\|x_0-x^*\right\|+\mu\sigma+\left\|g_0\right\|_*\\ &\le LR_1+LR_0+\mu\sigma+\left\|g_0\right\|_* \le \frac{\lambda_1}{2}\end{aligned}$$

假设对所有 $t\le k$ 都成立。我们证明

$$\left\|\nabla f(x_{k+1})\right\|_* \le \frac{\lambda_{k+1}}{2}$$

由 3.7 可得

$$\left\|x_{k+1}-x^*\right\| \le \sqrt{2\mathbf{D}_\psi\left(x^*,x_{k+1}\right)} \le R_1+8AC_1=2R_1$$

Thus

$$\begin{aligned}\|\nabla f(x_{k+1})\|_* &= \|\nabla f(x_{k+1}) - \nabla f(x^*)\|_* + \|\nabla f(x_0) - \nabla f(x^*)\|_* \\ &\quad + \|\nabla f(x_0) - g_0\|_* + \|g_0\|_* \\ &\le L\|x_{k+1} - x^*\| + L\|x_0 - x^*\| + \mu\sigma + \|g_0\|_* \\ &\le 2LR_1 + LR_0 + \mu\sigma + \|g_0\|_* \le \frac{\lambda_{k+1}}{2}\end{aligned}$$

符合所需。因此，由引理 3.6 可得

$$\eta \sum_{t=1}^{T} \Delta_{t+1} + \mathbf{D}_\psi(x^*, x_{T+1}) \le 2R_1^2$$

从而得到

$$\frac{1}{T}\sum_{t=2}^{T+1} \Delta_t \le \frac{2R_1^2}{\eta} = 48R_1 \max\left\{26^{\frac{1}{p}} T^{\frac{1-p}{p}} \sigma\gamma^{\frac{p-1}{p}}; 2\left(2LR_1 + LR_0 + \mu\sigma + \|g_0\|_*\right) T^{-1}\gamma\right\}$$

2. 我们可以遵循类似的步骤。注意，(η_t) 是一个递减序列。我们还使用事实 3.8 来验证命题 3.7 的第二个条件。证明略。 □

事实 3.8。 我们有 $\sum_{t=1}^{\infty} \frac{1}{2t\,(1+(\log t)^2)} < 1$.

注释 3.9。 在定理 3.1 中，我们使用到最优解的初始距离 R_1 来设置步长和裁剪参数。通常无法获得这一信息，但可以避免使用它。例如，对于我们知道定义域半径有 R 上界的约束问题，可以将定理 3.1 中的 R_1 替换为 R，而不改变依赖关系。对于一般问题，定理 3.10 中的参数选择不需要知道任何常数 T、σ、δ 或 R_1。然而，我们需要一个温和的假设，即知道 $\|\nabla f(x_1)\|_*$ 的上界 ∇_1，并且该上界可以被高精度地估计（有关讨论请参见假设 5）。我们还注意到，当 σ 未知时，收敛保证会失去对该常数的自适应性。

定理 3.10。 假设 f 满足假设 *(1)*、*(2)*、*(3)*、*(4)* 和 *(5)*。令 $\gamma =$ 最大值 $\{$对数 $\frac{1}{\delta}$； $1\}$; $R_1 = \sqrt{2\mathbf{D}_\psi}$ (x^*, x_1)，并假设 ∇_1 是 $\|\nabla f(x_1)\|_*$ 的上界。我们选择 λ_t 和 η_t，使得

$$\begin{aligned}\lambda_t &= \max\left\{\left(52t(1+\log t)^2 c_2\right)^{1/p}; 2\left(L\max_{i\le t}\|x_i - x_1\| + \nabla_1\right); \frac{Lc_1}{6}\right\}, \text{ and} \\ \eta_t &= \frac{c_1}{24\lambda_t} = \frac{c_1}{24}\min\left\{\left(52t(1+\log t)^2 c_2\right)^{-1/p}; \frac{1}{2\left(L\max_{i\le t}\|x_i - x_1\| + \nabla_1\right)}; \frac{6}{Lc_1}\right\},\end{aligned}$$

其中，常数 c_1 和 c_2 用于确保维度的正确性。于是，以至少 $1-\delta$ 的概率，我们有

$$\begin{aligned}\frac{1}{T}\sum_{t=2}^{T+1}\Delta_t &\le \frac{8}{Tc_1}\left(R_1 + \frac{c_1}{3}\left(\gamma + \frac{2\sigma^p}{c_2}\right)\right)^2 \max\left\{\left(52T(1+\log T)^2 c_2\right)^{1/p}; 4R_1 L + \frac{2c_1}{3}L\left(\gamma + \frac{2\sigma^p}{c_2}\right) + 2\nabla_1; \frac{Lc_1}{6}\right\} \\ &= \widetilde{O}\left(T^{\frac{1-p}{p}}\right).\end{aligned}$$

证明。 注意到 $\eta_t \le \frac{1}{4L}$。以至少 $1-\delta$ 的概率，事件 $E(\delta)$ 会发生。在此事件的条件下，在 3.7 中，我们选择

$$C_1 = \frac{c_1}{24}; \quad C_2 = \frac{1}{26c_2}; \quad C_3 = \frac{1}{52c_2}; \quad A = \gamma + \frac{2\sigma^p}{c_2}$$

我们验证命题 3.7 的条件

$$\begin{aligned}\lambda_t\eta_t &= C_1\\ \sum_{t=1}^{T}\left(\frac{1}{\lambda_t}\right)^p &\le \sum_{t=1}^{T}\frac{1}{52t(1+\log t)^2c_2} \le \frac{1}{26c_2} = C_2\\ \left(\frac{1}{\lambda_t}\right)^{2p} &\le \frac{1}{52tc_2}\left(\frac{1}{\lambda_t}\right)^p \le C_3\left(\frac{1}{\lambda_t}\right)^p\\ \max\left\{\log\frac{1}{\delta}+26\sigma^pC_2+\frac{2\sigma^{2p}C_2C_3}{A};1\right\} &= \max\left\{\log\frac{1}{\delta}+\frac{\sigma^p}{c_2}+\frac{\sigma^p}{c_2};1\right\} \le A\end{aligned}$$

其中我们有 $\frac{2\sigma^{2p}C_2C_3}{A} \le 2\sigma^{2p}C_2C_3 \times \frac{c_2}{2\sigma^p} \le \frac{\sigma^p}{c_2}$。

$$\begin{aligned}\|\nabla f(x_t)\|_* &= \|\nabla f(x_t)-\nabla f(x_1)\|_* + \|\nabla f(x_1)\|_*\\ &\le L\|x_t-x_1\|_* + \|\nabla f(x_1)\|_* \le \frac{\lambda_{k+1}}{2}\end{aligned}$$

因此，根据引理 3.6，我们有

$$\begin{aligned}\eta_T\sum_{t=1}^{T}\Delta_{t+1}+\mathbf{D}_\psi(x^*,x_{T+1}) &\le \frac{1}{2}(R_1+8AC_1)^2\\ &= \frac{1}{2}\left(R_1+\frac{c_1}{3}\left(\gamma+\frac{2\sigma^p}{c_2}\right)\right)^2\end{aligned}$$

从而得到

$$\begin{aligned}\frac{1}{T}\sum_{t=2}^{T+1}\Delta_t &\le \frac{1}{2T\eta_T}\left(R_1+\frac{c_1}{3}\left(\gamma+\frac{2\sigma^p}{c_2}\right)\right)^2\\ &= \frac{8}{Tc_1}\left(R_1+\frac{c_1}{3}\left(\gamma+\frac{2\sigma^p}{c_2}\right)\right)^2\max\left\{\left(52T(1+\log T)^2c_2\right)^{1/p};2\left(L\max_{i\le T}\|x_i-x_1\|+\nabla_1\right);\frac{L}{8}\right\}\end{aligned}$$

注意到

$$\begin{aligned}\|x_i-x_1\| &\le \|x_i-x^*\|+\|x_1-x^*\|\\ &\le 2R_1+\frac{c_1}{3}\left(\gamma+\frac{2\sigma^p}{c_2}\right)\end{aligned}$$

从而得到最终的收敛速率。 □

4 截断加速随机镜像下降

本节将裁剪随机镜像下降的分析扩展到裁剪加速随机镜像下降的情形（算法 2）。我们将看到，只需稍作修改，分析过程基本相同。算法 2 给出了加速随机镜像下降的裁剪版本（见 [15]），其中使用裁剪梯度 $\nabla f(x_t)$ 而非随机梯度 $\nabla f(x_t)$ 来更新迭代点。

我们采用以下附加假设：

(5') 全局最小化器：我们假设 $\nabla f(x^*) = 0$。

换言之，我们假设全局最小化器位于问题的定义域内。该假设与 [7, 29] 的研究一致。

我们首先给出已知时间范围时的收敛保证。未知 T 的情况将在附录中讨论。

算法 2裁剪加速随机镜像下降

参数：初始值　　点 $y_1 = z_1$，步长 $\{\eta_t\}$，裁剪参数 $\{\lambda_t\}$, ψ 关于 $\|\cdot\|$ 是 1-强凸的

对于 $t = 1$ 到 T，执行

Set $\alpha_t = \frac{2}{t+1}$

$x_t = (1-\alpha_t)\, y_t + \alpha_t z_t$

$\widetilde{\nabla} f(x_t) = \min\left\{1, \frac{\lambda_t}{\left\|\widehat{\nabla} f(x_t)\right\|_*}\right\} \widehat{\nabla} f(x_t)$

