基于动量的非凸随机梯度下降的高概率界限

李少杰 刘 勇
（中国人民大学高瓴人工智能学院 北京 100872）
（北京市大数据管理与分析方法重点实验室 北京 100872）

摘 要 基于动量的随机梯度下降（Stochastic Gradient Descent with Momentum，SGDM）在机器学习中得到了广泛应用，但其理论性质尚缺乏深入理解。在非凸领域，现有文献对SGDM的分析主要集中在期望意义上，而高概率的分析相对较少。高概率结果的重要性在于它适用于样本空间中的最坏情况。针对这一问题，本文为SGDM提供了高概率的收敛界限和泛化界限，推导出的收敛界限与现有的期望结果相匹配，并且据我们所知，推导出的泛化界限是SGDM的首次提出。此外，同时考虑收敛和泛化有助于理解SGDM在实际应用中的优良性能，本文的理论结果解释了两个新近提出的SGDM算法的优越性。最后，本文通过数值实验验证了理论分析所用假设的合理性，并且验证了所用假设如何影响泛化界限的变化速率。

关键词 随机梯度下降；优化界限；泛化界限；非凸优化
中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2025.00763

High Probability Bounds of Non-Convex Stochastic Gradient Descent with Momentum

LI Shao-Jie LIU Yong
（Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872）
（Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing 100872）

Abstract Stochastic gradient descent with momentum (SGDM) has been widely used in machine learning, but its theoretical properties are poorly understood. In non-convex domains, the existing literature mainly focuses on the analysis of SGDM in expectation, while the analysis in high probability is relatively scarce. The importance of high-probability results is that they hold for the worst-case scenario in the sample space. To address this problem, this paper provides high-probability convergence and generalization bounds for SGDM. The derived convergence bounds match the existing in-expectation results, and to our best knowledge, the derived generalization bounds are the first to be proposed for SGDM. In addition, considering both convergence and generalization helps to understand the excellent empirical performance of SGDM in practice, and our theoretical results explain the superiority of two recently proposed SGDM algorithms. Finally, this paper validates the reasonableness of the assumptions used in the theoretical analysis through numerical experiments and examines how these assumptions influence the variation rate of the generalization bounds.

Keywords stochastic gradient descent; optimization bound; generalization bound; non-convex optimization

收稿日期：2024-09-06；在线发布日期：2025-02-20。本课题得到国家自然科学基金面上项目（62476277）、国家重点研发计划（2024YFE0203200）和CCF-阿里妈妈科技袋基金（CCF-ALIMAMA OF 2024008）资助。**李少杰**，博士，主要研究领域为机器学习理论、大模型理论。E-mail：2020000277@ruc.edu.cn。 **刘 勇**（通信作者），博士，副教授，中国计算机学会（CCF）会员，主要研究领域为大模型学习算法与理论、机器学习基础理论。E-mail：liuyonggsai@ruc.edu.cn。

1 引 言

随机优化在现代机器学习和数据驱动的优化中起着至关重要的作用，因为许多机器学习问题可以转化为随机优化问题[1]。在过去几十年中，随机优化算法取得了显著进展，其中SGDM算法因其简单性及每次更新的低计算复杂度而在自然语言理解、计算机视觉和语音识别等众多学习任务中取得了巨大成功[2-3]。SGDM通常在每次迭代更新时向随机梯度下降（Stochastic Gradient Descent，SGD）方法中添加一个动量项，即当前迭代点与上一次迭代点之间的差异。这种设计的直观想法是，如果从上一次迭代到当前迭代的方向是合理的，SGDM会利用动量参数带来的惯性效应。SGDM的优越经验性能吸引了众多研究者对其理论性质进行深入探讨。

SGDM的理论性质可以从收敛界限和泛化界限两个角度进行研究。前者关注学习算法如何根据训练样本优化模型，而后者则关注从训练样本学习到的模型在测试样本上的表现[4-5]。从收敛界限的角度来看，现有文献对于SGDM在非凸领域的收敛性研究大多是期望分析[6]。然而，期望的界限仅能提供一种平均性能的见解，并不能充分捕捉到随机优化算法在单次或少数几次运行中的表现，这与随机优化算法的概率特性密切相关。此外，期望的界限也不能排除极端不良结果的存在[7]。在诸如神经网络等实际应用场景中，由于训练过程通常需要数小时乃至数天，算法往往仅运行一次，因此在单次运行的算法理论研究中，高概率界限相比期望界限更具吸引力[7]。据我们所知，目前只有两项研究探讨了SGDM的高概率收敛界限[6,8]。从泛化界限的角度来看，关于SGDM的泛化研究相对稀缺。在非凸领域中尚未建立相应的高概率界限。因此，为了理解SGDM在实际应用中展现出的优秀性能，高概率界限的分析显得尤为迫切。

基于此，本文考虑在非凸设置下建立SGDM的高概率收敛界限和泛化界限。本文研究两种新近提出的SGDM算法，一个使用隐式梯度传输更新（Implicit Gradient Transport）[9]，另一个使用海森校正动量（Hessian-Corrected Momentum）[10]。这两种SGDM算法通过结合二阶光滑性（smoothness）能实现比基本的SGDM算法[11]更好的收敛性能，突破$\mathcal{O}(1/T^{1/4})$的限制，因为对于具有光滑损失的一阶随机优化算法而言，$\mathcal{O}(1/T^{1/4})$的收敛界限已经被证实为最优的[12]，其中T是迭代次数。

本文的贡献可以总结如下：对于带有隐式梯度传输更新的SGDM，我们提供了$\tilde{\mathcal{O}}(1/T^{2/7})$阶数的收敛界限，并且给出了$\tilde{\mathcal{O}}(d^{1/4}/n^{1/4})$阶数的泛化界限，这个泛化界限与SGD算法的相同[13]，其中d是模型维度，n是训练样本个数。对于使用海森校正动量的SGDM，我们进一步将收敛界限从$\tilde{\mathcal{O}}(1/T^{2/7})$改善到$\tilde{\mathcal{O}}(1/T^{1/3})$，并给出了相同阶数的泛化界限。值得注意的是，本文的证明思路同样适用于基本的SGDM算法[11]。相比于基本的SGDM算法，本文的理论结果分析揭示了这两种SGDM算法在保持泛化性能的同时，能够实现更快的收敛速率。同时，本文的分析揭示了这两种算法具有更优的迭代复杂度。数值实验证实了我们的理论发现。

2 相关研究

本节回顾SGDM的收敛分析和泛化分析的相关工作。

2.1 SGDM的收敛分析

动量方法历史悠久，最早由Polyak引入[11]，其通过将前一次更新的加权版本与当前的梯度更新相结合。动量方法的初衷是为了加速在凸优化中梯度下降的收敛速度。随后的研究[14]讨论了动量方法在非凸优化中的重要性。由于该方法在实践中取得的显著成功，许多研究者致力于对其收敛性进行分析，不过这些研究主要集中于期望的收敛界限上[6]。尽管SGDM在非凸领域收敛性的文献多集中在期望分析上，但这些工作为SGDM提供了非常有趣的理论见解。比如，近年来已有许多工作开始研究动量如何减少梯度估计的方差来提高收敛的稳定性[15-16]。还有一些研究提出了统一的分析框架，在分析SGDM收敛界限的同时，考虑了方差的影响[17]、算法在不同步长[18]和加速条件下[19]的表现。近期还有研究进一步探讨了收敛界限，分析了动量在不同加速算法（如AdaGrad、Adam等）下的收敛性能，通过严谨的理论框架系统性地研究了SGDM的收敛界限[20]。然而，关于SGDM在非凸领域的高概率收敛分析，目前仅有两项相关研究[6,8]。具体而言，文献[6]在次高斯梯度噪声（轻尾条件）下研究了Polyak动量，文献[8]研究了动量和梯度截断。这两项研究均提供了$\tilde{\mathcal{O}}(1/T^{1/4})$阶的收敛界限。文献[6]

中讨论到，尚不清楚这一收敛速率是否可以改善并扩展到超出次高斯梯度噪声的更一般设置。本文考虑重尾的随机梯度噪声设置（扩展次高斯梯度噪声至次威布尔噪声），并通过研究新近提出的SGDM算法给出了高概率的快速收敛速率。

2.2　SGDM的泛化分析

与收敛分析相比，动量方法的泛化界限研究相对较少[21]。文献[22-23]的分析仅限于平方损失函数，且难以扩展到一般损失函数。针对此，文献[24]通过算法稳定性工具研究了一般的凸损失函数。多年来，算法稳定性已经成为学习算法泛化分析的标准工具[25]。然而，文献[24]的分析显示，存在某些凸损失函数使得SGDM的泛化界限在运行多个周期后变得无界。同样，文献[26]为Nesterov加速梯度算法（NAG）建立了紧的算法稳定性界限。该研究同样表明，在一般的凸光滑设置中，NAG的算法稳定性随着迭代次数的增加呈指数增长。具体来说，NAG的算法稳定性下界为$\Omega(\exp(T)/n)$，这表明在经过$\mathcal{O}(\log n)$步之后此算法稳定性界限变成无意义的$\Omega(1)$，即随迭代次数的增加以指数速度崩溃。综上所述，文献[24,26]都表明，SGDM很容易表现出算法的不稳定性。因此，关于SGDM在一般损失函数下的泛化界限仍然缺失。本文旨在研究这一问题，并在非凸设置下以高概率的形式给出SGDM的泛化界限。

3　主要结果

本节先介绍相关的问题设置，包含一些必要的数学符号以及假设，然后介绍本文的理论成果，包括隐式梯度传输的SGDM以及海森校正动量的SGDM，最后总结和比较相关的结果。

3.1　问题设置

设$\mathcal{X}$是$\mathbb{R}^d$中的一个参数空间，$\mathbb{P}$是定义在样本空间$\mathcal{Z}$上的概率分布。定义函数$f:\mathcal{X}\times\mathcal{Z}\mapsto\mathbb{R}_+$，本文考虑以下随机优化

$$\min_{x\in\mathcal{X}} F(\boldsymbol{x}):=\mathbb{E}_{z\sim\mathbb{P}}[f(\boldsymbol{x};z)],$$

其中，F通常称为总体风险，f可能是非凸的，$\mathbb{E}_{z\sim\mathbb{P}}$表示关于从$\mathbb{P}$中抽取的随机变量$z$的期望。在实践中，$\mathbb{P}$通常是不可知的，我们得到的是从$\mathbb{P}$中以独立同分布方式采样的数据集$S=\{z_1,z_2,\cdots,z_n\}$。因此，实践中通常优化以下经验风险：

$$\min_{x\in\mathcal{X}} F_S(\boldsymbol{x}):=\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x};z_i)$$

SGDM被广泛采用优化经验风险$F_S(\boldsymbol{x})$[14]，其伪代码见算法1。

算法1.　SGDM

输入：步长$\{\eta_t\}_t$，数据集$S=\{z_1,z_2,\cdots,z_n\}$和动量参数γ

初始化：$\boldsymbol{x}_1=0,\boldsymbol{m}_0=0$

1. FOR $t=1,2,\cdots,T$ DO
2. 从集合$\{j:j\in[n]\}$中以均匀分布采样j_t
3. 更新$\boldsymbol{m}_t=\gamma\boldsymbol{m}_{t-1}+(1-\gamma)\nabla f(\boldsymbol{x}_t;z_{j_t})$
4. 更新$\boldsymbol{x}_{t+1}=\boldsymbol{x}_t-\eta_t\boldsymbol{m}_t$
5. END FOR

在算法1的步骤3中，SGDM将动量项$\boldsymbol{m}_{t-1}$与SGD的梯度估计$\nabla f(\boldsymbol{x}_t;z_{j_t})$以动量参数$\gamma$加权后相加。结合步骤4中的更新，SGDM的更新可以理解为

$$\boldsymbol{x}_{t+1}=\boldsymbol{x}_t-\eta_t(1-\gamma)\nabla f(\boldsymbol{x}_t;z_{j_t})+\frac{\eta_t\gamma}{\eta_{t-1}}(\boldsymbol{x}_t-\boldsymbol{x}_{t-1}) \tag{1}$$

3.2　符号和假设

为了研究SGDM的理论性质，需要引入一些必要的符号和假设。令$B=\sup_{z\in\mathcal{Z}}\|\nabla f(0;z)\|$，其中$\nabla f(\cdot;z)$表示$f$关于第一个参数的梯度，而$\|\cdot\|$表示欧几里得范数。对于任意$R>0$，定义$B(\boldsymbol{x}_0,R):=\{\boldsymbol{x}\in\mathbb{R}^d:\|\boldsymbol{x}-\boldsymbol{x}_0\|\leqslant R\}$，它表示以$\boldsymbol{x}_0\in\mathbb{R}^d$为中心，半径为$R$的球体。如果存在常数$c$、$c'>0$使得$ca\leqslant b\leqslant c'a$，我们记$a\asymp b$。本文还将使用标准的数量级符号，如$\mathcal{O}(\cdot)$、$\tilde{\mathcal{O}}(\cdot)$和$\Omega(\cdot)$。

假设1.　可微函数f是一个（可能的）非凸函数，且对于任意$z\in\mathcal{Z}$，$\boldsymbol{x}\mapsto f(\boldsymbol{x};z)$是$L$-光滑的。

一个可微函数$g:\mathcal{X}\mapsto\mathbb{R}$被称为$L$-光滑的当且仅当对于每对$\boldsymbol{x}_1,\boldsymbol{x}_2$，有以下不等式成立

$$\|\nabla g(\boldsymbol{x}_1)-\nabla g(\boldsymbol{x}_2)\|\leqslant L\|\boldsymbol{x}_1-\boldsymbol{x}_2\|,$$

其中∇是梯度算子。光滑性假设在学习理论中是标准的[27]。令$\langle\cdot,\cdot\rangle$为内积，光滑性有两个等价的有用性质[28]：

$$g(\boldsymbol{x}_1)-g(\boldsymbol{x}_2)\leqslant\langle\boldsymbol{x}_1-\boldsymbol{x}_2,\nabla g(\boldsymbol{x}_2)\rangle+\frac{1}{2}L\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2,$$

$$(2L)^{-1}\|\nabla g(\boldsymbol{x})\|^2\leqslant g(\boldsymbol{x})-\inf_{\boldsymbol{x}} g(\boldsymbol{x})。$$

此外，我们引入二阶光滑性假设。

假设2.　F_S是ρ-二阶光滑的。

二阶光滑性是个与光滑性类似的条件。一个可微函数$g:\mathcal{X}\mapsto\mathbb{R}$被称为$\rho$-二阶光滑的当且仅当对于

每对$\boldsymbol{x}_1,\boldsymbol{x}_2$和$\boldsymbol{y}$,有以下不等式成立

$$\left\|\left(\nabla^2 g(\boldsymbol{x}_1)-\nabla^2 g(\boldsymbol{x}_2)\right)\boldsymbol{y}\right\|^2\leqslant\rho\|\boldsymbol{x}_1-\boldsymbol{x}_2\|\|\boldsymbol{y}\|$$

光滑性意味着梯度是Lipschitz连续的,而二阶光滑性则意味着海森是Lipschitz连续的,请参阅著作[29]。

接下来对随机梯度的噪声做出假设。

假设3. 假设存在某个正数K和$\theta\geqslant 1/2$,梯度噪声$\nabla f(\boldsymbol{x}_t;z_{j_t})-\nabla F_S(\boldsymbol{x}_t)$满足

$$\mathbb{E}_{j_t}\left[\exp\left(\left(\frac{\left\|\nabla f\left(\boldsymbol{x}_t;z_{j_t}\right)-\nabla F_S(\boldsymbol{x}_t)\right\|}{K}\right)^{\frac{1}{\theta}}\right)\right]\leqslant 2 \quad (2)$$

文献[6]对随机梯度噪声做出了如下假设:

$$\mathbb{E}_{j_t}\left[\exp\left(\frac{\left\|\nabla f(x_t;z_{j_t})-\nabla F_S(x_t)\right\|^2}{K^2}\right)\right]\leqslant 2 \quad (3)$$

这意味着噪声分布的尾部由高斯分布的尾部主导。相比之下,假设3中的公式(2)将次高斯噪声推广到了更为广泛的分布类别,包括次指数分布(即$\theta=1$)和重尾分布(即$\theta>1$)。更高的尾部参数θ对应于更重的尾部[30-31]。实际上,式(2)中的分布被称为次威布尔(sub-Weibull)分布[30]:若对于某个正数K和θ,随机变量X满足$\mathbb{E}\left[\exp\left(\left(\frac{|X|}{K}\right)^{\frac{1}{\theta}}\right)\right]\leqslant 2$,则称$X$为具有尾部参数$\theta$的次威布尔随机变量,记作$X\sim \mathrm{sub}W(\theta,K)$。因此,本文中的学习界限适用于广泛的重尾分布,并且得出的理论界限能够展示从次高斯/次指数(即轻尾)变量转向重尾变量的过程中,重尾梯度噪声对收敛和泛化速率的影响。此外,近年来的很多工作表明,随机优化算法的噪声往往比次高斯噪声更重[32-33]。例如,有实证研究表明,在全连接神经网络[34]、卷积神经网络[33]以及递归神经网络[35-36]中,梯度噪声通常表现出重尾行为。此外,诸如Bert等大语言模型同样表现出重尾行为[35]。因此,为了更具现实性地分析,研究在重尾条件下随机优化算法的理论性质是必要的。

我们还将需要假设二阶的梯度噪声满足次威布尔分布。

假设4. 假设存在某个正数K'和$\theta\geqslant 1/2$,对于任意向量$\boldsymbol{y}\in\mathbb{R}^d$,二阶的梯度噪声满足

$$\mathbb{E}_{j_t}\left[\exp\left(\left(\frac{\left\|\nabla^2 f\left(\boldsymbol{x}_t;z_{j_t}\right)\boldsymbol{y}-\nabla^2 F_S(\boldsymbol{x}_t)\boldsymbol{y}\right\|}{\|\boldsymbol{y}\|K'}\right)^{\frac{1}{\theta}}\right)\right]\leqslant 2$$

我们将在第4.2节验证假设4的合理性。

3.3 具有隐式梯度传输的SGDM

首先给出算法的伪代码。

算法2. 具有隐式梯度传输的SGDM

输入:步长$\{\eta_t\}_t$,数据集$S=\{z_1,z_2,\cdots,z_n\}$和动量参数γ

初始化:$\boldsymbol{x}_1=0,\boldsymbol{x}_0=0,\boldsymbol{m}_0=0$

1. FOR $t=1,2,\cdots,T$ DO
2. 从集合$\{j:j\in[n]\}$中以均匀分布采样j_t
3. 更新$\boldsymbol{w}_t=\boldsymbol{x}_t+\frac{\gamma}{1-\gamma}(\boldsymbol{x}_t-\boldsymbol{x}_{t-1})$
4. 更新$\boldsymbol{m}_t=\gamma\boldsymbol{m}_{t-1}+(1-\gamma)\nabla f(\boldsymbol{w}_t;z_{j_t})$
5. 更新$\boldsymbol{x}_{t+1}=\boldsymbol{x}_t-\eta_t\frac{\boldsymbol{m}_t}{\|\boldsymbol{m}_t\|}$
6. END FOR

算法2将隐式梯度传输结合到SGDM中。具体而言,步骤4在移动点$\boldsymbol{w}_t=\boldsymbol{x}_t+\frac{\gamma}{1-\gamma}(\boldsymbol{x}_t-\boldsymbol{x}_{t-1})$处估计随机梯度$\nabla f(\boldsymbol{w}_t;z_{j_t})$,融合了梯度更新过去的信息。$\boldsymbol{w}_t$的这种更新方法被称为隐式梯度传输,由文献[37]引入。相比于使用标准动量更新的估计,在隐式梯度传输和二阶光滑性假设下,$\boldsymbol{m}_t$是对$\nabla F_S(\boldsymbol{x}_t)$的一个偏置更小的估计[9]。步骤5中的$\frac{\boldsymbol{m}_t}{\|\boldsymbol{m}_t\|}$是动量的归一化,有助于简化理论证明。若忽略归一化操作,算法2中的更新可以理解为

$$\boldsymbol{x}_{t+1}=\boldsymbol{x}_t+\frac{\eta_t\gamma}{\eta_{t-1}}(\boldsymbol{x}_t-\boldsymbol{x}_{t-1})-\eta_t(1-\gamma)\nabla f\left(\boldsymbol{x}_t+\frac{\gamma}{1-\gamma}(\boldsymbol{x}_t-\boldsymbol{x}_{t-1});z_{j_t}\right) \quad (4)$$

与(1)相比,(4)在估计梯度时引入了过去的信息$\boldsymbol{x}_t-\boldsymbol{x}_{t-1}$。接下来,我们将为算法2提供高概率的收敛界限和泛化界限。

定理1. 设$\boldsymbol{x}_t$是由算法2生成的迭代序列。令步长为$\eta_t=\eta=\frac{a}{T^{5/7}}$且$1-\gamma=\frac{b}{T^{4/7}}$,其中$a,b$为任意正数且满足$1-\gamma\leqslant 1$。假设条件1、2和3成立。那么对于任意的$\delta\in(0,1)$,以$1-\delta$的概率,有以下不等式成立

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|=\mathcal{O}\left(\frac{1}{T^{2/7}}\log^{(\theta+\frac{1}{2})}(T/\delta)\right).$$

算法1的收敛界限是$\tilde{\mathcal{O}}(1/T^{1/4})$阶[6]。相比之下,定理1中的收敛界限是更快的$\tilde{\mathcal{O}}(1/T^{2/7})$阶数。此外,定理1中的界限与文献[9]中的期望收敛界限

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla F_S(\boldsymbol{x}_t)\right\|=\mathcal{O}\left(\frac{1}{T^{2/7}}\right)$$

相匹配，仅相差一个对数项，详见文献[9]中的定理3。该对数项是由于高概率分析带来的。

目前已有多项研究为随机梯度算法提供了高概率的收敛界限保证。我们将定理1与相关的文献作比较。其中，文献[38]同样研究非凸设置下的高概率收敛界限，并考虑了次威布尔梯度噪声，但其研究的是SGD算法，其定理15显示，当假设1和3成立，有界梯度假设$\left\|\nabla f(\boldsymbol{x}_t;z)\right\|\leqslant G$成立，有如下的高概率收敛界限

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla F_S(\boldsymbol{x}_t)\right\|=\mathcal{O}\left(\frac{\log^{\frac{1}{2}}(T)\log^{\theta}\frac{1}{\delta}+\log^{\min\{0,\theta-1\}}\left(\frac{T}{\delta}\right)\log^{\frac{1}{2}}\frac{1}{\delta}}{T^{1/4}}\right)$$

与此结果相比，定理1不需要有界梯度假设。尽管有界梯度假设是随机优化理论分析中的常用假设，也称为f的Lipschitz连续性假设，但在优化领域里，通常认为这一假设过于严格。因此，收敛性分析通常致力于去除该假设[6]。文献[13]同样研究了次威布尔梯度噪声下非凸SGD的高概率收敛界限，其定理3.1将有界梯度假设改进为一个弱化的条件$\eta_t\left\|\nabla F_S(\boldsymbol{x}_t)\right\|\leqslant G$，$t\in N$，此假设更弱的原因是$\eta_t$通常随迭代次数的增加而趋向于0。在额外的曲率条件下，文献[6]进一步提供了更快速的收敛界限。同时，文献[6]还探讨了泛化界限，我们将在后续进行讨论。与文献[13,38]相比，定理1研究的是更为复杂的动量算法。此外，定理1完全地移除了有界梯度假设，这意味着其提供了更具现实意义的收敛性分析。最近，从一个连续时间的视角研究随机优化问题，其证明当步长趋向于零时，SGDM的轨迹以L_2范数收敛为确定性的二阶常微分方程(ODE)。基于此发现，文献[39]构造出SGDM离散时间下的Lyapunov函数，揭示了SGDM的内在动态特性，然后将其应用于SGDM的收敛性分析，为SGDM提供了一种新颖的任意时刻收敛性保证。具体而言，当假设1成立，函数f是凸的，随机梯度满足$\mathbb{E}_{j_t}\left[\left\|\nabla f(\boldsymbol{x}_t;z_{j_t})\right\|^2\right]\leqslant\left\|\nabla F_S(\boldsymbol{x}_t)\right\|^2+\sigma^2$，并且随机梯度的噪声满足与文献[6]相同的次高斯假设(3)，文献[39]的定理5显示，以至少$1-\delta$的概率，有随后的收敛界限成立：