$z_{t+1} = \arg\min_{x\in\mathcal{X}} \left\{\eta_t \left\langle \widetilde{\nabla} f(x_t), x\right\rangle + \mathbf{D}_\psi\left(x, z_t\right)\right\}$

$y_{t+1} = (1-\alpha_t)\, y_t + \alpha_t z_{t+1}$

定理 4.1. 假设 f 满足假设 *(1)*、*(2)*、*(3)*、*(4)* 和 *(5')*。令 $\gamma =$最大值 $\left\{\text{对数}\frac{1}{\delta};\ 1\right\}$；且 $R_1 = \sqrt{2\mathbf{D}_\psi\left(x^*, x_1\right)}$。对于已知的 T，我们选择常数 c、λ_t 和 η_t，使得

$$c = \max\left\{10^4; \frac{4\left(T+1\right)\left(\frac{26T}{\gamma}\right)^{\frac{1}{p}}\sigma}{\gamma L R_1}\right\}$$

$$\lambda_t = \frac{cR_1\gamma L\alpha_t}{8} = \max\left\{\frac{10^4 R_1\gamma L}{6(t+1)}; \frac{T+1}{t+1}\left(\frac{26T}{\gamma}\right)^{1/p}\sigma\right\}$$

$$\eta_t = \frac{1}{3c\gamma^2 L\alpha_t} = \frac{R_1}{24\gamma}\min\left\{\frac{4(t+1)}{10^4 R_1\gamma L}; \frac{t+1}{T+1}\left(\frac{26T}{\gamma}\right)^{-1/p}\sigma^{-1}\right\}$$

那么，以至少 $1-\delta$ 的概率

$$f\left(y_{T+1}\right) - f\left(x^*\right) \le 6\max\left\{10^4 L\gamma^2 R_1^2 (T+1)^{-2}; 4R_1\left(T+1\right)^{-1}\left(26T\right)^{\frac{1}{p}}\gamma^{\frac{p-1}{p}}\sigma\right\}$$

注 4.2. 加速算法的一个特点是在两种情形之间进行插值：当 σ 较大时，算法达到与非加速算法相同的 $O\left(T^{\frac{1-p}{p}}\right)$ 收敛；然而，当 σ 足够小时，算法达到加速的 $O\left(T^{-2}\right)$ 速率。

我们也通过以下引理对加速随机镜像下降进行基本分析。

引理4.3. 假设 f 满足假设 *(1)*、*(2)*、*(3)*、*(4)* 和 $\eta_t \le \frac{1}{2L\alpha_t}$，则算法 *1* 输出的迭代序列 $(x_t)_{t\ge 1}$ 满足以下条件

$$\begin{aligned}
&\frac{\eta_t}{\alpha_t}\left(f\left(y_{t+1}\right) - f\left(x^*\right)\right) - \frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right) - f\left(x^*\right)\right) + \mathbf{D}_\psi\left(x^*, z_{t+1}\right) - \mathbf{D}_\psi\left(x^*, z_t\right)\\
\le&\eta_t\left\langle\theta_t^u, x^* - z_t\right\rangle + \eta_t\left\langle\theta_t^b, x^* - z_t\right\rangle + 2\eta_t^2\left(\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right) + 2\eta_t^2\left\|\theta_t^b\right\|_*^2 + 2\eta_t^2\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]
\end{aligned}$$

与上一节类似，我们定义以下变量

$$\begin{aligned}
Z_t = z_t\Bigg(&\frac{\eta_t}{\alpha_t}\left(f\left(y_{t+1}\right) - f\left(x^*\right)\right) - \frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right) - f\left(x^*\right)\right) + \mathbf{D}_\psi\left(x^*, z_{t+1}\right) - \mathbf{D}_\psi\left(x^*, z_t\right)\\
&- \eta_t\left\langle\theta_t^b, x^* - z_t\right\rangle - 2\eta_t^2\left\|\theta_t^b\right\|_*^2 - 2\eta_t^2\mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\Bigg) - \left(\frac{3}{8\lambda_t^2} + 24 z_t^2\eta_t^4\lambda_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|^2 \mid \mathcal{F}_{t-1}\right]
\end{aligned}$$

where $z_t = \dfrac{1}{2\eta_t\lambda_t\max_{i\le t}\sqrt{2\mathbf{D}_\psi\left(x^*, x_i\right)} + 16Q\eta_t^2\lambda_t^2}$

对于常数 $Q \ge 1$，以及

$$S_t = \sum_{i=1}^{t} Z_i$$

按照相同的分析，我们可以得到引理 4.4 和命题 4.5 中的结论，证明过程从略。在证明引理 4.4 时，唯一需要注意的步骤是对以下求和进行界定

$$\sum_{t=1}^{k} \frac{z_t \eta_t}{\alpha_t} \left(f\left(y_{t+1}\right) - f\left(x^*\right) \right) - \frac{z_t \eta_t \left(1-\alpha_t\right)}{\alpha_t} \left(f\left(y_t\right) - f\left(x^*\right) \right)$$

注意，如果我们假设 $\frac{\eta_{t-1}}{\alpha_{t-1}} \geq \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t}$，由于 z_t 是递减序列且 $\alpha_1 = 0$，我们可以用最后一项 $\frac{z_k \eta_k}{\alpha_k}(f(y_{k+1}) - f(x^*))$ 给上述求和一个下界，从而得到所需不等式。

引理 4.4。假设对于所有 $t \geq 1$，η_t 满足 $\frac{\eta_{t-1}}{\alpha_{t-1}} \geq \frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t}$。对于任意 $\delta > 0$，令 $E(\delta)$ 表示对所有 $1 \leq k \leq T$ 都成立的事件

$$\begin{aligned}
&\frac{z_k \eta_k}{\alpha_k} \left(f\left(y_{k+1}\right) - f\left(x^*\right) \right) + z_k \mathbf{D}_\psi \left(x^*, x_{k+1}\right) \\
\leq & z_1 \mathbf{D}_\psi \left(x^*, x_1\right) + \log \frac{1}{\delta} + \sum_{t=1}^{k} z_t \eta_t \left\langle x^* - x_t, \theta_t^b \right\rangle + 2 \sum_{t=1}^{k} z_t \eta_t^2 \left\| \theta_t^b \right\|_*^2 \\
& + \sum_{t=1}^{k} \left(\left(2 z_t \eta_t^2 + \frac{3}{8 \lambda_t^2} + 24 z_t^2 \eta_t^4 \lambda_t^2 \right) \mathbb{E}\left[\left\| \theta_t^u \right\|_*^2 \mid \mathcal{F}_{t-1} \right] \right)
\end{aligned}$$

于是 概率 $[E(\delta)] \geq 1 - \delta$。

命题 4.5。我们假设事件 $E(\delta)$ 发生。假设对于某个 $\ell \leq T$，存在常数 C_1 和 C_2，使得对于所有 $t \leq \ell$

1. $\lambda_t \eta_t = C_1$
2. $\sum_{t=1}^{\ell} \left(\frac{1}{\lambda_t} \right)^p \leq C_2$
3. $\left(\frac{1}{\lambda_t} \right)^{2p} \leq C_3 \left(\frac{1}{\lambda_t} \right)^p$
4. $\left\| \nabla f\left(x_t\right) \right\|_* \leq \frac{\lambda_t}{2}$

于是
对于所有 $t \leq \ell + 1$

$$\frac{\eta_t}{\alpha_t} \left(f\left(y_{t+1}\right) - f\left(x^*\right) \right) + \mathbf{D}_\psi \left(x^*, z_{t+1}\right) \leq \frac{1}{2} \left(R_1 + 8 A C_1 \right)^2$$

对于 $A \geq \max\left\{ \log \frac{1}{\delta} + 26 \sigma^p C_2 + \frac{2 \sigma^{2p} C_2 C_3}{A}; 1 \right\}$

我们现在可以证明定理 4.1。

证明。注意到 $\eta_t \leq \frac{1}{2c\gamma^2 L \alpha_t} \leq \frac{1}{2L\alpha_t}$ and

$$\begin{aligned}
\frac{\eta_{t-1}}{\alpha_{t-1}} &= \frac{t^2}{8c\gamma^2 L} \\
\frac{\eta_t \left(1-\alpha_t\right)}{\alpha_t} &= \frac{(t+1)(t-1)}{8c\gamma^2 L}
\end{aligned}$$

因此 $\frac{\eta_{t-1}}{\alpha_{t-1}} \geq \frac{\eta_t}{\left(1-\alpha_t\right) \alpha_t}$。我们有，事件 $E(\delta)$ 以至少 $1-\delta$ 的概率发生。在此事件上进行条件化，在 3.7 中我们选择

$$C_1 = \frac{R_1}{24\gamma}; \quad C_2 = \frac{\gamma}{26\sigma^p}; \quad C_3 = \frac{\gamma}{26 T \sigma^p}; \quad A = 3\gamma$$

我们验证命题 4.5 的条件。前四个条件与上一节中的情况类似，均成立。我们将通过归纳法证明，对所有 $t \geq 1$, $\left\| \nabla f(x_t) \right\|_* \leq \frac{\lambda_t}{2}$，且 $\max\left\{ \left\| x_t - x^* \right\|, \left\| y_t - x^* \right\|, \left\| z_t - x^* \right\| \right\} \leq 2R_1$，均成立。

对于 $t=1$，注意到 $x_1=y_1=z_1$，因此我们有

$$\|\nabla f(x_1)\|_* = \|\nabla f(x_1)-\nabla f(x^*)\|_* \le LR_1 \le \frac{\lambda_1}{2}$$

现在假设　　结论对 $1\le t\le k$ 成立，我们证明它对 $t=k+1$ 也成立。由命题 4.5 可知　　t

$$\frac{2\eta_k}{\alpha_k} f\left(y_{k+1}\right) - f\left(x^*\right) + \left\|z_{k+1}-x^*\right\|^2 \le 4R_1^2$$

此外

$$\begin{aligned}\left\|y_{k+1}-x^*\right\| &\le (1-\alpha_k)\left\|y_k-x^*\right\| + \alpha_k\left\|z_{k+1}-x^*\right\| \le 2R_1\\ \left\|x_{k+1}-x^*\right\| &\le (1-\alpha_k)\left\|y_{k+1}-x^*\right\| + \alpha_k\left\|z_{k+1}-x^*\right\| \le 2R_1\end{aligned}$$

对于 $k\ge 1$，我们有 $\alpha_{k+1}=\frac{2}{k+2}<1$；　$\frac{\alpha_{k+1}}{1-\alpha_{k+1}}=\frac{2}{k}\le\frac{4}{k+2}\le 2\alpha_{t+1}$ 且 $\alpha_t\le\frac{3}{2}\alpha_{t+1}$。因此 e

$$\begin{aligned}\|\nabla f(x_{k+1})\|_* &\le \|\nabla f(x_{k+1})-\nabla f(y_{k+1})\|_* + \|\nabla f(y_{k+1})-\nabla f(x^*)\|_*\\ &\le L\left\|x_{k+1}-y_{k+1}\right\| + \sqrt{2L\left(f\left(y_{k+1}\right)-f\left(x^*\right)\right)}\\ &\le \frac{L\alpha_{k+1}\left\|x_{k+1}-z_{k+1}\right\|}{1-\alpha_{k+1}} + 2R_1\sqrt{\frac{L\alpha_t}{2\eta_t}}\\ &\le 4LR_1\frac{\alpha_{k+1}}{1-\alpha_{k+1}} + 2\sqrt{\frac{3}{2}}c\gamma R_1 L\alpha_t\\ &\le 8\gamma LR_1\alpha_{t+1} + 3\sqrt{\frac{3}{2}}c\gamma LR_1\alpha_{t+1}\\ &\le (8+3\sqrt{\frac{3}{2}}c)R_1\gamma L\alpha_{t+1}\\ &= \frac{16(8+3\sqrt{\frac{3}{2}}c)\lambda_{t+1}}{2c} \le \frac{\lambda_{t+1}}{2}.\end{aligned}$$

正如所需。因此我们有

$$\frac{\eta_T}{\alpha_T}\left(f\left(y_{T+1}\right)-f\left(x^*\right)\right) + \mathbf{D}_\psi\left(x^*, x_{T+1}\right) \le 2R_1^2$$

由此得到

$$\begin{aligned}f\left(y_{T+1}\right)-f\left(x^*\right) &\le \frac{2R_1^2\alpha_T}{\eta_T} = 6R_1^2c\gamma^2L\alpha_T^2\\ &= 6\max\left\{10^4L\gamma^2R_1^2(T+1)^{-2}; 6R_1\left(T+1\right)^{-1}\left(26T\right)^{\frac{1}{p}}\gamma^{\frac{p-1}{p}}\sigma\right\}.\end{aligned}$$

□

5 非凸函数的裁剪随机梯度下降

本节研究非凸函数下裁剪随机梯度下降的收敛。在此设定中，我们只考虑定义域为 $\mathbb{R}^d$ 且范数为 ℓ_2 范数的情形。算法3中的一般框架由 [7] 提出（不过针对的是凸目标函数），并在 [2, 24, 19, 29] 中进行了研究。我们注意到，尽管这些后续工作在各种设定下实现了近乎最优的 $\widetilde{O}\left(T^{\frac{2-2p}{3p-2}}\right)$ 时间依赖性，但它们都存在同样的局限：将集中不等式作为黑箱使用，并且要求已知时间范围。我们在这方面改进了这些工作。我们将再次证明，白箱方法是一种强大的方法，既能对高概率下的收敛进行紧确分析，也能很好地推广到不同情形。

我们将证明已知 T 时的以下保证。未知 T 时的结论见附录。

算法3裁剪随机梯度下降

参数：初始点 x_1、步长 $\{\eta_t\}$、裁剪参数 $\{\lambda_t\}$，对于 $t=1$ 至 T 执行

$$\widetilde{\nabla} f(x_t) = \min\left\{1, \frac{\lambda_t}{\|\widehat{\nabla} f(x_t)\|}\right\} \widehat{\nabla} f(x_t)$$
$$x_{t+1} = x_t - \eta_t \widetilde{\nabla} f(x_t)$$

定理 5.1. 假设 f 满足假设 *(1')*、*(2)*、*(3)*、*(4)*。令 $\gamma = \max\left\{\log\frac{1}{\delta}; 1\right\}$ 且 $\Delta_1 = f(x_1) - f^*$。对于已知的 T，我们选择 λ_t 和 η_t，使得

$$\lambda_t = \lambda = \max\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}} T^{\frac{1}{3p-2}} \sigma^{\frac{p}{p-1}}; 2\sqrt{90L\Delta_1}; 32^{\frac{1}{p}} \sigma T^{\frac{1}{3p-2}}\right\}$$
$$\eta_t = \eta = \frac{\sqrt{\Delta_1} T^{\frac{1-p}{3p-2}}}{8\lambda\sqrt{L\gamma}} = \frac{\sqrt{\Delta_1}}{8\sqrt{L\gamma}} \min\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{-1}{p-1}} T^{\frac{-p}{3p-2}} \sigma^{\frac{-p}{p-1}}; \frac{T^{\frac{1-p}{3p-2}}}{2\sqrt{90L\Delta_1}}; \frac{T^{\frac{-p}{3p-2}}}{32^{1/p}\sigma}\right\}$$

那么，以至少 $1-\delta$ 的概率

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \le 720\sqrt{\Delta_1 L\gamma} \max\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}} T^{\frac{2-2p}{3p-2}} \sigma^{\frac{p}{p-1}}; 2\sqrt{90L\Delta_1} T^{\frac{1-2p}{3p-2}}; 32^{1/p}\sigma T^{\frac{2-2p}{3p-2}}\right\} = O\left(T^{\frac{2-2p}{3p-2}}\right).$$

Th算法3的分析从以下引理开始，其证明见附录 .

引理 5.2. 假设函数 f 满足假设 *(1')*、*(2)*、*(3)*、*(4)*，并且 $\eta_t \le \frac{1}{L}$，则对所有$t \ge 1$,

$$\begin{aligned}\frac{\eta_t}{2}\|\nabla f(x_t)\|^2 \le \Delta_t - \Delta_{t+1} + \left(L\eta_t^2 - \eta_t\right)\langle \nabla f(x_t), \theta_t^u\rangle + \frac{3\eta_t}{2}\left\|\theta_t^b\right\|^2 \\ + L\eta_t^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta^u\|^2 \mid \mathcal{F}_{t-1}\right]\right) + L\eta_t^2 \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\end{aligned}$$

与凸分析类似，为了克服先前工作中归纳式论证所面临的问题，我们定义以下术语

$$\begin{aligned}Z_t = z_t\left(\frac{\eta_t}{2}\|\nabla f(x_t)\|^2 + \Delta_{t+1} - \Delta_t - \frac{3\eta_t}{2}\left\|\theta_t^b\right\|^2 - L\eta_t^2 \mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\ - \left(3z_t^2 L\eta_t^2\Delta_t + 6L^2 z_t^2 \eta_t^4 \lambda_t^2\right)\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\end{aligned}$$
$$\text{where } z_t = \frac{1}{2P_t\eta_t\lambda_t \max_{i\le t}\sqrt{2L\Delta_i} + 8Q_t L\eta_t^2\lambda_t^2}$$

对于 $P_t, Q_t \in \mathcal{F}_{t-1} \ge 1$ 和

$$S_t = \sum_{i=1}^{t} Z_i$$

下面的引理以高概率保证所有时刻 t 的所有项之间存在一种关系。该引理的证明结合了标准技术和前文展示的技术，因此推迟到附录中。

引理 5.3. 对于任意 $\delta > 0$，令 $E(\delta)$ 表示如下事件：对于所有 $1 \le k \le T$

$$\begin{aligned}\frac{1}{2}\sum_{t=1}^{k} z_t\eta_t\|\nabla f(x_t)\|^2 + z_k\Delta_{k+1} \le z_1\Delta_1 + \log\frac{1}{\delta} + \sum_{t=1}^{k}\frac{3z_t\eta_t}{2}\left\|\theta_t^b\right\|^2 \\ + \sum_{t=1}^{k}\left(\left(3z_t^2 L\eta_t^2\Delta_t + 6L^2 z_t^2\eta_t^4\lambda_t^2 + z_t L\eta_t^2\right)\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right)\end{aligned}$$