同时对于所有的$t\geqslant 0$，$F_S(\boldsymbol{x}_t)-\min F_S=\mathcal{O}\left(\frac{\log(t+2)\left(C_1(t)+C_2(t)\log\frac{1}{\delta}\right)}{(t+1)^{1/2}}\right)$，

其中，$C_1(t)$和$C_2(t)$是关于t的系数。与此结果相比，定理1研究的是非凸设置下的高概率收敛界限。虽然文献[39]的理论框架很有趣，但尚不确定其理论发现是否能推广至非凸的随机优化，而非凸的随机优化更符合现实。并且，文献[39]只研究了SGDM的收敛界限，没有涉及SGDM的泛化性能分析。本文为SGDM同时提供了高概率的收敛界限和泛化界限，收敛和泛化的联合视角有助于系统地理解SGDM在实际应用中的优良性能。

接下来为算法3.2提供高概率的泛化界限。

定理2.　设$\boldsymbol{x}_t$是由算法3.2生成的迭代序列。令步长为$\eta_t=\eta=\frac{a}{T^{5/7}}$且$1-\gamma=\frac{b}{T^{4/7}}$，其中$a,b$为任意正数且满足$1-\gamma\leqslant 1$。假设条件1、2和3成立。设置$T\asymp\left(\frac{n}{d}\right)^{\frac{7}{8}}$。那么对于任意的$\delta\in(0,1)$，以$1-\delta$的概率，有以下不等式成立：

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla F(\boldsymbol{x}_t)\right\|=\mathcal{O}\left(\left(\frac{d}{n}\right)^{\frac{1}{4}}\log^{\left(\theta+\frac{1}{2}\right)}\left(\frac{n}{d\delta}\right)\right).$$

运用本文的证明技术，可以得到当$T\asymp n/d$，算法1的泛化界限是$\tilde{\mathcal{O}}\left((d/n)^{\frac{1}{4}}\right)$阶。相比之下，定理2中的泛化界限的阶数也为$\tilde{\mathcal{O}}\left(d/n\right)^{\frac{1}{4}}\Big)$。因此可以得出结论，算法2在保持与原始SGDM类似的泛化性能的同时，拥有更快的收敛速度。此外，定理2显示，算法2的迭代复杂度为$T\asymp(n/d)^{\frac{7}{8}}$，优于算法1的$T\asymp n/d$。这些理论发现很好地解释了文献[9]中报告的算法2优越的实验性能。在相关工作中，文献[13]研究了次威布尔梯度噪声下非凸SGD的高概率泛化界限，但如前所述，其结果需要假设$\eta_t\left\|\nabla F_S(\boldsymbol{x}_t)\right\|\leqslant G$，$t\in N$。相比之下，定理2研究的是更为复杂的动量算法，动量项的存在使得其理论分析更为复杂。此外，定理1不需要有界梯度类型的假设，提供了更具现实意义的泛化性能分析。

3.4 具有海森校正动量的SGDM

算法3. 具有海森校正动量的SGDM

输入：步长$\{\eta_t\}_t$，数据集$S=\{z_1,z_2,\cdots,z_n\}$和动量参数γ

初始化：$\boldsymbol{x}_1=0,\boldsymbol{x}_0=0,\boldsymbol{m}_0=0$

1. FOR $t=1,\cdots,T$ DO
2. 从集合$\{j:j\in[n]\}$中以均匀分布采样j_t
3. 更新$\boldsymbol{m}_t=\gamma(\boldsymbol{m}_{t-1}+\nabla^2 f(\boldsymbol{x}_t;z_{j_t})(\boldsymbol{x}_t-\boldsymbol{x}_{t-1}))+(1-\gamma)\nabla f(\boldsymbol{x}_t;z_{j_t})$
4. 更新$\boldsymbol{x}_{t+1}=\boldsymbol{x}_t-\eta_t\frac{\boldsymbol{m}_t}{\|\boldsymbol{m}_t\|}$
5. END FOR

算法3将海森校正动量引入SGDM。具体而言，第3步使用海森-向量积$\nabla^2 f(\boldsymbol{x}_t;z_{j_t})(\boldsymbol{x}_t-\boldsymbol{x}_{t-1})$来“校正”SGDM中动量$\boldsymbol{m}_t$的偏差项。近年来，海森-向量积被用作在不产生显著计算开销的情况下实现某些二阶信息的优势[40-41]。第5步中的$\frac{\boldsymbol{m}_t}{\|\boldsymbol{m}_t\|}$同样是对动量的归一化。如果忽略归一化，算法3中的更新可以理解为

$$\boldsymbol{x}_{t+1}=\boldsymbol{x}_t-\eta_t(1-\gamma)\nabla f(\boldsymbol{x}_t;z_{j_t})-\eta_t\gamma\nabla^2 f(\boldsymbol{x}_t;z_{j_t})(\boldsymbol{x}_t-\boldsymbol{x}_{t-1})+\frac{\eta_t\gamma}{\eta_{t-1}}(\boldsymbol{x}_t-\boldsymbol{x}_{t-1}) \tag{5}$$

与(1)中的更新相比，(5)融合了二阶梯度信息$\nabla^2 f(\boldsymbol{x}_t;z_{j_t})(\boldsymbol{x}_t-\boldsymbol{x}_{t-1})$。接下来，我们将为算法3提供高概率的收敛界限和泛化界限。

表1 结果总结

文献	假设	指标	界限
文献[6]	$1,\theta=\frac{1}{2}$	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|$	$\mathcal{O}\left(\left(\frac{\log(T/\delta)\log T}{\sqrt{T}}\right)^{\frac{1}{2}}\right)$
	$1,\theta=\frac{1}{2}$	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|$	$\max\left\{\mathcal{O}\left(\left(\frac{d\log^{\frac{3}{2}}(T/\delta)}{\sqrt{T}}\right)^{\frac{1}{2}}\right),\mathcal{O}\left(\left(\frac{d^2\log^2(T/\delta)}{T}\right)^{\frac{1}{2}}\right)\right\}$
文献[9]	1,2，一阶有界方差梯度噪声	$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F_S(\boldsymbol{x}_t)\|$	$\mathcal{O}\left(\frac{1}{T^{2/7}}\right)$
文献[10]	1,2，一阶和二阶有界方差梯度噪声，有界梯度	$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F_S(\boldsymbol{x}_t)\|$	$\mathcal{O}\left(\frac{1}{T^{1/3}}\right)$
本文	$1,2,\theta\geqslant\frac{1}{2}$	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|$	$\mathcal{O}\left(\frac{1}{T^{2/7}}\log^{\left(\theta+\frac{1}{2}\right)}(T/\delta)\right)$
	$1,2,\theta\geqslant\frac{1}{2}$	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|$	$\mathcal{O}\left(\left(\frac{d}{n}\right)^{\frac{1}{4}}\log^{\left(\theta+\frac{1}{2}\right)}\left(\frac{n}{d\delta}\right)\right)$
	$1,2,3,4,\theta\geqslant\frac{1}{2}$	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|$	$\mathcal{O}\left(\frac{1}{T^{1/3}}\log^{\left(\theta+\frac{1}{2}\right)}(T/\delta)\right)$
	$1,2,3,4,\theta\geqslant\frac{1}{2}$	$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|$	$\mathcal{O}\left(\left(\frac{d}{n}\right)^{\frac{1}{4}}\log^{\left(\theta+\frac{1}{2}\right)}\left(\frac{n}{d\delta}\right)\right)$

定理3. 设$\boldsymbol{x}_t$为算法3生成的迭代序列。令步长为$\eta_t=\eta=\frac{a}{T^{2/3}}$且$1-\gamma=\frac{b}{T^{5/9}}$，其中$a,b$为任意正数且满足$1-\gamma\leqslant 1$。假设条件1、2、3和4成立。那么对于任意的$\delta\in(0,1)$，以概率$1-\delta$，有以下不等式成立

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|=\mathcal{O}\left(\frac{1}{T^{1/3}}\log^{\left(\theta+\frac{1}{2}\right)}(T/\delta)\right).$$

定理3中的收敛界限的阶数为$\tilde{\mathcal{O}}\left(\frac{1}{T^{1/3}}\right)$。这一界限与文献[10]中的期望收敛界限

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F_S(\boldsymbol{x}_t)\|=\mathcal{O}\left(\frac{1}{T^{1/3}}\right)$$

相匹配，仅存在对数项的区别，详见文献[10]的定理5。此外，可以看出，定理3中收敛界限展示出比定理1中的收敛界限更快的收敛速度。

我们现在讨论几种常见的自适应梯度算法(adaptive gradient algorithms)，包括AdaGrad[42]、RMSProp[43]和Adam[44]。在非凸优化中，相关研究通常为自适应的梯度算法提供期望的收敛界限。目前，AdaGrad的高概率收敛界限已经得到了一些研究成果，通常分为两种形式：实值版本(AdaGrad-

Norm)和元素版本(AdaGrad)[42]。文献[45]针对次高斯梯度噪声(3),为AdaGrad-Norm提供了能适应噪声水平的收敛界限,但该结果依赖于有界梯度假设。文献[46]改进了这一结果,在相同的噪声假设下,移除了有界梯度假设,并将结果扩展到AdaGrad的元素版本。文献[6]研究了AdaGrad的元素版本,在次高斯梯度噪声下,为具有动量的AdaGrad变体(延迟步长)提供了高概率收敛界限,且该结果同样不需要有界梯度假设。最近,RMSProp的非凸优化研究同样取得了进展。文献[47]为RMSProp提供了高概率收敛界限,该结果假设有界梯度和次高斯梯度噪声。如何在类似条件下为Adam算法提供高概率收敛界限,是一个极为重要的问题。由于Adam结合了动量和RMSProp算法的优势,我们认为结合本文中关于动量的证明技术和RMSProp的分析技术[47],有可能为Adam提供高概率的结果。然而,Adam独特的自适应学习率形式增加了分析难度。传统自适应梯度方法(如AdaGrad和RMSProp)的步长随时间递减,现有AdaGrad和RMSProp的高概率分析大多依赖于这一递减特性,而Adam的步长是非单调的。此外,Adam采用了偏差修正的动量机制,这使其更新方向变为对真实梯度的复杂有偏估计。不过,我们相信本文使用的处理动量导致的有偏估计的迭代技术会为Adam的分析提供帮助。

接下来为算法3提供高概率的泛化界限。

定理4.　设$\boldsymbol{x}_t$为算法3生成的迭代序列。令步长为$\eta=\frac{a}{T^{2/3}}$且$1-\gamma=\frac{b}{T^{5/9}}$,其中$a,b$为任意正数且满足$1-\gamma\leqslant 1$。假设条件1、2、3和4成立。设置$T\asymp(\frac{n}{d})^{\frac{3}{4}}$。那么对于任意$\delta\in(0,1)$,以概率$1-\delta$,有以下不等式成立:

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla F(\boldsymbol{x}_t)\right\|=\mathcal{O}\left(\left(\frac{d}{n}\right)^{\frac{1}{4}}\log^{\left(\theta+\frac{1}{2}\right)}\left(\frac{n}{d\delta}\right)\right).$$

定理4中的泛化界限阶数为$\tilde{\mathcal{O}}((d/n)^{\frac{1}{4}})$阶,其速率与定理2相似。通过综合考虑收敛界限和泛化界限,可以得出结论:算法3具有与算法2相似的泛化性能,但其收敛速度更快。此外,算法3的迭代复杂度$T\asymp(n/d)^{\frac{3}{4}}$也优于定理2的$T\asymp(n/d)^{\frac{7}{8}}$。这些理论结果很好地解释了文献[10]中所报告的算法3在实验中的优越性能。

还有,文献[10]的结果需要有界梯度假设$\left\|\nabla f(\boldsymbol{x}_t;z)\right\|\leqslant G$,见表1。如前所述,这个假设是个过强的假设[6],我们的分析成功地移除了这一假设。我们在表1中提供了本文获得的结果以及相关工作的结果,这里只列出了最相关的文献。文献[6]的第二个结果是针对SGDM的一种变体算法推导而来,即延迟的带动量的AdaGrad,其步长不包含当前的梯度。从表1中可以看出,本文提供了一系列相关工作不涉及的高概率泛化界限以及快速率的高概率收敛界限。