那么 概率$[E(\delta)] \ge 1-\delta$。

现在我们明确 η_t 和 λ_t 的取值。下面的引理给出了 η_t 和 λ_t 的一般选取条件，使其在时间 T 内达到正确的收敛速率。

命题 5.4。我们假设事件 E 发生。设对于某个 $\ell \leq T$，存在常数C_1、C_2 和 C_3，使得对于所有 $t \leq \ell$

1. $\lambda_t \eta_t \sqrt{2L} \leq C_1$
2. $\frac{1}{L\eta_t}\left(\frac{1}{\lambda_t}\right)^p \leq C_2$
3. $\sum_{t=1}^{T} L\left(\frac{1}{\lambda_t}\right)^p \lambda_t^2 \eta_t^2 \leq C_3$
4. $\|\nabla f(x_t)\| \leq \frac{\lambda_t}{2}$

那么对于所有 $t \leq \ell + 1$

$$\frac{1}{2}\sum_{i=1}^{t} \eta_i \left\|\nabla f(x_i)\right\|^2 + \Delta_{t+1} \leq \left(\sqrt{\Delta_1} + 2\sqrt{A}C_1\right)^2$$

对于常数 $A \geq$ 最大值 $\left\{64\left(\text{对数 } \frac{1}{\delta} + \frac{60\sigma^p C_3}{C_1^2}\right)^2 + \frac{48\sigma^{2p}C_2C_3 + 140\sigma^p C_3}{C_1^2};\quad 1\right\}$。

最后，我们给出定理 3.1 的证明，它是命题 3.7 的直接推论。

$\frac{1}{1-p}$证明。注意到 $\eta \leq \frac{T^{3p-2}}{16\sqrt{90}L\gamma} \leq \frac{1}{L}$。我们以至少 $1-\delta$ 的概率有事件 $E(\delta)$ 发生。对该事件进行条件化后，我们验证引理 5.4 的条件。我们选取以下常数

$$C_1 = \frac{\sqrt{\Delta_1}}{4\sqrt{2}\gamma};\quad C_2 \leq \frac{1}{\sigma^p};\quad C_3 \leq \frac{\Delta_1}{2048\sigma^p\gamma};\quad A = 256\gamma^2$$

我们在引理 C.1 中验证，对于这些常数的选择，命题 5.4 的条件 (1)-(3) 均得到满足。此外，我们有

$$\begin{aligned}
&64\left(\log\frac{1}{\delta} + \frac{60\sigma^p C_3}{C_1^2}\right)^2 + \frac{48\sigma^{2p}C_2C_3 + 140\sigma^p C_3}{C_1^2} \\
&= 64\left(\log\frac{1}{\delta} + 60\log\frac{1}{\delta}\frac{32}{\Delta_1}\frac{\Delta_1}{2048}\right)^2 + \left(48\frac{\Delta_1}{2048} + 140\frac{\Delta_1}{2048}\right)\frac{32}{\Delta_1} \\
&\leq 256\gamma^2 = A
\end{aligned}$$

我们只需证明对所有 t, $\|\nabla f(x_t)\| \leq \frac{\lambda_t}{2}$。我们将通过归纳法证明这一点。事实上，对于基例，我们有 $\|\nabla f(x_1)\| = \sqrt{2L\Delta_1} \leq \frac{\lambda_1}{2}$。假设对所有 $t \leq k$ 该结论均成立。我们将证明$\|\nabla f(x_{k+1})\| \leq \frac{\lambda_{k+1}}{2}$。根据引理 5.4 和归纳假设

$$\Delta_{k+1} \leq \left(\sqrt{\Delta_1} + 2\sqrt{A}C_1\right) \leq \left(\sqrt{\Delta_1} + \frac{\sqrt{\Delta_1}}{2\sqrt{2}\gamma} \times 16\gamma\right)^2 \leq 45\Delta_1$$

因此我们得到

$$\|\nabla f(x_{k+1})\| = \sqrt{2L\Delta_{k+1}} \leq \sqrt{90L\Delta_1} \leq \frac{\lambda_{k+1}}{2}$$

这是所需的。根据引理 3.6，我们有

$$\frac{\eta}{2}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 + \Delta_{k+1} \leq 45\Delta_1$$

因此

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \frac{90\Delta_1}{\eta T} = 720\sqrt{\Delta_1 L}\gamma \max\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}} T^{\frac{2-2p}{3p-2}}\sigma^{\frac{p}{p-1}}; 2\sqrt{90L\Delta_1}T^{\frac{1-2p}{3p-2}}; 32^{\frac{1}{p}}\sigma T^{\frac{2-2p}{3p-2}}\right\}.$$

□

6 结论

本文提出了一种新方法，用于分析存在重尾噪声时的各种梯度裁剪算法。该分析可以轻松扩展到非光滑凸目标函数的情形。 [6] 中已采用与凸优化类似的技术，研究了重尾噪声下变分不等式的高概率收敛问题。将我们的方法扩展到这一设定的问题留待未来研究。

参考文献

[1] Ashok Cutkosky。任意时刻在线到批量转换、乐观算法与加速。载于国际机器学习会议（*ICML*），机器学习研究论文集第97卷，第1446–1454页。PMLR，2019年。[2] Ashok Cutkosky 和 Harsh Mehta。重尾条件下非凸随机优化的高概率界。神经信息处理系统进展，34：4883–4895，2021年。[3] Kacha Dzhaparidze 和 JH Van Zanten。关于鞅的伯恩斯坦型不等式。随机过程及其应用，93(1)：109–117，2001年。[4] David A Freedman。关于鞅的尾概率。概率年刊，第100–118页，1975年。[5] Saeed Ghadimi 和 Guanghui Lan。非凸随机规划的随机一阶与零阶方法。*SIAM*优化杂志，23(4)：2341–2368，2013年。[6] Eduard Gorbunov、Marina Danilova、David Dobre、Pavel Dvurechenskii、Alexander Gasnikov 和 Gauthier Gidel。重尾噪声下用于变分不等式的随机裁剪方法。神经信息处理系统进展，35：31319–31332，2022年。[7] Eduard Gorbunov、Marina Danilova 和 Alexander Gasnikov。通过加速梯度裁剪实现重尾噪声下的随机优化。神经信息处理系统进展，33：15042–15053，2020年。[8] Eduard Gorbunov、Marina Danilova、Innokentiy Shibaev、Pavel Dvurechensky 和 Alexander Gasnikov。重尾噪声下非光滑随机优化的近乎最优高概率复杂度界。*arXiv* 预印本 *arXiv:2106.05958*，2021年。[9] Mert Gurbuzbalaban、Umut Simsekli 和 Lingjiong Zhu。随机梯度下降中的重尾现象。载于国际机器学习会议，第3964–3975页。PMLR，2021年。[10] Nicholas JA Harvey、Christopher Liaw、Yaniv Plan 和 Sikander Randhawa。非光滑随机梯度下降的紧确分析。载于学习理论会议，第1579–1613页。PMLR，2019年。[11] Elad Hazan 和 Satyen Kale。突破遗憾最小化的瓶颈：随机强凸优化的最优算法。机器学习研究杂志，15(1)：2489–2512，2014年。[12] Sham M Kakade 和 Ambuj Tewari。关于在线强凸规划算法的泛化能力。神经信息处理系统进展，21，2008年。[13] Ahmed Khaled 和 Peter Richtárik。非凸领域中随机梯度下降的更优理论。*arXiv* 预印本 *arXiv:2002.03329*，2020年。[14] Guanghui Lan。随机复合优化的一种最优方法。数学规划，133(1)：365–397，2012年。[15] Guanghui Lan。机器学习的一阶与随机优化方法。Springer，2020年。

[16] Shaojie Li 和 Yong Liu。重尾条件下非凸随机梯度下降的高概率保证。载于国际机器学习会议，第12931–12963页。PMLR，2022年。[17] Xiaoyu Li 和 Francesco Orabona。带动量的自适应随机梯度下降的高概率分析。*arXiv* 预印本 *arXiv:2007.14294*，2020年。[18] Zijian Liu、Ta Duy Nguyen、Thien Hang Nguyen、Alina Ene 和 Huy Lê Nguyen。随机梯度方法的高概率收敛。*arXiv* 预印本 *arXiv:2302.14843*，2023年。[19] Zijian Liu、Jiawei Zhang 和 Zhengyuan Zhou。以（少量）结构突破下界：重尾噪声下非凸随机优化的加速。*arXiv* 预印本 *arXiv:2302.06763*，2023年。[20] Liam Madden、Emiliano Dall'Anese 和 Stephen Becker。非凸随机梯度下降的高概率收敛与一致稳定性界。*arXiv* 预印本 *arXiv:2006.05610*，2020年。[21] Stanislav Minsker。巴拿赫空间中的几何中位数与稳健估计。*Bernoulli*，第2308–2335页，2015年。[22] Alexander V Nazin、Arkadi S Nemirovsky、Alexandre B Tsybakov 和 Anatoli B Juditsky。基于镜像下降法的稳健随机优化算法。《自动化与遥控》，80(9)：1607–1627，2019年。[23] Arkadi Nemirovski、Anatoli Juditsky、Guanghui Lan 和 Alexander Shapiro。用于随机规划的稳健随机逼近方法。*SIAM*优化杂志，19(4)：1574–1609，2009年。[24] Ta Duy Nguyen、Thien Hang Nguyen、Alina Ene 和 Huy Le Nguyen。重尾噪声下裁剪随机梯度下降的高概率收敛。*arXiv* 预印本 *arXiv:2302.05437*，2023年。[25] Daniela A Parletta、Andrea Paudice、Massimiliano Pontil 和 Saverio Salzo。重尾噪声下随机次梯度方案的高概率界。*arXiv* 预印本 *arXiv:2208.08567*，2022年。[26] Razvan Pascanu、Tomas Mikolov 和 Yoshua Bengio。理解梯度爆炸问题。*CoRR*，*abs/1211.5063*，2(417)：1，2012年。[27] Maxim Raginsky 和 Alexander Rakhlin。黑箱凸优化的信息复杂度：通过反馈信息论重新审视。载于*2009*年第*47*届 *Allerton* 通信、控制与计算年会（*Allerton*），第803–510页。IEEE，2009年。[28] Alexander Rakhlin、Ohad Shamir 和 Karthik Sridharan。使梯度下降成为强凸随机优化的最优方法。*arXiv* 预印本 *arXiv:1109.5647*，2011年。[29] Abdurakhmon Sadiev、Marina Danilova、Eduard Gorbunov、Samuel Horváth、Gauthier Gidel、Pavel Dvurechensky、Alexander Gasnikov 和 Peter Richtárik。随机优化与变分不等式的高概率界：无界方差情形。*arXiv* 预印本 *arXiv:2302.00999*，2023年。[30] Umut Şimşekli、Mert Gürbüzbalaban、Thanh Huy Nguyen、Gaël Richard 和 Levent Sagun。深度神经网络随机梯度下降的重尾理论。*arXiv* 预印本 *arXiv:1912.00018*，2019年。[31] Umut Simsekli、Levent Sagun 和 Mert Gurbuzbalaban。深度神经网络中随机梯度噪声的尾指数分析。第5827–5837页，2019年。[32] Nuri Mert Vural、Lu Yu、Krishna Balasubramanian、Stanislav Volgushev 和 Murat A Erdogdu。镜像下降再显神威：无限噪声方差下的最优随机凸优化。载于学习理论会议，第65–102页。PMLR，2022年。