4　数值实验

4.1　泛化界限随参数θ的变化

本小节通过数值实验来验证算法2的泛化界限$\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla F(\boldsymbol{x}_t)\right\|^2$随参数$\theta$的变化情况。算法3的结果也可以通过类似实验进行验证。

设$F_S(\boldsymbol{x})$和$F_{S'}(\boldsymbol{x})$分别为基于训练数据集S和测试数据集S'的经验风险,可知$F_{S'}(\boldsymbol{x})=\frac{1}{|S'|}\sum_{z\in S'}f(\boldsymbol{x};z)$,其中$|S'|$是集合$S'$的基数。我们使用$F_{S'}(\boldsymbol{x})$作为对总体风险$F$的一个良好近似。实验中使用了六个来自LIBSVM的数据集:Heart、Fourclass、German、Australian、Diabetes和Phishing[48]。每个数据集的80%用作训练数据,其余20%用作测试数据。为考察参数θ的影响,我们考虑了$\theta\in\{1/2,1,5\}$。

根据式(4),动量的更新公式为$\boldsymbol{m}_t=\gamma\boldsymbol{m}_{t-1}+(1-\gamma)(\nabla F_S(\boldsymbol{x}_t)+\nabla f(\boldsymbol{x}_t;z_{j_t})-\nabla F_S(\boldsymbol{x}_t))=\gamma\boldsymbol{m}_{t-1}+(1-\gamma)(\nabla F_S(\boldsymbol{x}_t)+\boldsymbol{e}_t)$,其中$\boldsymbol{e}_t=\nabla f(\boldsymbol{x}_t;z_{j_t})-\nabla F_S(\boldsymbol{x}_t)$。在训练过程的每次更新中,我们独立且同分布地从次威布尔分布中为每个维度抽取随机变量,以模拟假设3中的梯度噪声$\boldsymbol{e}_t$。如果随机向量$\boldsymbol{e}_t$的每个坐标均服从次威布尔分布,则$\|\boldsymbol{e}_t\|$也是一个次威布尔随机变量,这可以通过文献[49]的引理3.4和文献[50]的命题2.1(c)进行证明。由于我们假设随机梯度是准确梯度的无偏估计,我们对分布进行偏移和缩放,以获得均值为零且方差等于1的随机向量。

实验采用二元分类的广义线性模型$\ell(\langle\boldsymbol{x},\boldsymbol{x}\rangle)$,其中,$\ell$是逻辑斯蒂函数$\ell(s)=(1+e^{-s})^{-1}$。第一个实验研究Hube损失,其形式为$f(\boldsymbol{x},z)=\frac{1}{2}\left(\ell(\langle\boldsymbol{x},x\rangle)-y\right)^2$,当$|\ell(\langle\boldsymbol{x},x\rangle)-y|\leqslant\tau$时,否则

为$\tau\left(|\ell(\langle \boldsymbol{x},x\rangle)-y|-\frac{1}{2}\tau\right)$。设定$\tau=0.1,\gamma=0.9$和$\eta_t=0.1t^{-\frac{1}{2}}$,实验重复100次,并报告结果的平均值。泛化界限$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|^2$随迭代次数的变化如图1所示。实验结果与定理2的泛化界限一致,θ的增加会导致泛化性能下降。第二个实验研究平方损失,其形式为$f(\boldsymbol{x},z)=(\ell(\langle \boldsymbol{x},x\rangle)-y)^2$。在这种情况下,指标$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|^2$随迭代次数的变化趋势如图2所示。实验结果同样表明$\theta$的增加将导致更差的泛化结果,与定理2的泛化界限一致。此外值得注意的是,定理2中对于θ的依赖是$\log^{\left(\theta+\frac{1}{2}\right)}(T/\delta)$,仅是对数阶的。因此,当$\theta$取不同值($\theta=1/2,\theta\in(1/2,1]$,$\theta>1$)时,对泛化界限的影响很小,这也与图1和图2中揭示的泛化性能的衰减趋势一致。

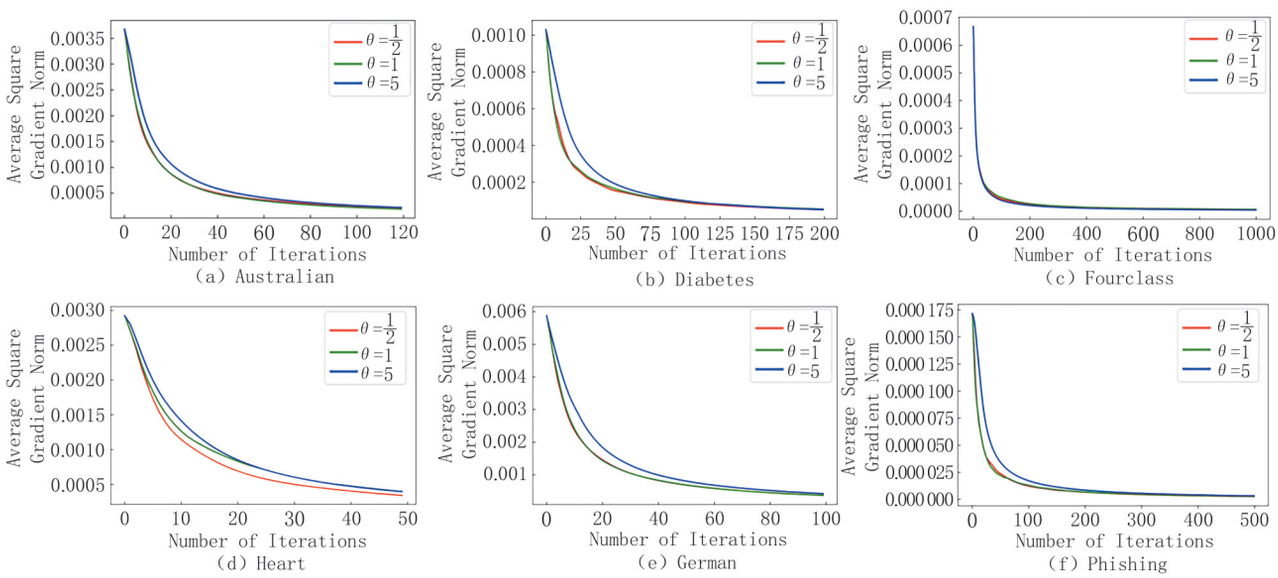

图1 Huber损失下,不同的选择$\left(\theta\in\{1/2,\ 1,\ 5\}\right.$、泛化指标$\left.\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|^2\right)$与遍历次数的关系

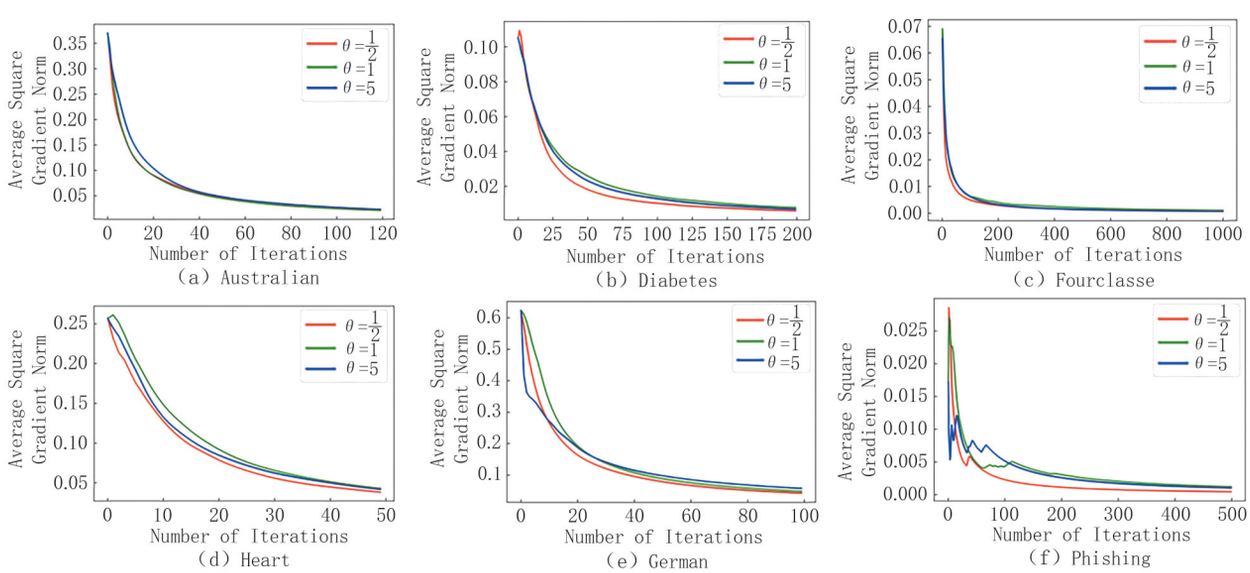

图2 平方损失下,不同的选择$\left(\theta\in\{1/2,\ 1,5\}\right.$、泛化指标$\left.\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|^2\right)$与遍历次数的关系

在数值实验中,我们同样可以考虑泛化指标的分布函数,或者其他能够分析概率速度的度量。具体地,定理2中的泛化界限可以等价地写成:对于任意的$t\geqslant 0$,

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|\geqslant t\right)=\mathcal{O}\left(\left(\frac{n}{d}\right)\exp\left\{-\left(\left(\frac{n}{d}\right)^{\frac{1}{4}}t\right)^{\frac{1}{\theta+1/2}}\right\}\right)。$$

对于泛化指标$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|$，其分布函数$F(t)$可以写成

$$F(t)=1-\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|\geqslant t\right)=\begin{cases}0 & \text{if } t\leqslant 0\\ \mathcal{O}\left(\left(\frac{n}{d}\right)\exp\left\{-\left(\left(\frac{n}{d}\right)^{\frac{1}{4}}t\right)^{\frac{1}{\theta+1/2}}\right\}\right) & \text{if } t>0\end{cases}$$

我们同样使用测试数据集S'上的经验风险$F_{S'}(\boldsymbol{x})$作为总体风险F的一个良好近似，并据此计算$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|$。通过设计数值实验，我们可以观察$F(t)$关于$t$的增长速度是否和$\mathcal{O}(\left(\frac{n}{d}\right)\exp\{-\left(\left(\frac{n}{d}\right)^{\frac{1}{4}}t\right)^{\frac{1}{\theta+1/2}}\})$的速率一致。此外，我们还可以探讨$F(t)$是否随$\theta$的增加而增加，这个增长是因为$1/(\theta+1/2)$项导致的。

4.2　二阶梯度噪声的重尾性质

本小节进一步通过数值实验来验证假设4中二阶梯度噪声满足次威布尔分布的合理性。我们研究算法3，因为仅算法3使用了假设4。

我们对$\|\nabla^2 f(\boldsymbol{x}_t;z_i)\boldsymbol{y}-\nabla^2 F_S(\boldsymbol{x}_t)\boldsymbol{y}\|/\|\boldsymbol{y}\|$的分布进行数值研究，其中$\boldsymbol{x}_t$是算法3产生的最后一次迭代，$i=1,2,\cdots,n$。我们根据$\|\nabla^2 f(\boldsymbol{x}_t;z_1)\boldsymbol{y}-\nabla^2 F_S(\boldsymbol{x}_t)\boldsymbol{y}\|/\|\boldsymbol{y}\|,\cdots,\|\nabla^2 f(\boldsymbol{x}_t;z_n)\boldsymbol{y}-\nabla^2 F_S(\boldsymbol{x}_t)\boldsymbol{y}\|/\|\boldsymbol{y}\|$的值绘制相应的概率密度分布直方图，同时使用SciPy库中的scipy.stats.weibull_min函数拟合次威布尔分布的比例(scale)、形状(shape)和位置(loc)等参数，并绘制相应的次威布尔分布的概率密度函数曲线。