[33] Hongjian Wang、Mert Gurbuzbalaban、Lingjiong Zhu、Umut Simsekli 和 Murat A Erdogdu。无限噪声方差下随机梯度下降的收敛速率。神经信息处理系统进展，34:18866–18877，2021年。

[34] Jingzhao Zhang、Sai Praneeth Karimireddy、Andreas Veit、Seungyeon Kim、Sashank Reddi、Sanjiv Kumar 和 Suvrit Sra。为什么自适应方法适用于注意力模型？神经信息处理系统进展，33:15383–15393，2020年。

[35] Jiujia Zhang 和 Ashok Cutkosky。在重尾下以高概率获得无参数遗憾。载于神经信息处理系统进展。

第3节中缺失的证明

引理 A.1。 假设 $\eta_t \leq \frac{1}{4L}$ 并假设 f 满足假设 *(1)*、*(2)*、*(3)* 以及以下条件

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + G \|y - x\| + \frac{L}{2} \|y - x\|^2, \quad \forall y, x \in \mathcal{X}. \tag{6}$$

那么，由算法 *1* 生成的迭代序列$(x_t)^{t\geq 1}$ 满足以下条件：

$$\begin{aligned}\eta_t \Delta_{t+1} \leq\ & \mathbf{D}_\psi\left(x^*, x_t\right) - \mathbf{D}_\psi\left(x^*, x_{t+1}\right) + \eta_t \left\langle x^* - x_t, \theta_t^u \right\rangle + \eta_t \left\langle x^* - x_t, \theta_t^b \right\rangle \\ & + 2\eta_t^2 \left(\|\theta_t^u\|_*^2 - \mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1} \right] \right) + 2\eta_t^2 \mathbb{E}\left[\|\theta_t^u\|_*^2 \mid \mathcal{F}_{t-1} \right] + 2\eta_t^2 \left\|\theta_t^b\right\|_*^2 + 2G^2\eta_t^2\end{aligned}$$

证明。 由条件 (6) 和凸性可得，

$$\begin{aligned}f\left(x_{t+1}\right) - f\left(x^*\right) &\leq \underbrace{f\left(x_{t+1}\right) - f\left(x_t\right)}_{\text{condition (6)}} + \underbrace{f\left(x_t\right) - f\left(x^*\right)}_{\text{convexity}} \\ &\leq \left\langle \nabla f\left(x_t\right), x_{t+1} - x_t \right\rangle + \frac{L}{2} \left\|x_t - x_{t+1}\right\|^2 + G \left\|x_t - x_{t+1}\right\| + \left\langle \nabla f\left(x_t\right), x_t - x^* \right\rangle \\ &= \left\langle \nabla f\left(x_t\right), x_{t+1} - x^* \right\rangle + \frac{L}{2} \left\|x_t - x_{t+1}\right\|^2 + G \left\|x_t - x_{t+1}\right\| \\ &= \left\langle \theta_t, x^* - x_{t+1} \right\rangle + \left\langle \widetilde{\nabla} f(x_t), x_{t+1} - x^* \right\rangle + \frac{L}{2} \left\|x_t - x_{t+1}\right\|^2 + G \left\|x_t - x_{t+1}\right\|\end{aligned}$$

根据最优性条件，我们有

$$\left\langle \eta_t \widetilde{\nabla} f(x_t) + \nabla_x \mathbf{D}_\psi\left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle \geq 0$$

因此

$$\left\langle \eta_t \widetilde{\nabla} f(x_t), x_{t+1} - x^* \right\rangle \leq \left\langle \nabla_x \mathbf{D}_\psi\left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle$$

注意到

$$\begin{aligned}\left\langle \nabla_x \mathbf{D}_\psi\left(x_{t+1}, x_t\right), x^* - x_{t+1} \right\rangle &= \left\langle \nabla \psi\left(x_{t+1}\right) - \nabla \psi\left(x_t\right), x^* - x_{t+1} \right\rangle \\ &= \mathbf{D}_\psi\left(x^*, x_t\right) - \mathbf{D}_\psi\left(x_{t+1}, x_t\right) - \mathbf{D}_\psi\left(x^*, x_{t+1}\right)\end{aligned}$$

Thus

$$\begin{aligned}\eta_t \left\langle \widetilde{\nabla} f(x_t), x_{t+1} - x^* \right\rangle &\leq \mathbf{D}_\psi\left(x^*, x_t\right) - \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x_{t+1}, x_t\right) \\ &\leq \mathbf{D}_\psi\left(x^*, x_t\right) - \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \frac{1}{2} \left\|x_{t+1} - x_t\right\|^2\end{aligned}$$

其中我们利用了由 ψ 的强凸性可知，$\mathbf{D}\psi\left(x_{t+1}, x_t\right) \geq \frac{1}{2} \left\|x_{t+1} - x_t\right\|^2$。

结合这两个不等式，并利用假设 $L\eta_t \leq \frac{1}{4}$，我们得到

$$\begin{aligned}
&\eta_t \Delta_{t+1} + \mathbf{D}_\psi\left(x^*, x_{t+1}\right) - \mathbf{D}_\psi\left(x^*, x_t\right) \\
&\leq \eta_t \left\langle \theta_t, x^* - x_{t+1} \right\rangle + \frac{L\eta_t}{2} \left\| x_t - x_{t+1} \right\|^2 + G\eta_t \left\| x_t - x_{t+1} \right\| - \frac{1}{2} \left\| x_{t+1} - x_t \right\|^2 \\
&\leq \eta_t \left\langle \theta_t, x^* - x_t \right\rangle + \eta_t \left\langle \theta_t, x_t - x_{t+1} \right\rangle - \frac{3}{8} \left\| x_{t+1} - x_t \right\|^2 + G\eta_t \left\| x_t - x_{t+1} \right\| \\
&\leq \eta_t \left\langle \theta_t, x^* - x_t \right\rangle + \eta_t^2 \left\| \theta_t \right\|_*^2 + 2G^2\eta_t^2 \\
&\leq \eta_t \left\langle \theta_t^u + \theta_t^b, x^* - x_t \right\rangle + 2\eta_t^2 \left\| \theta_t^u \right\|_*^2 + 2\eta_t^2 \left\| \theta_t^b \right\|_*^2 + 2G^2\eta_t^2
\end{aligned}$$

我们得到所需不等式。 □

B 第4节缺失的证明

引理*4.3*的证明。我们有

$$\begin{aligned}
f\left(y_{t+1}\right) - f\left(x^*\right) &= \underbrace{f\left(y_{t+1}\right) - f\left(x_t\right)}_{\text{smoothness}} + \underbrace{f\left(x_t\right) - f\left(x^*\right)}_{\text{convexity}} \\
&\leq \left\langle \nabla f\left(x_t\right), y_{t+1} - x_t \right\rangle + \frac{L}{2} \left\| y_{t+1} - x_t \right\|^2 \\
&\quad + \alpha_t \left\langle \nabla f\left(x_t\right), x_t - x^* \right\rangle + \left(1 - \alpha_t\right)\left(f\left(x_t\right) - f\left(x^*\right)\right) \\
&= \underbrace{\left(1 - \alpha_t\right)\left\langle \nabla f\left(x_t\right), y_t - x_t \right\rangle}_{\text{convexity}} + \alpha_t \left\langle \nabla f\left(x_t\right), z_{t+1} - x^* \right\rangle \\
&\quad + \frac{L\alpha_t^2}{2} \left\| z_{t+1} - z_t \right\|^2 + \left(1 - \alpha_t\right)\left(f\left(x_t\right) - f\left(x^*\right)\right) \\
&\leq \left(1 - \alpha_t\right)\left(f\left(y_t\right) - f\left(x_t\right)\right) + \left(1 - \alpha_t\right)\left(f\left(x_t\right) - f\left(x^*\right)\right) \\
&\quad + \alpha_t \left\langle \theta_t, x^* - z_{t+1} \right\rangle + \alpha_t \left\langle \widetilde{\nabla} f(x_t), z_{t+1} - x^* \right\rangle + \frac{L\alpha_t^2}{2} \left\| z_{t+1} - z_t \right\|^2 \\
&\leq \left(1 - \alpha_t\right)\left(f\left(y_t\right) - f\left(x^*\right)\right) + \alpha_t \left\langle \theta_t, x^* - z_{t+1} \right\rangle \\
&\quad + \alpha_t \left\langle \widetilde{\nabla} f(x_t), z_{t+1} - x^* \right\rangle + \frac{L\alpha_t^2}{2} \left\| z_{t+1} - z_t \right\|^2
\end{aligned}$$

根据最优性条件，我们有

$$\left\langle \eta_t \widetilde{\nabla} f(x_t) + \nabla_x \mathbf{D}_\psi\left(z_{t+1}, z_t\right), x^* - z_{t+1} \right\rangle \geq 0$$

因此

$$\left\langle \eta_t \widetilde{\nabla} f(x_t), z_{t+1} - x^* \right\rangle \leq \left\langle \nabla_x \mathbf{D}_\psi\left(z_{t+1}, z_t\right), x^* - z_{t+1} \right\rangle$$

注意，

$$\begin{aligned}
\left\langle \nabla_x \mathbf{D}_\psi\left(z_{t+1}, z_t\right), x^* - z_{t+1} \right\rangle &= \left\langle \nabla\psi\left(z_{t+1}\right) - \nabla\psi\left(z_t\right), x^* - z_{t+1} \right\rangle \\
&= \mathbf{D}_\psi\left(x^*, z_t\right) - \mathbf{D}_\psi\left(z_{t+1}, z_t\right) - \mathbf{D}_\psi\left(x^*, z_{t+1}\right)
\end{aligned}$$

Thus

$$\begin{aligned}
\eta_t \left\langle \widetilde{\nabla} f(x_t), z_{t+1} - x^* \right\rangle &\leq \mathbf{D}_\psi\left(x^*, z_t\right) - \mathbf{D}_\psi\left(x^*, z_{t+1}\right) - \mathbf{D}_\psi\left(z_{t+1}, z_t\right) \\
&\leq \mathbf{D}_\psi\left(x^*, z_t\right) - \mathbf{D}_\psi\left(x^*, z_{t+1}\right) - \frac{1}{2} \left\| z_{t+1} - z_t \right\|^2
\end{aligned}$$

其中我们利用了根据 ψ 的强凸性，$\mathbf{D}\psi\left(z_{t+1}, z_t\right) \geq \frac{1}{2}\left\|z_{t+1}-z_t\right\|^2$。

$$
\begin{aligned}
f\left(y_{t+1}\right)-f\left(x^*\right) \leq & \left(1-\alpha_t\right)\left(f\left(y_t\right)-f\left(x^*\right)\right)+\alpha_t\left\langle\theta_t, x^*-z_{t+1}\right\rangle \\
& +\frac{\alpha_t}{\eta_t} \mathbf{D}_\psi\left(x^*, z_t\right)-\frac{\alpha_t}{\eta_t} \mathbf{D}_\psi\left(x^*, z_{t+1}\right)+\left(\frac{L \alpha_t^2}{2}-\frac{\alpha_t}{2 \eta_t}\right)\left\|z_{t+1}-z_t\right\|^2
\end{aligned}
$$

两边除以$\frac{\alpha_t}{\eta_t}$，并利用条件 $L \eta_t \alpha_t \leq \frac{1}{2}$，可得

$$
\begin{aligned}
\frac{\eta_t}{\alpha_t}\left(f\left(y_{t+1}\right)-f\left(x^*\right)\right)+\mathbf{D}_\psi\left(x^*, z_{t+1}\right)-\mathbf{D}_\psi\left(x^*, z_t\right) \leq & \frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)+\eta_t\left\langle\theta_t, x^*-z_t\right\rangle \\
& +\eta_t\left\langle\theta_t, z_t-z_{t+1}\right\rangle-\frac{1-L \eta_t \alpha_t}{2}\left\|z_{t+1}-z_t\right\|^2 \\
\leq & \frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)+\eta_t\left\langle\theta_t, x^*-z_t\right\rangle \\
& +\frac{\eta_t^2\left\|\theta_t\right\|_*^2}{2\left(1-L \eta_t \alpha_t\right)} \\
\leq & \frac{\eta_t\left(1-\alpha_t\right)}{\alpha_t}\left(f\left(y_t\right)-f\left(x^*\right)\right)+\eta_t\left\langle\theta_t^u+\theta_t^b, x^*-z_t\right\rangle \\
& +2 \eta_t^2\left\|\theta_t^u\right\|_*^2+2 \eta_t^2\left\|\theta_t^b\right\|_*^2
\end{aligned}
$$

视需要而定。 □

定理 B.1. 假设 f满足假设 *(1)*、*(2)*、*(3)*、*(4)* 和 *(5')*。令 $\gamma=$最大值 $\left\{\text{对数 } \frac{1}{\delta};\ \ 1\right\}$；并且$R_1=\sqrt{2} \mathbf{D} \psi\left(x^*\right.$, $\left.x_1\right)$。对于未知的 T，我们选择 c_t、 λ_t 和 η_t，使得

$$
\begin{aligned}
c_t & =\max \left\{10^4 ; \frac{4(t+1)\left(\frac{52 t(1+\log t)^2}{\gamma}\right)^{\frac{1}{p}} \sigma}{\gamma L R_1}\right\} \\
\lambda_t & =\frac{c_t R_1 \gamma L \alpha_t}{8}=\max \left\{\frac{10^4 R_1 \gamma L}{4(t+1)} ;\left(\frac{52 t(1+\log t)^2}{\gamma}\right)^{1 / p} \sigma\right\} \\
\eta_t & =\frac{1}{3 c_t \gamma^2 L \alpha_t}=\frac{R_1}{24 \gamma} \min \left\{\frac{4(t+1)}{10^4 R_1 \gamma L} ;\left(\frac{52 t(1+\log t)^2}{\gamma}\right)^{-1 / p} \sigma^{-1}\right\}
\end{aligned}
$$

那么，以至少 $1-\delta$ 的概率

$$
f\left(y_{T+1}\right)-f\left(x^*\right) \leq 6 \max \left\{10^4 L \gamma^2 R_1^2(T+1)^{-2} ; 4 R_1(T+1)^{-1}\left(52 t(1+\log t)^2\right)^{\frac{1}{p}} \gamma^{\frac{p-1}{p}} \sigma\right\} .
$$

证明。 按照定理 4.1 证明中的类似步骤，并注意到 (c_t) 是一个递增序列，我们得到收敛速率。

□

C 第 5 节中遗漏的证明

引理 *5.2* 的证明。 由 f 的光滑性以及更新 $x_{t+1}=x_t-\frac{1}{\eta_t} \widetilde{\nabla} f(x_t)$，我们有

$$
\begin{aligned}
& f(x_{t+1}) - f(x_t) \\
\leq & \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
= & -\eta_t \left\langle \nabla f(x_t), \widetilde{\nabla} f(x_t) \right\rangle + \frac{L\eta_t^2}{2} \left\| \widetilde{\nabla} f(x_t) \right\|^2 \\
= & -\eta_t \langle \nabla f(x_t), \theta_t + \nabla f(x_t) \rangle + \frac{L\eta_t^2}{2} \|\theta_t + \nabla f(x_t)\|^2 \\
= & -\eta_t \|\nabla f(x_t)\|^2 - \eta_t \langle \nabla f(x_t), \theta_t \rangle + \frac{L\eta_t^2}{2} \|\theta_t\|^2 + \frac{L\eta_t^2}{2} \|\nabla f(x_t)\|^2 + L\eta_t^2 \langle \nabla f(x_t), \theta_t \rangle \\
= & -\left(\eta_t - \frac{L\eta_t^2}{2}\right) \|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2} \|\theta_t\|^2 + \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t \rangle \\
= & -\left(\eta_t - \frac{L\eta_t^2}{2}\right) \|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2} \|\theta_t\|^2 + \underbrace{\left(L\eta_t^2 - \eta_t\right)}_{\leq 0} \left\langle \nabla f(x_t), \theta_t^u + \theta_t^b \right\rangle .
\end{aligned}
$$

利用柯西-施瓦茨不等式，我们有 $\left\langle \nabla f(x_t), \theta_t^b \right\rangle \leq \frac{1}{2} \|\nabla f(x_t)\|^2 + \frac{1}{2} \left\|\theta_t^b\right\|^2$ 因此，推得

$$
\begin{aligned}
\Delta_{t+1} - \Delta_t \leq & -\left(\frac{2\eta_t - L\eta_t^2}{2}\right) \|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2} \|\theta_t\|^2 + \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t^u \rangle \\
& + \frac{\eta_t - L\eta_t^2}{2} \|\nabla f(x_t)\|^2 + \frac{\eta_t - L\eta_t^2}{2} \left\|\theta_t^b\right\|^2 \\
\leq & -\frac{\eta_t}{2} \|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2} \|\theta_t\|^2 + \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t^u \rangle + \frac{\eta_t}{2} \left\|\theta_t^b\right\|^2 \\
\leq & -\frac{\eta_t}{2} \|\nabla f(x_t)\|^2 + L\eta_t^2 \|\theta_t^u\|^2 + \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t^u \rangle + \left(L\eta_t^2 + \frac{\eta_t}{2}\right) \left\|\theta_t^b\right\|^2 \\
\leq & -\frac{\eta_t}{2} \|\nabla f(x_t)\|^2 + L\eta_t^2 \|\theta_t^u\|^2 + \left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t^u \rangle + \frac{3\eta_t}{2} \left\|\theta_t^b\right\|^2
\end{aligned}
$$