实验采用二元分类的广义线性模型$\ell(\langle\boldsymbol{x},\boldsymbol{x}\rangle)$，其中$l$是逻辑斯蒂函数$\ell(s)=(1+e^{-s})^{-1}$，使用交叉熵损失，其形式为$f(\boldsymbol{x},z)=-[y\log\ell(\langle\boldsymbol{x},\boldsymbol{x}\rangle)+(1-y)\log(1-\ell(\langle\boldsymbol{x},\boldsymbol{x}\rangle))]$。与第4.1节类似，实验采用六个来自LIBSVM的数据集：Heart、Fourclass、German、Australian、Diabetes和Phishing[48]，并设定$\gamma=0.9$和$\eta_t=0.1t^{-\frac{1}{2}}$。对于$\|\nabla^2 f(\boldsymbol{x}_t;z_i)\boldsymbol{y}-\nabla^2 F_S(\boldsymbol{x}_t)\boldsymbol{y}\|/\|\boldsymbol{y}\|$中的随机向量$\boldsymbol{y}$，实验采用均匀分布在区间$[0,1]$上生成。实验结果如图3所示。这些直方图显示，算法3的二阶梯度噪声的范数具有

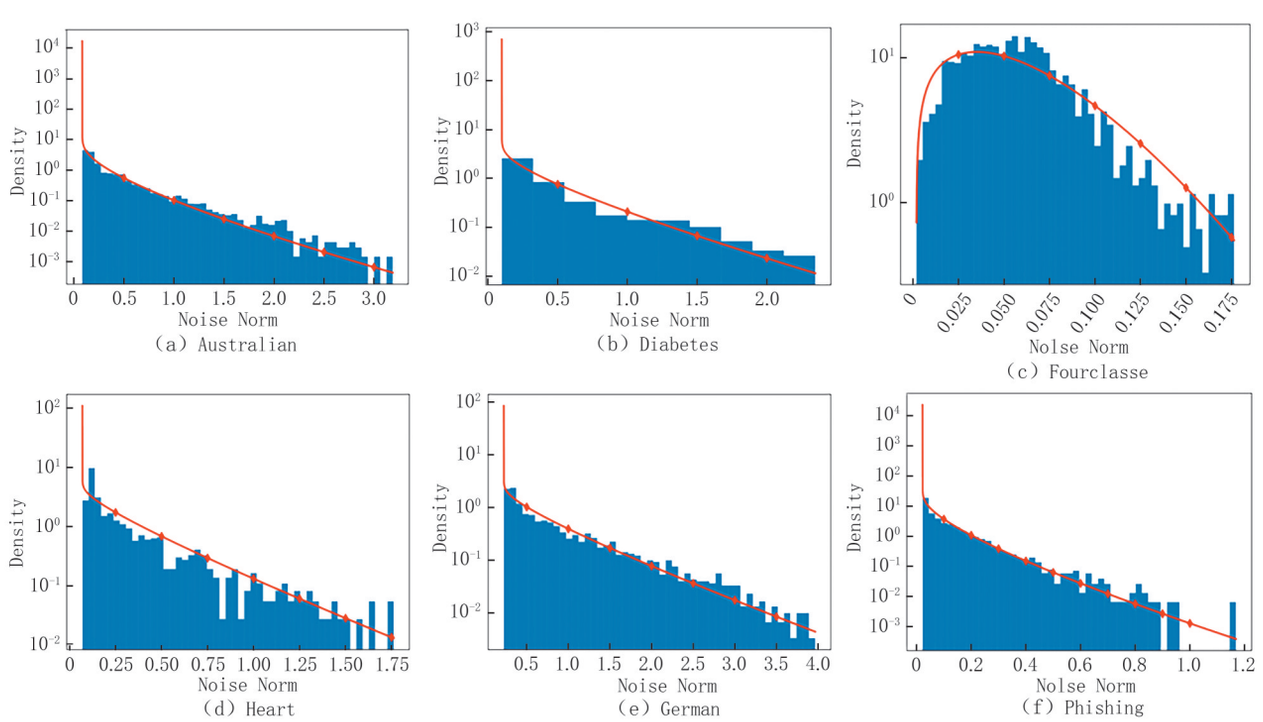

图3　不同数据集的$\|\nabla^2 f(\boldsymbol{x}_t;z_i)\boldsymbol{y}-\nabla^2 F_S(\boldsymbol{x}_t)\boldsymbol{y}\|/\|\boldsymbol{y}\|$的概率密度分布直方图以及相应的次威布尔分布的概率密度函数曲线(红线)

与次威布尔分布相似的变化趋势，说明其可以很好地用次威布尔分布拟合，验证了假设4中二阶梯度噪声满足次威布尔分布的合理性。

5 结 论

本文研究了非凸情境下基于动量的随机梯度下降算法的学习界限。我们研究了高概率的界限，相比于期望的界限，高概率的结果对样本空间最坏的情况也成立。并且，我们同时提供了收敛界限和泛化界限，这使得能从优化和泛化的联合作用出发，系统地分析SGDM的性能。此外，我们考虑梯度噪声服从一种重尾的次威布尔分布，得出的理论界限能够展示从次高斯/次指数（即轻尾）变量转向重尾变量的过程中重尾梯度噪声对收敛速率和泛化性能的影响。最后，数值实验验证了我们的理论结果。我们相信，本文的理论发现能够为非凸SGDM算法的学习保证提供深刻的见解。

致 谢 感谢编辑和审稿人宝贵的意见。本研究得到国家自然科学基金面上项目(No. 62476277)、国家重点研发计划(No. 2024YFE0203200)和CCF-阿里妈妈科技资基金(No. CCF-ALIMAMA OF 2024008)的资助。

参 考 文 献

[1] Bottou L, Curtis F E, Nocedal J. Optimization methods for large-scale machine learning. Siam Review, 2018, 60(2): 223-311

[2] Sun Tao, Li Dong-Sheng. Nonconvex low rank and total variation regularized model and algorithm for image deblurring. Chinese Journal of Computers, 2020, 43(4):643-652(in Chinese)
(孙涛, 李东升. 基于非凸的全变分和低秩混合正则化的图像去模糊模型和算法. 计算机学报, 2020, 43(4): 643-652)

[3] Liu Ji-Yuan, Liu Xin-Wang, Cai Zhi-Ping, et al. On the correlation measurement of data representations. Chinese Journal of Computers, 2024, 47(7): 1568-1581 (in Chinese)
(刘吉元, 刘新旺, 蔡志平等. 数据表示的相关性度量方法. 计算机学报, 2024, 47(7): 1568-1581)

[4] Li Xiang, Chen Shuo, Yang Jian. Generalization bound regularizer: A unified perspective for understanding weight decay. Chinese Journal of Computers, 2021, 44(10): 2122-2134 (in Chinese)
(李翔, 陈硕, 杨健. 泛化界正则项: 理解权重衰减正则形式的统一视角. 计算机学报, 2021, 44(10): 2122-2134)

[5] Li S, Liu Y. Learning rates for nonconvex pair wise learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9996-10011

[6] Li X, Orabona F. A high probability analysis of adaptive sgd with momentum. arXiv preprint arXiv:2007.14294, 2020

[7] Harvey N J, Liaw C, PLAN Y, et al. Tight analyses for non-smooth stochastic gradient descent//Proceedings of the Conference on Learning Theory. Phoenix, USA, 2019: 1579-1613

[8] Cutkosky A, Mehta H. High-probability bounds for non-convex stochastic optimization with heavy tails//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021: 4883-4895

[9] Cutkosky A, Mehta H. Momentum improves normalized sgd// Proceedings of the International Conference on Machine Learning. Virtual, 2020: 2260-2268

[10] Tran H, Cutkosky A. Better SGD using second-order momentum//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 3530-3541

[11] Polyak B T. Some methods of speeding up the convergence of iteration methods. Ussr Computational Mathematics and Mathematical Physics, 1964, 4(5): 1-17

[12] Arjevani Y, Carmon Y, Duchi J C, et al.Lower bounds for non-convex stochastic optimization. Mathematical Programming, 2023, 199(1): 165-214

[13] Li S, Liu Y. High probability guarantees for nonconvex stochastic gradient descent with heavy tails//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 12931-12963

[14] Sutskever I, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning//Proceedings of the International Conference on Machine Learning. Atlanta, USA, 2013: 1139-1147

[15] Liu H, Tian X. Sgem: Stochastic gradient with energy and momentum. Numerical Algorithms, 2024, 95(4): 1583-1610

[16] Tran-Dinh Q, Pham N H, Phan D T, et al. A hybrid stochastic optimization framework for composite nonconvex optimization. Mathematical Programming, 2022, 191(2): 1005-1071

[17] Cutkosky A, Orabona F. Momentum-based variance reduction in non-convex sgd// Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 15236-15245

[18] Liu Y, Gao Y, Yin W. An improved analysis of stochastic gradient descent with momentum//Advances in Neural Information Processing Systems.Virtual, 2020: 18261-18271

[19] Gitman I, Lang H, Zhang P, et al. Understanding the role of momentum in stochastic gradient methods//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019:9633-9643

[20] Deng Y, Song Z, Yang C. Enhancing stochastic gradient descent: A unified framework and novel acceleration methods for faster convergence. arXiv preprint arXiv:2402.01515, 2024

[21] Tran H, Cutkosky A. Momentum aggregation for private non-convex erm// Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 10996-11008

[22] Ong M Y. Understanding generalization[M. S. Dissertation].

Massachusetts Institute of Technology, USA, 2017

[23] Chen Y, Jin C, Yu B. Stability and convergence trade-off of iterative optimization algorithms. arXiv preprint arXiv: 1804.01619, 2018

[24] Ramezani-Kebrya A, Antonakopoulosk, Cevher V, et al. On the generalization of stochastic gradient descent with momentum. Journal of Machine Learning Research, 2024, 25(22): 1-56

[25] Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge, USA: Cambridge University Press, 2014

[26] Attia A, Koren T. Algorithmic instabilities of accelerated gradient descent// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021:1204-1214

[27] Li S, Liu Y. High probability analysis for non-convex stochastic optimization with clipping//Proceedings of the European Conference on Artificial Intelligence. Kraków, Poland, 2023: 1406-1413

[28] Nesterov Y. Introductory lectures on convex optimization: A basic course. New York, USA: Springer, 2014

[29] Nesterov Y, Polyak B T. Cubic regularization of newton method and its global performance. Mathematical Programming, 2006, 108(1): 177-205

[30] Vladimirova M, Girard S, Nguyen H, et al. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. Stat, 2020, 9(1): e318

[31] Kuchibhotla A K, Chakrabortty A. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. Information and Inference: A Journal of the IMA, 2022, 11(4): 1389-1456

[32] Simsekli U, Sagun L, Gurbuzbalaban M. A tail-index analysis of stochastic gradient noise in deep neural networks// Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 5827-5837

[33] Şimşekli U, Gürbüzbalaban M, Nguyent H, et al. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. arXiv preprint arXiv:1912.00018, 2019

[34] Gurbuzbalaban M, Hu Y. Fractional moment-preserving initialization schemes for training deep neural networks// Proceedings of the International Conference on Artificial Intelligence and Statistics. Virtual, 2021: 2233-2241

[35] Zhang J, Karimireddy S P, Veit A, et al. Why are adaptive methods good for attention models?// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 15383-15393

[36] Wang H, Gürbüzbalaban M, Zhu L, et al.Convergence rates of stochastic gradient descent under infinite noise variance// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021: 18866-18877

[37] Arnold S, Manzagol P A, Babanezhad Harikandeh R, et al. Reducing the variance in online optimization by transporting past gradients//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 5391-5402

[38] Madden L, Dall′anese E, Becker S. High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. Journal of Machine Learning Research, 2024, 25(241):1-36

[39] Feng Y, Jiang Y, Wang T, et al. The anytime convergence of stochastic gradient descent with momentum: From a continuous-time perspective. arXiv preprint arXiv:2310.19598, 2024

[40] Tripuraneni N, Stern M, Jin C, et al. Stochastic cubic regularization for fast nonconvex optimization//Proceedings of the Advances in Neural Information Processing Systems. Montréal Canada, 2018: 2899-2908

[41] Zhou D, Xu P, Gu Q. Stochastic variance-reduced cubic regularization methods. Journal of Machine Learning Research, 2019, 20(134): 1-47

[42] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 2011, 12(7): 2121-2159

[43] Tieleman T, Hinton G. Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. University of Toronto, Canada, Technical Report: 6, 2012

[44] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014

[45] Kavis A, Levy K Y, Cevher V. High probability bounds for a class of nonconvex algorithms with adagrad stepsize. arXiv preprint arXiv:2204.02833, 2022

[46] Liu Z, Nguyen T D, Nguyen T H, et al. High probability convergence of stochastic gradient methods//Proceedings of the International Conference on Machine Learning. Hawaii, USA, 2023: 21884-21914