其中，最后一个不等式由 $\eta_t \leq \frac{1}{L}$ 得出。重新整理后，即得引理。 □

引理 *5.3* 的证明。我们有

$$
\begin{aligned}
& \mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \exp\left(\left(3z_t^2 L\eta_t^2 \Delta_t + 6L^2 z_t^2 \eta_t^4 \lambda_t^2\right) \mathbb{E}\left[\left\|\theta_t^u\right\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
\overset{(a)}{\leq} & \mathbb{E}\left[\exp\left(z_t \left(\left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t^u \rangle + L\eta_t^2 \left(\left\|\theta_t^u\right\|^2 - \mathbb{E}\left[\left\|\theta^u\right\|^2 \mid \mathcal{F}_{t-1}\right]\right)\right)\right) \mid \mathcal{F}_{t-1}\right] \\
\overset{(b)}{\leq} & \exp\left(\mathbb{E}\left[\frac{3}{4}\left(z_t \left(\left(L\eta_t^2 - \eta_t\right) \langle \nabla f(x_t), \theta_t^u \rangle + L\eta_t^2 \left(\left\|\theta_t^u\right\|^2 - \mathbb{E}\left[\left\|\theta^u\right\|^2 \mid \mathcal{F}_{t-1}\right]\right)\right)\right)^2 \mid \mathcal{F}_{t-1}\right]\right) \\
\overset{(c)}{\leq} & \exp\left(\mathbb{E}\left[\frac{3}{2} z_t^2 \eta_t^2 \|\nabla f(x_t)\|^2 \left\|\theta_t^u\right\|^2 \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[\frac{3}{2} L^2 z_t^2 \eta_t^4 \left\|\theta_t^u\right\|^4 \mid \mathcal{F}_{t-1}\right]\right) \\
\overset{(d)}{\leq} & \exp\left(3z_t^2 L\eta_t^2 \Delta_t \mathbb{E}\left[\left\|\theta_t^u\right\|^2 \mid \mathcal{F}_{t-1}\right] + 6L^2 z_t^2 \eta_t^4 \lambda_t^2 \mathbb{E}\left[\left\|\theta_t^u\right\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
= & \exp\left(\left(3z_t^2 L\eta_t^2 \Delta_t + 6L^2 z_t^2 \eta_t^4 \lambda_t^2\right) \mathbb{E}\left[\left\|\theta_t^u\right\|^2 \mid \mathcal{F}_{t-1}\right]\right)
\end{aligned}
$$

对于 (*a*)，我们使用引理 5.2。对于 (*b*)，我们使用引理 2.2。注意，

$$
\mathbb{E}\left[\langle \nabla f(x_t), \theta_t^u \rangle\right] = \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 - \mathbb{E}\left[\left\|\theta_t^u\right\|_*^2 \mid \mathcal{F}_{t-1}\right]\right] = 0
$$

并且由于对于一个 L-光滑函数， $\|\theta_t^u\| \le 2\lambda_t$ 和 $\|\nabla f(x_t)\| \le \sqrt{2L\Delta_t}$，我们有

$$\begin{aligned}
&\left|\left(L\eta_t^2 - \eta_t\right)\langle \nabla f(x_t), \theta_t^u\rangle + L\eta_t^2\left(\|\theta_t^u\|^2 - \mathbb{E}\left[\|\theta^u\|^2 \mid \mathcal{F}_{t-1}\right]\right)\right| \\
\le& 2\eta_t\lambda_t\|\nabla f(x_t)\| + L\eta_t^2\left(\|\theta_t^u\|^2 + \mathbb{E}\left[\|\theta^u\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
\le& 2\eta_t\lambda_t\|\nabla f(x_t)\| + 8L\eta_t^2\lambda_t^2 \\
\le& 2\eta_t\lambda_t\sqrt{2L\Delta_t} + 8L\eta_t^2\lambda_t^2
\end{aligned}$$

因此 $z_t \le \frac{1}{2\eta_t\lambda_t\sqrt{2L\Delta_t}+8L\eta_t^2\lambda_t^2}$。对于 (c)，我们使用 $(a+b)^2 \le 2a^2+2b^2$ 以及 $\mathbb{E}\left[(X-\mathbb{E}[X])^2\right] \le \mathbb{E}\left[X^2\right]$。对于 (d)，我们使用 $\|\nabla f(x_t)\|^2 \le 2L\Delta_t$ 以及 $\|\theta_t^u\| \le 2\lambda_t$。于是得到

$$\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \le 1$$

因此

$$\begin{aligned}
\mathbb{E}\left[\exp\left(S_t\right) \mid \mathcal{F}_{t-1}\right] &= \exp\left(S_{t-1}\right)\mathbb{E}\left[\exp\left(Z_t\right) \mid \mathcal{F}_{t-1}\right] \\
&\le \exp\left(S_{t-1}\right)
\end{aligned}$$

这意味着 $(S_t)_{t\ge 1}$ 是一个超鞅。根据 Ville 不等式，对所有 $k \ge 1$，我们有

$$\Pr\left[S_k \ge \log\frac{1}{\delta}\right] \le \delta\mathbb{E}\left[\exp\left(S_1\right)\right] \le \delta$$

换言之，以至少 $1-\delta$ 的概率，对所有 $k \ge 1$

$$\sum_{t=1}^{k} Z_t \le \log\frac{1}{\delta}$$

代入 Z_t 的定义，我们得到

$$\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{k} z_t\eta_t\|\nabla f(x_t)\|^2 + \sum_{t=1}^{k}\left(z_t\Delta_{t+1} - z_t\Delta_t\right) \\
\le& \log\frac{1}{\delta} + \sum_{t=1}^{k}\frac{3z_t\eta_t}{2}\left\|\theta_t^b\right\|^2 \\
&+ \sum_{t=1}^{k}\left(\left(3z_t^2L\eta_t^2\Delta_t + 6L^2z_t^2\eta_t^4\lambda_t^2 + z_tL\eta_t^2\right)\mathbb{E}\left[\|\theta_t^u\|^2 \mid \mathcal{F}_{t-1}\right]\right)
\end{aligned}$$

注意到 z_t 是一个递减序列，因此上述不等式左端可由以下量界定

$$\begin{aligned}
\text{LHS} &= \frac{1}{2}\sum_{t=1}^{k} z_t\eta_t\|\nabla f(x_t)\|^2 + z_k\Delta_{k+1} - z_1\Delta_1 + \sum_{t=2}^{k}\left(z_{k-1}-z_k\right)\Delta_k \\
&\ge \frac{1}{2}\sum_{t=1}^{k} z_t\eta_t\|\nabla f(x_t)\|^2 + z_k\Delta_{k+1} - z_1\Delta_1
\end{aligned}$$

我们得到了所需不等式。 □

命题 *5.4* 的证明。 我们将对 k 进行归纳法证明：

$$\frac{1}{2}\sum_{i=1}^{k}\eta_i\|\nabla f(x_i)\|^2 + \Delta_{k+1} \le \left(\sqrt{\Delta_1} + 2\sqrt{A}C_1\right)^2$$

基例 $k=0$ 是显然的。假设该命题对所有 $t \le k \le \ell$ 均成立。现在证明它对 $k+1$ 也成立。回顾:

$$z_t = \frac{1}{2P_t\eta_t\lambda_t \max_{i\le t}\sqrt{2L\Delta_i} + 8Q_tL\eta_t^2\lambda_t^2}$$

我们选择

$$P_t = \frac{C_1}{\lambda_t\eta_t\sqrt{2L}} \ge 1$$
$$Q_t = \frac{C_1^2\sqrt{A}}{2L\eta_t^2\lambda_t^2} \ge 1$$

我们有

$$z_t = \frac{1}{2C_1 \max_{i\le t}\sqrt{\Delta_i} + 4C_1^2\sqrt{A}}$$

现在我们可以注意到，$(z_t)_{t\ge 1}$ 是一个递减序列。根据归纳假设

$$\begin{aligned}\frac{z_t}{z_k} &\le \frac{2C_1\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)+4C_1^2\sqrt{A}}{2C_1\sqrt{\Delta_1}+4C_1^2\sqrt{A}}\\ &= \frac{\sqrt{\Delta_1}+4\sqrt{A}C_1}{\sqrt{\Delta_1}+2\sqrt{A}C_1} \le 2\end{aligned}$$

根据 λ_t 的选择，对于所有 $t\le k$, $\|\nabla f(x_t)\| \le \frac{\lambda_t}{2}$，我们可以应用引理 2.1，得到 e

$$\begin{aligned}\left\|\theta_t^b\right\| &\le 4\sigma^p\lambda_t^{1-p};\\ \mathbb{E}\left[\left\|\theta_t^u\right\|^2 \mid \mathcal{F}_{t-1}\right] &\le 40\sigma^p\lambda_t^{2-p}.\end{aligned}$$

Thus

$$\begin{aligned}&\frac{1}{2}z_k\sum_{t=1}^{k}\eta_t\left\|\nabla f(x_t)\right\|^2 + z_k\Delta_{k+1}\\ \le& z_1\Delta_1 + \log\frac{1}{\delta} + \sum_{t=1}^{k}\frac{3z_t\eta_t}{2}\left\|\theta_t^b\right\|^2\\ &+\sum_{t=1}^{k}\left(\left(3z_t^2L\eta_t^2\Delta_t + 6L^2z_t^2\eta_t^4\lambda_t^2 + z_tL\eta_t^2\right)\mathbb{E}\left[\left\|\theta_t^u\right\|^2 \mid \mathcal{F}_{t-1}\right]\right)\\ \le& z_1\Delta_1 + \log\frac{1}{\delta} + 24\sigma^{2p}\sum_{t=1}^{k}z_t\eta_t\lambda_t^2\left(\frac{1}{\lambda_t}\right)^{2p}\\ &+40\sigma^p\sum_{t=1}^{k}\left(\left(3z_t^2\Delta_t + 6z_t^2L\eta_t^2\lambda_t^2 + z_t\right)L\eta_t^2\lambda_t^2\left(\frac{1}{\lambda_t}\right)^{p}\right)\end{aligned}$$

由于 $\frac{z_t}{z_k} \leq 2$，我们有

$$
\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{k}\eta_t \left\|\nabla f(x_t)\right\|^2 + \Delta_{k+1} \\
\leq &\frac{z_1\Delta_1}{z_k} + \frac{1}{z_k}\log\frac{1}{\delta} + 48\sigma^{2p}\sum_{t=1}^{k}\eta_t\lambda_t^2\left(\frac{1}{\lambda_t}\right)^{2p} \\
&+ 80\sigma^p\sum_{t=1}^{k}\left(\left(3z_t\Delta_t + 6z_tL\eta_t^2\lambda_t^2 + 1\right)L\eta_t^2\lambda_t^2\left(\frac{1}{\lambda_t}\right)^p\right) \\
\overset{(a)}{\leq} &\frac{\sqrt{\Delta_1}+4\sqrt{A}C_1}{\sqrt{\Delta_1}+2\sqrt{A}C_1}\Delta_1 + 2C_1\left(\sqrt{\Delta_1}+4\sqrt{A}C_1\right)\log\frac{1}{\delta} + 48\sigma^{2p}C_2\sum_{t=1}^{k}L\eta_t^2\lambda_t^2\left(\frac{1}{\lambda_t}\right)^p \\
&+ 80\sigma^p\sum_{t=1}^{k}\left(\left(\frac{3\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)^2}{2C_1\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)} + \frac{6}{8Q_t} + 1\right)L\eta_t^2\lambda_t^2\left(\frac{1}{\lambda_t}\right)^p\right) \\
\overset{(b)}{\leq} &\Delta_1 + 2\sqrt{\Delta_1}\sqrt{A}C_1 + 2C_1\left(\sqrt{\Delta_1}+4\sqrt{A}C_1\right)\log\frac{1}{\delta} + 48\sigma^{2p}C_2C_3 \\
&+ 80\sigma^p\left(\frac{3\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)}{2C_1} + \frac{7}{4}\right)C_3 \\
\leq &\Delta_1 + 2\sqrt{\Delta_1}\sqrt{A}C_1 + 2C_1\left(\sqrt{\Delta_1}+4\sqrt{A}C_1\right)\left(\log\frac{1}{\delta} + \frac{60\sigma^pC_3}{C_1^2}\right) \\
&+ 48\sigma^{2p}C_2C_3 + 140\sigma^pC_3 \\
\overset{(c)}{\leq} &\Delta_1 + 2\sqrt{\Delta_1}\sqrt{A}C_1 + 2C_1\left(\sqrt{\Delta_1}+4\sqrt{A}C_1\right)\frac{\sqrt{A}}{8} + AC_1^2 \\
\leq &\left(\sqrt{\Delta_1}+2\sqrt{A}C_1\right)^2
\end{aligned}
$$

对于（a），我们使用 $\left(\frac{1}{\lambda_t}\right)^p \leq C_2L\eta_t$ 和归纳假设。对于（b），我们使用 $\sum_{t=1}^{T}L\left(\frac{1}{\lambda_t}\right)^p\lambda_t^2\eta_t^2 \leq C_3$ 和 $Q_t \geq 1$。对于（c），我们有

$$
\begin{aligned}
\log\frac{1}{\delta} + \frac{60\sigma^pC_3}{C_1^2} &\leq \frac{\sqrt{A}}{8} \\
48\sigma^{2p}C_2C_3 + 140\sigma^pC_3 &\leq AC_1^2
\end{aligned}
$$

因为

$$
A \geq 64\left(\log\frac{1}{\delta} + \frac{60\sigma^pC_3}{C_1^2}\right)^2 + \frac{48\sigma^{2p}C_2C_3 + 140\sigma^pC_3}{C_1^2}.
$$

证明至此完成。 □

引理 C.1。 定理 *5.1* 中 η_t 和 λ_t 的取值满足命题 *5.4* 的条件 *(1)-(3)*，因为

$$
\begin{aligned}
C_1 &= \frac{\sqrt{\Delta_1}}{4\sqrt{2}\gamma} \\
C_2 &= \frac{1}{\sigma^p} \\
C_3 &= \frac{\Delta_1}{2048\sigma^p\gamma}.
\end{aligned}
$$

证明。We 验证第一种情况。第二种情况完全相同。首先，我们有 $p>1$ hen ce

$$\eta_t\lambda_t\sqrt{2L}=\frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\sqrt{L}\gamma}\sqrt{2L}\leq\frac{\sqrt{\Delta_1}}{4\sqrt{2}\gamma}=C_1$$

由于 $\eta_t=\frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\lambda_t\sqrt{L}\gamma}$、 $p>1$ 和 $\lambda_t\geq\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}}T^{\frac{1}{3p-2}}\sigma^{\frac{p}{p-1}}$

$$\begin{aligned}\eta_t\lambda_t^p&=\frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\sqrt{L}\gamma}\lambda_t^{p-1}\\&\geq\frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\sqrt{L}\gamma}\frac{8\gamma}{\sqrt{L\Delta_1}}T^{\frac{p-1}{3p-2}}\sigma^p\\&=\frac{\sigma^p}{L}\end{aligned}$$

这给出

$$\frac{1}{L\eta_t}\left(\frac{1}{\lambda_t}\right)^p\leq\frac{1}{\sigma^p}=C_2.$$

最后，我们有 $\lambda_t\geq 32^{1/p}\sigma T^{\frac{1}{3p-2}}$，因此

$$\left(\frac{1}{\lambda_t}\right)^p T^{\frac{p}{3p-2}}\leq\frac{1}{32\sigma^p}.$$

因此，

$$\begin{aligned}\sum_{t=1}^T L\left(\frac{1}{\lambda_t}\right)^p\lambda_t^2\eta_t^2&=\sum_{t=1}^T L\left(\frac{1}{\lambda_t}\right)^p\left(\frac{\sqrt{\Delta_1}T^{\frac{1-p}{3p-2}}}{8\sqrt{L}\gamma}\right)^2\\&\leq TL\left(\frac{1}{\lambda_t}\right)^p T^{\frac{2-2p}{3p-2}}\frac{\Delta_1}{64L\gamma}\\&=\left(\frac{1}{\lambda_t}\right)^p T^{\frac{p}{3p-2}}\frac{\Delta_1}{64\gamma^2}\\&\leq\frac{1}{32\sigma^p}\frac{\Delta_1}{64\gamma^2}\leq\frac{\Delta_1}{2048\sigma^p\gamma}.\end{aligned}$$

□

定理 C.2。 假设 f 满足假设 *(1')*、*(2)*、*(3)*、*(4)*。令 $\gamma=$最大 $\left\{\text{对数}\frac{1}{\delta};\ 1\right\}$ 并令 $\Delta_1=f(x_1)-f^*$。对于未知的 T，我们选择 λ_t 和 η_t，使得

$$\lambda_t=\max\left\{\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}}\left(2t\left(1+\log t\right)^2\right)^{\frac{1}{3p-2}}\sigma^{\frac{p}{p-1}};2\sqrt{90L\Delta_1};32^{\frac{1}{p}}\sigma\left(2t\left(1+\log t\right)^2\right)^{\frac{1}{3p-2}}\right\}$$

$$\eta_t=\frac{\sqrt{\Delta_1}\left(2t\left(1+\log t\right)^2\right)^{\frac{1-p}{3p-2}}}{8\lambda_t\sqrt{L}\gamma}$$

那么，以至少 $1-\delta$ 的概率

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T\left\|\nabla f(x_t)\right\|^2\leq 720\sqrt{\Delta_1 L}\gamma\max\Bigg\{&\left(\frac{8\gamma}{\sqrt{L\Delta_1}}\right)^{\frac{1}{p-1}}\left(2\left(1+\log T\right)^2\right)^{\frac{p}{3p-2}}\sigma^{\frac{p}{p-1}}T^{\frac{2-2p}{3p-2}};\\&2\sqrt{90L\Delta_1}\left(2\left(1+\log T\right)^2\right)^{\frac{p-1}{3p-2}}T^{\frac{1-2p}{3p-2}};32^{\frac{1}{p}}\sigma\left(2\left(1+\log T\right)^2\right)^{\frac{p}{3p-2}}T^{\frac{2-2p}{3p-2}}\Bigg\}.\end{aligned}$$

证明。 完全按照已知 T 时的相同步骤，并注意到 η_t 是递减的，我们即可得到收敛保证。

□