[47] Zhou D, Chen J, Cao Y, et al. On the convergence of adaptive gradient methods for nonconvex optimization. arXiv preprint arXiv:1808.05671, 2018

[48] Chang C C, Lin C J. Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27

[49] Bastianello N, Madden L, Carli R, et al. Astochastic operator framework for inexact static and online optimization. arXiv preprint arXiv:2105.09884, 2021

[50] Kim S, Madden L, Dall′anese E. Convergence of the inexact online gradient and proximal-gradient under the polyak-łojasiewicz condition. arXiv preprint arXiv:2108.03285, 2021

[51] Li C J. A note on concentration inequality for vector-valued martingales with weak exponential-type tail. arXiv preprint arXiv:1809.02495, 2021

[52] Lei Y, Tang K. Learning rates for stochastic gradient descent with nonconvex objectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4505-4511

附录A. 辅助引理

前两个引理是次威布尔的鞅差序列的集中不等式。

引理1 （定理2[51]）. 令$\theta\in(0,\infty)$。假设($\boldsymbol{X}_i\in\mathbb{R}^d, i=1,2,\cdots,N$)是相对于滤波(*filtration*)$\mathcal{F}_i$的鞅差序列，即$\mathbb{E}[\boldsymbol{X}_i|\mathcal{F}_{i-1}]=0$，并满足以下弱指数尾部条件：对于某些$\theta>0$和所有$i=1,2,\cdots,N$，以及某

个标量$K_i>0$，我们有$\mathbb{E}\left[\exp\left(\left\|\frac{\boldsymbol{X}_i}{K_i}\right\|^{\frac{1}{\theta}}\right)\right]\leqslant 2$。假设对于每个$i=1,2,\cdots,N,K_i<\infty$。则，对于任意$N\geqslant 1$和$t>0$，有以下不等式成立

$$P\left(\max_{n\leqslant N}\left\|\sum_{i=1}^{n}\boldsymbol{X}_i\right\|\geqslant t\right)\leqslant 4\left[3+(3\theta)^{2\theta}\frac{128\sum_{i=1}^{N}K_i^2}{t^2}\right]\exp\left\{-\left(\frac{t^2}{64\sum_{i=1}^{N}K_i^2}\right)^{\frac{1}{2\theta+1}}\right\}。$$

引理2 （定理11[38]）. 令$(\Omega,\mathcal{F},(\mathcal{F}_i),P)$是一个域流概率空间。设$(\xi_i)$和$(K_i)$适应于$(\mathcal{F}_i)$，$n\in\mathbb{N}$，并且对于所有$i\in[n]$，假设$K_{i-1}\geqslant 0$，$\mathbb{E}[\xi_i|\mathcal{F}_{i-1}]=0$以及

$$\mathbb{E}\left[\exp\left((|\xi_i|/K_{i-1})^{1/\theta}\right)|\mathcal{F}_{i-1}\right]\leqslant 2,$$

其中，$\theta\geqslant 1/2$。如果$\theta>1/2$，假设存在(m_i)使得$K_{i-1}\leqslant m_i$。

如果$\theta=1/2$，令$a=2$。则对于所有$x,\beta\geqslant 0$，$\alpha>0$以及$\lambda\in\left[0,\frac{1}{2\alpha}\right]$，

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i\geqslant x\text{且}\sum_{i=1}^{k}aK_{i-1}^2\leqslant\alpha\sum_{i=1}^{k}\xi_i+\beta\right\}\right)\leqslant\exp(-\lambda x+2\lambda^2\beta),$$

并且对于所有$x,\beta,\lambda\geqslant 0$，

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i\geqslant x\text{且}\sum_{i=1}^{k}aK_{i-1}^2\leqslant\beta\right\}\right)\leqslant\exp\left(-\lambda x+\frac{\lambda^2}{2}\beta\right)。$$

如果$\theta\in\left(\frac{1}{2},1\right]$，令$a=(4\theta)^{2\theta}e^2$且$b=(4\theta)^{\theta}e$。则对于所有$x,\beta\geqslant 0,\alpha\geqslant b\max_{i\in[n]}m_i$以及$\lambda\in\left[0,\frac{1}{2\alpha}\right]$，

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i\geqslant x\text{且}\sum_{i=1}^{k}aK_{i-1}^2\leqslant\alpha\sum_{i=1}^{k}\xi_i+\beta\right\}\right)\leqslant\exp(-\lambda x+2\lambda^2\beta),$$

并且对于所有$x,\beta\geqslant 0$以及$\lambda\in\left[0,\frac{1}{b\max_{i\in[n]}m_i}\right]$，

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i\geqslant x\text{且}\sum_{i=1}^{k}aK_{i-1}^2\leqslant\beta\right\}\right)\leqslant\exp\left(-\lambda x+\frac{\lambda^2}{2}\beta\right)。$$

如果$\theta>1$，令$\delta\in(0,1)$，$a=(2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3}$，$b=2\log^{\theta-1}(n/\delta)$。则对于所有$x,\beta\geqslant 0,\alpha\geqslant b\max_{i\in[n]}m_i$以及$\lambda\in\left[0,\frac{1}{2\alpha}\right]$，

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i\geqslant x\text{且}\sum_{i=1}^{k}aK_{i-1}^2\leqslant\alpha\sum_{i=1}^{k}\xi_i+\beta\right\}\right)\leqslant\exp(-\lambda x+2\lambda^2\beta)+2\delta,$$

并且对于所有$x,\beta\geqslant 0$以及$\lambda\in\left[0,\frac{1}{b\max_{i\in[n]}m_i}\right]$，

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i\geqslant x\text{且}\sum_{i=1}^{k}aK_{i-1}^2\leqslant\beta\right\}\right)\leqslant\exp\left(-\lambda x+\frac{\lambda^2}{2}\beta\right)+2\delta。$$

最后一个引理描述了总体梯度∇F与经验梯度∇F_S之间的一致偏差。

引理3 （推论2[52]）. 记$B_R=B(0,R)$。令$\delta\in(0,1)$，$S=\{z_1,z_2,\cdots,z_n\}$是一组独立同分布的样本。假设条件1成立。则以至少$1-\delta$的概率有以下不等式成立

$$\sup_{x\in B_R}\|\nabla F(\boldsymbol{x})-\nabla F_S(\boldsymbol{x})\|\leqslant\frac{(LR+B)}{\sqrt{n}}\times\left(2+2\sqrt{48e\sqrt{2}\,(\log 2+d\log(3e))}+\sqrt{2\log\left(\frac{1}{\delta}\right)}\right),$$

其中$B=\sup_{z\in\mathcal{Z}}\|\nabla f(0;z)\|$，$L$为光滑性参数。

附录B. 定理1的证明

证明. 定义$\epsilon_t=\boldsymbol{m}_t-\nabla F_S(\boldsymbol{x}_t)$，根据假设1，有

$$F_S(\boldsymbol{x}_{t+1})-F_S(\boldsymbol{x}_t)\leqslant\langle\boldsymbol{x}_{t+1}-\boldsymbol{x}_t,\nabla F_S(\boldsymbol{x}_t)\rangle+\frac{1}{2}L\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2=-\left\langle\eta_t\frac{\boldsymbol{m}_t}{\|\boldsymbol{m}_t\|},\nabla F_S(\boldsymbol{x}_t)\right\rangle+\frac{1}{2}L\eta_t^2。$$

首先考虑$-\left\langle\frac{\boldsymbol{m}_t}{\|\boldsymbol{m}_t\|},\nabla F_S(\boldsymbol{x}_t)\right\rangle$。当$\|\epsilon_t\|\leqslant\frac{1}{2}\|\nabla F_S(\boldsymbol{x}_t)\|$时，有$\|\epsilon_t+\nabla F_S(\boldsymbol{x}_t)\|\leqslant\frac{3}{2}\|\nabla F_S(\boldsymbol{x}_t)\|$，此时

$$-\left\langle\frac{\boldsymbol{m}_t}{\|\boldsymbol{m}_t\|},\nabla F_S(\boldsymbol{x}_t)\right\rangle=-\left\langle\frac{\epsilon_t+\nabla F_S(\boldsymbol{x}_t)}{\|\epsilon_t+\nabla F_S(\boldsymbol{x}_t)\|},\nabla F_S(\boldsymbol{x}_t)\right\rangle\leqslant-\left\langle\frac{\|\nabla F_S(\boldsymbol{x}_t)\|^2}{\|\epsilon_t+\nabla F_S(\boldsymbol{x}_t)\|}+\frac{\|\nabla F_S(\boldsymbol{x}_t)\|\|\epsilon_t\|}{\|\epsilon_t+\nabla F_S(\boldsymbol{x}_t)\|}\right\rangle\leqslant-\frac{1}{2}\frac{\|\nabla F_S(\boldsymbol{x}_t)\|^2}{\|\epsilon_t+\nabla F_S(\boldsymbol{x}_t)\|}\leqslant-\frac{1}{3}\|\nabla F_S(\boldsymbol{x}_t)\|。$$

当$\|\epsilon_t\|>\frac{1}{2}\|\nabla F_S(\boldsymbol{x}_t)\|$时，有

$$-\left\langle \frac{\boldsymbol{m}_t}{\|\boldsymbol{m}_t\|}, \nabla F_S(\boldsymbol{x}_t)\right\rangle \leqslant \|\nabla F_S(\boldsymbol{x}_t)\| = -\frac{1}{3}\|\nabla F_S(\boldsymbol{x}_t)\| + \frac{4}{3}\|\nabla F_S(\boldsymbol{x}_t)\| \leqslant -\frac{1}{3}\|\nabla F_S(\boldsymbol{x}_t)\| + \frac{8}{3}\|\epsilon_t\|$$

结合这两种情况可以推出

$$-\left\langle \frac{\boldsymbol{m}_t}{\|\boldsymbol{m}_t\|}, \nabla F_S(\boldsymbol{x}_t)\right\rangle \leqslant -\frac{1}{3}\|\nabla F_S(\boldsymbol{x}_t)\| + \frac{8}{3}\|\epsilon_t\|,$$

这表明

$$F_S(\boldsymbol{x}_{t+1}) - F_S(\boldsymbol{x}_t) \leqslant -\frac{1}{3}\eta_t\|\nabla F_S(\boldsymbol{x}_t)\| + \frac{8}{3}\eta_t\|\epsilon_t\| + \frac{1}{2}L\eta_t^2。$$

从$t=1$至$t=T$累加，并根据$\eta_t=\eta$，可以得到

$$\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\| \leqslant 3\frac{F_S(\boldsymbol{x}_1) - F_S(\boldsymbol{x}_{T+1})}{\eta} + 8\sum_{t=1}^{T}\|\epsilon_t\| + \frac{3LT}{2}\eta。$$

现在对$\sum_{t=1}^{T}\|\epsilon_t\|$进行界定。首先定义$\epsilon'_t = \nabla f(\boldsymbol{w}_t; z_{j_t}) - \nabla F_S(\boldsymbol{w}_t)$和$Z(a,b) = \nabla F_S(a) - \nabla F_S(b) - \nabla^2 F_S(b)(a-b)$，那么对于任意的$t \geqslant 1$有以下递归公式：

$$\begin{aligned}\boldsymbol{m}_{t+1} &= \gamma(\epsilon_t + \nabla F_S(\boldsymbol{x}_t)) + (1-\gamma)\nabla f(\boldsymbol{w}_{t+1}; z_{j_{t+1}}) = \\ &\gamma(\nabla F_S(\boldsymbol{x}_{t+1}) + \nabla^2 F_S(\boldsymbol{x}_{t+1})(\boldsymbol{x}_t - \boldsymbol{x}_{t+1})) + \\ &\gamma(Z(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) + \epsilon_t) + (1-\gamma)(\epsilon'_{t+1} + \\ &Z(\boldsymbol{w}_{t+1}, \boldsymbol{x}_{t+1})) + (1-\gamma)(\nabla F_S(\boldsymbol{x}_{t+1}) + \\ &\nabla^2 F_S(\boldsymbol{x}_{t+1})(\boldsymbol{x}_t - \boldsymbol{x}_{t+1})) = \nabla F_S(\boldsymbol{x}_{t+1}) + \gamma(\epsilon_t + \\ &Z(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})) + (1-\gamma)(\epsilon'_{t+1} + Z(\boldsymbol{w}_{t+1}, \boldsymbol{x}_{t+1})),\end{aligned}$$

其中，最后一个方程是由$\boldsymbol{w}_t$的定义得出的。根据ϵ_t的定义，可以推出

$$\epsilon_{t+1} = \gamma(\epsilon_t + Z(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})) + (1-\gamma)(\epsilon'_{t+1} + Z(\boldsymbol{w}_{t+1}, \boldsymbol{x}_{t+1}))$$

通过设定$\boldsymbol{m}_0 = 0$和$\boldsymbol{x}_0 = \boldsymbol{x}_1$，有$\epsilon_0 = -\nabla F_S(\boldsymbol{x}_1)$，因此上述$\epsilon_{t+1}$的递归关系在$t=0$时也成立。此时，展开递归关系可以得到

$$\epsilon_{t+1} = (1-\gamma)\sum_{\tau=0}^{t}\gamma^\tau\epsilon'_{t+1-\tau} + \gamma\sum_{\tau=0}^{t}\gamma^\tau Z(\boldsymbol{x}_{t-\tau}, \boldsymbol{x}_{t-\tau+1}) + (1-\gamma)\sum_{\tau=0}^{t}\gamma^\tau Z(\boldsymbol{w}_{t-\tau+1}, \boldsymbol{x}_{t-\tau+1}) + \gamma^{t+1}\epsilon_0。$$

相应地，可以得到

$$\|\epsilon_{t+1}\| = (1-\gamma)\left\|\sum_{\tau=0}^{t}\gamma^\tau\epsilon'_{t+1-\tau}\right\| + \gamma^{t+1}\|\epsilon_0\| + \gamma\left\|\sum_{\tau=0}^{t}\gamma^\tau Z(\boldsymbol{x}_{t-\tau}, \boldsymbol{x}_{t-\tau+1})\right\| + (1-\gamma)\left\|\sum_{\tau=0}^{t}\gamma^\tau Z(\boldsymbol{w}_{t-\tau+1}, \boldsymbol{x}_{t-\tau+1})\right\|$$

由于F_S满足假设2，根据泰勒定理可以得到$\|Z(a,b)\| \leqslant \rho\|a-b\|^2$，利用这个性质，可以进一步推导出

$$\gamma\left\|\sum_{\tau=0}^{t}\gamma^\tau Z(\boldsymbol{x}_{t-\tau}, \boldsymbol{x}_{t-\tau+1})\right\| \leqslant \gamma\sum_{\tau=0}^{t}\gamma^\tau\rho\eta^2$$

以及

$$(1-\gamma)\left\|\sum_{\tau=0}^{t}\gamma^\tau Z(\boldsymbol{w}_{t-\tau+1}, \boldsymbol{x}_{t-\tau+1})\right\| \leqslant (1-\gamma)\sum_{\tau=0}^{t}\gamma^\tau\rho\eta^2\left(\frac{\gamma}{1-\gamma}\right)^2。$$

由于$\frac{\gamma^2}{1-\gamma} + \gamma \leqslant \frac{\gamma}{1-\gamma}$，将这些项结合起来可以得到

$$\|\epsilon_{t+1}\| \leqslant \gamma^{t+1}\|\epsilon_0\| + (1-\gamma)\left\|\sum_{\tau=0}^{t}\gamma^\tau\epsilon'_{t+1-\tau}\right\| + \frac{\gamma}{1-\gamma}\sum_{\tau=0}^{t}\gamma^\tau\rho\eta^2 \leqslant \gamma^{t+1}B + (1-\gamma)\left\|\sum_{\tau=0}^{t}\gamma^\tau\epsilon'_{t+1-\tau}\right\| + \frac{\gamma}{(1-\gamma)^2}\rho\eta^2,$$

其中，B的出现是因为$B = \sup_{z\in\mathcal{Z}}\|\nabla f(0; z)\|$。关于$\left\|\sum_{\tau=0}^{t}\gamma^\tau\epsilon'_{t+1-\tau}\right\|$，由于$\|\nabla f(\boldsymbol{w}_t; z_{j_t}) - \nabla F_S(\boldsymbol{w}_t)\|$是一个次威布尔随机变量，根据次威布尔变量的定义得到

$$\mathbb{E}\left[\exp\left(\frac{\gamma^\tau\|\nabla f(\boldsymbol{x}_{t+1-\tau}; z_{j_{t+1-\tau}}) - \nabla F_S(\boldsymbol{x}_{t+1-\tau})\|}{\gamma^\tau K}\right)^{\frac{1}{\theta}}\right] \leqslant 2,$$

这意味着$\gamma^\tau\|\nabla f(\boldsymbol{x}_{t+1-\tau}; z_{j_{t+1-\tau}}) - \nabla F_S(\boldsymbol{x}_{t+1-\tau})\| \sim \mathrm{sub}W(\theta, \gamma^\tau K)$。然后应用引理1推导出以下不等式

$$P\left(\max_{t\leqslant T}\left\|\sum_{\tau=0}^{t}\gamma^\tau\epsilon_{t+1-\tau}\right\| \geqslant x\right) \leqslant 4\left[3 + (3\theta)^{2\theta}\frac{128K^2\sum_{i=0}^{T}\gamma^{2i}}{x^2}\right]\exp\left\{-\left(\frac{x^2}{64K^2\sum_{i=0}^{T}\gamma^{2i}}\right)^{\frac{1}{2\theta+1}}\right\}。$$

令$4\exp\left\{-\left(\frac{x^2}{64K^2\sum_{i=0}^{T}\gamma^{2i}}\right)^{\frac{1}{2\theta+1}}\right\} = \delta$，可以得到$x = 8\log^{(\theta+\frac{1}{2})}(\frac{4}{\delta})K(\sum_{i=0}^{T}\gamma^{2i})^{\frac{1}{2}}$，因此以概率$1 - 3\delta - \frac{8(3\theta)^{2\theta}}{\log^{2\theta+1}\frac{4}{\delta}}\delta$，有

$$\max_{t\leqslant T}\left\|\sum_{\tau=0}^{t}\gamma^\tau\epsilon_{t+1-\tau}\right\| \leqslant 8\log^{(\theta+\frac{1}{2})}(\frac{4}{\delta})K(\sum_{i=0}^{T}\gamma^{2i})^{\frac{1}{2}}。$$

由于$\theta\geqslant 1/2$且$\delta\in(0,1)$,可以得知$\log^{2\theta+1}\frac{4}{\delta}>1$,这意味着以概率$1-3\delta-8(3\theta)^{2\theta}\delta$,有

$\max\limits_{t\leqslant T}\left\|\sum_{\tau=0}^{t}\gamma^{\tau}\epsilon_{t+1-\tau}\right\|\leqslant 8\log^{(\theta+\frac{1}{2})}(\frac{4}{\delta})K(\sum_{i=0}^{T}\gamma^{2i})^{\frac{1}{2}}$。

现在,以概率$1-\delta$,得出以下不等式

$$(1-\gamma)\max_{t\leqslant T}\left\|\sum_{\tau=0}^{t}\gamma^{\tau}\epsilon_{t+1-\tau}\right\|\leqslant$$

$$8\log^{(\theta+\frac{1}{2})}(\frac{4(3+8(3\theta)^{2\theta})}{\delta})K(\sum_{i=0}^{T}\gamma^{2i})^{\frac{1}{2}}(1-\gamma)\leqslant$$

$$8\log^{(\theta+\frac{1}{2})}(\frac{4(3+8(3\theta)^{2\theta})}{\delta})K\frac{(1-\gamma)}{(1-\gamma^2)^{1/2}}\leqslant$$

$$8\log^{(\theta+\frac{1}{2})}(\frac{4(3+8(3\theta)^{2\theta})}{\delta})K(1-\gamma)^{1/2},$$

将这个结果代入到$\|\epsilon_{t+1}\|$的不等式中,可以得出

$$\|\epsilon_{t+1}\|\leqslant\gamma^{t+1}B+\frac{\gamma}{(1-\gamma)^2}\rho\eta^2+$$

$$8\log^{(\theta+\frac{1}{2})}(\frac{4(3+8(3\theta)^{2\theta})}{\delta})K(1-\gamma)^{1/2},$$

将这个结果代入到$\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|$的不等式中,以概率$1-T\delta$可以得到

$$\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|\leqslant 3\frac{F_S(\boldsymbol{x}_1)-F_S(\boldsymbol{x}_{T+1})}{\eta}+$$

$$\frac{\gamma}{1-\gamma}8B+64\log^{(\theta+\frac{1}{2})}(\frac{4(3+8(3\theta)^{2\theta})}{\delta})$$

$$K(1-\gamma)^{1/2}T+\frac{8\gamma}{(1-\gamma)^2}\rho\eta^2T+\frac{3LT}{2}\eta。$$

设置$\eta=\frac{a}{T^{5/7}}$和$1-\gamma=\frac{b}{T^{4/7}}$,其中$a,b$是任意正数且满足$1-\gamma\leqslant 1$,至此,成功推导出以至少$1-\delta$的概率有

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|=\mathcal{O}\left(\frac{1}{T^{2/7}}\log^{(\theta+\frac{1}{2})}(T/\delta)\right)。$$

证毕。

附录C. 定理2的证明

证明. 首先,以概率$1-\delta$,有以下不等式成立

$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|\leqslant\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)-\nabla F_S(\boldsymbol{x}_t)\|+\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|\leqslant\frac{1}{T}\sum_{t=1}^{T}\max_{t\leqslant T}\|\nabla F(\boldsymbol{x}_t)-\nabla F_S(\boldsymbol{x}_t)\|+\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|\leqslant\max_{t\leqslant T}\|\nabla F(\boldsymbol{x}_t)-\nabla F_S(\boldsymbol{x}_t)\|+\mathcal{O}(\frac{1}{T^{2/7}}\log^{(\theta+\frac{1}{2})}(\frac{T}{\delta}))$。

然后,由于$\boldsymbol{x}_{t+1}=\boldsymbol{x}_t-\eta_t\frac{\boldsymbol{m}_t}{\|\boldsymbol{m}_t\|}$,可以得出

$$\boldsymbol{x}_{t+1}=-\sum_{i=1}^{t}\eta_i\frac{\boldsymbol{m}_i}{\|\boldsymbol{m}_i\|}$$

以及

$$\|\boldsymbol{x}_{t+1}\|\leqslant\sum_{i=1}^{t}\eta_i。$$

因为$\eta_t=\eta=\frac{a}{T^{5/7}}$,可以推出$\|\boldsymbol{x}_{t+1}\|\leqslant\frac{at}{T^{5/7}}$。为了简洁,令$\gamma=(2+2\sqrt{48e\sqrt{2}\,(\log 2+d\log(3e))}+\sqrt{2\log(\frac{1}{\delta})})$。根据引理3,以概率$1-\delta$,有

$$\max_{t\leqslant T}\|\nabla F(\boldsymbol{x}_t)-\nabla F_S(\boldsymbol{x}_t)\|\leqslant\frac{(LR_T+B)\gamma}{\sqrt{n}}\leqslant$$

$$\frac{(L\|\boldsymbol{x}_T\|+B)\gamma}{\sqrt{n}}\leqslant\frac{(L\frac{aT}{T^{5/7}}+B)\gamma}{\sqrt{n}}。$$

因此以概率$1-2\delta$,可以得到

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|=\mathcal{O}\left(\frac{T^{\frac{2}{7}}}{\sqrt{n}}\times(\sqrt{d}+\log^{\frac{1}{2}}(\frac{1}{\delta}))\right)+$$

$$\mathcal{O}\left(\frac{1}{T^{\frac{2}{7}}}\log^{(\theta+\frac{1}{2})}(\frac{T}{\delta})\right)。$$

取$T\asymp(\frac{n}{d})^{\frac{7}{8}}$,得到

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|=\mathcal{O}\left((\frac{d}{n})^{\frac{1}{4}}\log^{(\theta+\frac{1}{2})}(\frac{n}{d\delta})\right),$$

这意味着以至少$1-\delta$的概率我们有

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|=\mathcal{O}\left((\frac{d}{n})^{\frac{1}{4}}\log^{(\theta+\frac{1}{2})}(\frac{n}{d\delta})\right)。$$

证毕。

附录D. 定理3的证明

证明. 通过与定理1类似的证明,定义$\epsilon_t=\boldsymbol{m}_t-\nabla F_S(\boldsymbol{x}_t)$,可以得到以下不等式:

$$\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|\leqslant 3\frac{F_S(\boldsymbol{x}_1)-F_S(\boldsymbol{x}_{T+1})}{\eta}+$$

$$8\sum_{t=1}^{T}\|\epsilon_t\|+\frac{3LT}{2}\eta$$

现在来界定$\sum_{t=1}^{T}\|\epsilon_t\|$。首先定义$\epsilon_t'=\nabla f(\boldsymbol{x}_t;z_{j_t})-\nabla F_S(\boldsymbol{x}_t)$、$\boldsymbol{w}_t=\frac{\nabla^2 f(\boldsymbol{x}_t,z_{j_t})(\boldsymbol{x}_t-\boldsymbol{x}_{t-1})-\nabla^2 F_S(\boldsymbol{x}_t)(\boldsymbol{x}_t-\boldsymbol{x}_{t-1})}{\|\boldsymbol{x}_t-\boldsymbol{x}_{t-1}\|}$

以及$Z(a,b)=\nabla F_S(a)-\nabla F_S(b)-\nabla^2 F_S(b)(a-b)$,对于任意$t\geqslant 1$有如下的递推公式

$$\epsilon_{t+1}=\boldsymbol{m}_{t+1}-\nabla F_S(\boldsymbol{x}_{t+1})=\gamma(\boldsymbol{m}_t+\nabla^2 f(\boldsymbol{x}_{t+1};z_{j_{t+1}})(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t)-\nabla F_S(\boldsymbol{x}_{t+1}))+(1-\gamma)(\nabla f(\boldsymbol{x}_{t+1};z_{j_{t+1}})-\nabla F_S(\boldsymbol{x}_{t+1}))=\gamma(\boldsymbol{m}_t-\nabla F_S(\boldsymbol{x}_t)+\nabla F_S(\boldsymbol{x}_t)+\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|\boldsymbol{w}_{t+1}+\nabla^2 F_S(\boldsymbol{x}_{t+1})(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t)-\nabla F_S(\boldsymbol{x}_{t+1}))+(1-\gamma)\epsilon'_{t+1}=\gamma(\epsilon_t+\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|\boldsymbol{w}_{t+1}+Z(\boldsymbol{x}_t,\boldsymbol{x}_{t+1}))+(1-\gamma)\epsilon'_{t+1},$$

通过设定$\boldsymbol{m}_0=0$和$\boldsymbol{x}_0=\boldsymbol{x}_1$,可以得到$\epsilon_0=-\nabla F_S(\boldsymbol{x}_1)$。进一步地,容易验证上述$\epsilon_{t+1}$的递推关系同样适用于$t=0$。接下来展开递推关系得到

$$\epsilon_{t+1}=\gamma\sum_{\tau=0}^{t}\gamma^{\tau}\|\boldsymbol{x}_{t-\tau+1}-\boldsymbol{x}_{t-\tau}\|\boldsymbol{w}_{t-\tau+1}+\gamma^{t+1}\epsilon_0+(1-\gamma)\sum_{\tau=0}^{t}\gamma^{\tau}\epsilon'_{t+1-\tau}+\gamma\sum_{\tau=0}^{t}\gamma^{\tau}Z(\boldsymbol{x}_{t-\tau},\boldsymbol{x}_{t-\tau+1}),$$

因此有

$$\|\epsilon_{t+1}\|=\gamma\eta\left\|\sum_{\tau=0}^{t}\gamma^{\tau}\boldsymbol{w}_{t-\tau+1}\right\|+\gamma^{t+1}\|\epsilon_0\|+(1-\gamma)\left\|\sum_{\tau=0}^{t}\gamma^{\tau}\epsilon'_{t+1-\tau}\right\|+\gamma\left\|\sum_{\tau=0}^{t}\gamma^{\tau}Z(\boldsymbol{x}_{t-\tau},\boldsymbol{x}_{t-\tau+1})\right\|。$$

根据不等式$\|Z(a,b)\|\leqslant\rho\|a-b\|^2$,可以推出

$$\gamma\left\|\sum_{\tau=0}^{t}\gamma^{\tau}Z(\boldsymbol{x}_{t-\tau},\boldsymbol{x}_{t-\tau+1})\right\|\leqslant\gamma\sum_{\tau=0}^{t}\gamma^{\tau}\rho\eta^2。$$

关于$\left\|\sum_{\tau=0}^{t}\gamma^{\tau}\epsilon'_{t+1-\tau}\right\|$,由于$\|\nabla f(\boldsymbol{w}_t;z_{j_t})-\nabla F_S(\boldsymbol{w}_t)$是一个次威布尔随机变量,根据次威布尔变量的定义得到

$$\mathbb{E}[\exp(\frac{\gamma^{\tau}\|\nabla f(\boldsymbol{x}_{t+1-\tau};z_{j_{t+1-\tau}})-\nabla F_S(\boldsymbol{x}_{t+1-\tau})\|}{\gamma^{\tau}K})^{\frac{1}{\theta}}]\leqslant 2,$$

这意味着$\gamma^{\tau}\|\nabla f(\boldsymbol{x}_{t+1-\tau};z_{j_{t+1-\tau}})-\nabla F_S(\boldsymbol{x}_{t+1-\tau})\|\sim \mathrm{sub}W(\theta,\gamma^{\tau}K)$。那么类似地,以$1-\delta$的概率有下列不等式

$$(1-\gamma)\max_{t\leqslant T}\left\|\sum_{\tau=0}^{t}\gamma^{\tau}\epsilon_{t+1-\tau}\right\|\leqslant 8\log^{(\theta+\frac{1}{2})}(\frac{4(3+8(3\theta)^{2\theta})}{\delta})K(1-\gamma)^{1/2}。$$

关于$\left\|\sum_{\tau=0}^{t}\gamma^{\tau}\boldsymbol{w}_{t-\tau+1}\right\|$,需要使用假设4。根据假设4,显然有$\gamma^{\tau}\|\boldsymbol{w}_{t-\tau+1}\|\sim subW(\theta,\gamma^{\tau}K')$。通过类似的证明步骤,应用引理1可以得到以概率$1-\delta$,有

$$\gamma\eta\max_{t\leqslant T}\left\|\sum_{\tau=0}^{t}\gamma^{\tau}\boldsymbol{w}_{t-\tau+1}\right\|\leqslant 8\log^{(\theta+\frac{1}{2})}(\frac{4(3+8(3\theta)^{2\theta})}{\delta})K'\frac{\gamma\eta}{(1-\gamma^2)^{1/2}}。$$

结合上述界限,代入到$\|\epsilon_{t+1}\|$的不等式中,以至少$1-2\delta$的概率可以得到

$$\|\epsilon_{t+1}\|\leqslant\gamma^{t+1}B+\frac{\gamma}{1-\gamma}\rho\eta^2+8\log^{(\theta+\frac{1}{2})}(\frac{4(3+8(3\theta)^{2\theta})}{\delta})\max\{K,K'\}\times((1-\gamma)^{1/2}+\frac{\gamma\eta}{(1-\gamma^2)^{1/2}})。$$

将这个结果代入到$\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|$的不等式中,得到以至少$1-2T\delta$的概率,有

$$\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|\leqslant 3\frac{F_S(\boldsymbol{x}_1)-F_S(\boldsymbol{x}_{T+1})}{\eta}+\frac{\gamma}{1-\gamma}8B+\max\{K,K'\}((1-\gamma)^{1/2}+\frac{\gamma\eta}{(1-\gamma^2)^{1/2}})T64\log^{(\theta+\frac{1}{2})}(\frac{4(3+8(3\theta)^{2\theta})}{\delta})+\frac{8\gamma}{1-\gamma}\rho\eta^2T+\frac{3LT}{2}\eta。$$

设置$\eta=\frac{a}{T^{2/3}}$和$1-\gamma=\frac{b}{T^{5/9}}$,其中$a,b$是任意的正数且满足$1-\gamma\leqslant 1$。至此,成功推导出以至少$1-\delta$的概率,有

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|=\mathcal{O}\left(\frac{1}{T^{1/3}}\log^{(\theta+\frac{1}{2})}(T/\delta)\right)。$$

证毕。

附录E. 定理4的证明

证明.　与定理2的证明类似,以概率为$1-\delta$,有下述不等式成立

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|\leqslant\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)-\nabla F_S(\boldsymbol{x}_t)\|+\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\boldsymbol{x}_t)\|=\max_{t\leqslant T}\|\nabla F(\boldsymbol{x}_t)-\nabla F_S(\boldsymbol{x}_t)\|+\mathcal{O}\left(\frac{1}{T^{1/3}}\log^{(\theta+\frac{1}{2})}(T/\delta)\right)。$$

由于$\|\boldsymbol{x}_{t+1}\|\leqslant\sum_{i=1}^{t}\eta_i$以及$\eta_t=\eta=\frac{a}{T^{2/3}}$,可以得到$\|\boldsymbol{x}_{t+1}\|\leqslant\frac{at}{T^{2/3}}$。同样,令

$$\gamma=(2+2\sqrt{48e\sqrt{2}\,(\log 2+d\log(3e))}+\sqrt{2\log(\frac{1}{\delta})})。$$

根据引理3,以概率$1-\delta$,有

$$\max_{t\leqslant T}\|\nabla F(\boldsymbol{x}_t)-\nabla F_S(\boldsymbol{x}_t)\|\leqslant\frac{(LR_T+B)\gamma}{\sqrt{n}}\leqslant\frac{(L\|\boldsymbol{x}_T\|+B)\gamma}{\sqrt{n}}\leqslant\frac{(L\frac{aT}{T^{2/3}}+B)\gamma}{\sqrt{n}},$$

因此以概率$1-2\delta$,可以得到

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|=\mathcal{O}\left(\frac{T^{\frac{1}{3}}}{\sqrt{n}}\times(\sqrt{d}+\log^{\frac{1}{2}}(\frac{1}{\delta}))\right)+\mathcal{O}\left(\frac{1}{T^{1/3}}\log^{(\theta+\frac{1}{2})}(T/\delta)\right),$$

取 $T\asymp(\frac{n}{d})^{\frac{3}{4}}$，得到

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|=\mathcal{O}\left((\frac{d}{n})^{\frac{1}{4}}\log^{(\theta+\frac{1}{2})}(\frac{n}{d\delta})\right),$$

这意味着以至少 $1-\delta$ 的概率，有

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\boldsymbol{x}_t)\|=\mathcal{O}\left((\frac{d}{n})^{\frac{1}{4}}\log^{(\theta+\frac{1}{2})}(\frac{n}{d\delta})\right)。$$

证毕。

LI Shao-Jie, Ph. D. His research interests include machine learning theory and theories for large language models.

LIU Yong, Ph. D., associate professor. His research interests include algorithms and theories for large language model and fundamental theory of machine learning.

Background

SGDM is a widely used algorithms in optimization area. It enhances the basic stochastic gradient descent method by incorporating a momentum term, which helps accelerate convergence, particularly in the presence of noise and ill-conditioned landscapes common in high-dimensional non-convex optimization problems. Despite its practical effectiveness, the theoretical understanding of SGDM remains limited, especially when considering its behavior in non-convex domains.

Overally, the existing literature primarily focuses on analyzing SGDM in an expectation sense, which provides insights into its average case performance. However, the lack of high-probability analyses means that little is known about the algorithm′s performance in worst-case scenarios. Understanding high-probability convergence is essential because it offers stronger guarantees regarding the performance of optimization algorithms. This gap in the theoretical understanding of SGDM motivates the present study, which aims to provide high-probability convergence and generalization bounds for SGDM.

In this paper, we advance the understanding of SGDM by deriving high-probability convergence bounds that align with established in-expectation results. Importantly, we also introduce generalization bounds for SGDM, which, to our knowledge, are the first to be proposed for SGDM. By addressing both convergence and generalization, we not only illuminate the algorithm's effectiveness in practice but also elucidate the superior characteristics of two recently proposed SGDM algorithms, which have garnered attention in the machine learning community.

This research was supported by National Natural Science Foundation of China (No. 62476277), National Key Research and Development Program of China (No. 2024YFE0203200), and CCF-ALIMAMA TECH Kangaroo Fund (No. CCF-ALIMAMA OF 2024008). The convergence bounds and generalization bounds proposed in this paper are the key to understanding theoretical properties of SGDM. The authors always focus on the theories of machine learning and have published many highly-quality papers in this field.