arXiv:2305.05675v1 [cs.LG] 9 May 2023

UAdam：非凸随机优化的统一Adam类型算法框架

⋆

Yiming Jiang[a]、Jinlan Liu[a]、Dongpo Xu[a,*]、Danilo P. Mandic[b,*]

[a]教育部应用统计重点实验室、东北师范大学数学与统计学院，长春 130024，中国。[b]帝国理工学院电气与电子工程系，英国伦敦 SW7 2AZ

摘要

Adam 类算法已成为深度学习设置中优化问题的首选方法。然而，尽管取得了成功，其收敛性仍未得到充分理解。为此，我们提出了一个统一的Adam类型算法框架（称为UAdam）。该框架配备了二阶矩的一般形式，因此可以将Adam及其变体作为特例纳入其中，例如NAdam、AMSGrad、AdaBound、AdaFom和Adan。我们通过对UAdam在非凸随机设置下进行严格的收敛分析，对此提供了理论支持，结果表明，UAdam以$\mathcal{O}(1/T)$的速率收敛到驻点邻域。此外，随着β的增加，邻域大小会减小。重要的是，我们的分析仅要求一阶动量因子足够接近1，而不对二阶动量因子施加任何限制。理论结果还表明，通过选择适当的超参数，原始Adam可以实现收敛，从而为整类Adam 型算法的分析、应用和进一步发展提供理论保证。

关键词：非凸优化、统一Adam、收敛分析、方差递推估计、深度学习。

⋆本研究部分得到国家自然科学基金（项目编号：62176051、62272096）、国家重点研发计划（项目编号：2021YFA1003400）以及中央高校基本科研业务费（项目编号：2412020FZ024）的资助。

*通讯作者电子邮件地址：xudp100@nenu.edu.cn（Dongpo Xu），d.mandic@imperial.ac.uk（Danilo P. Mandic)

预印本已提交至《神经网络》。 2023年5月11日

1. 引言

深度神经网络在计算机视觉、图像识别和自然语言处理等诸多领域取得了巨大成功。训练深度神经网络通常考虑以下非凸随机优化设定

$$\min_{x\in\mathbb{R}^d} f(x) = \mathbb{E}_{\xi\sim\mathbb{P}}\left[f(x,\xi)\right], \tag{1}$$

其中， x 是待优化的模型参数， ξ 表示从某个未知概率分布 $\mathbb{P}$中抽取的随机变量， $f(x,\xi)$表示可微的非凸损失函数。求解式（1）中优化问题最常用的方法是一类基于随机梯度的一阶方法。随机梯度下降（SGD）算法 [8, 23] 因其简单高效而得到广泛应用；该算法沿负梯度方向更新模型，并按步长进行缩放。然而，SGD算法可能收敛缓慢，甚至陷入局部极小值，尤其是在模型规模较大且问题病态时。为解决这些问题，人们引入了动量技术，例如 Polyak [21], 提出的重球法（HB）加速算法，其更新规则如下：

$$\text{SHB: } x_{t+1} = x_t - \alpha\nabla f\left(x_t,\xi_t\right) + \beta\left(x_t - x_{t-1}\right), \tag{2}$$

其中， $x_0 = x_1 \in \mathbb{R}^d$、 $\alpha > 0$ 是步长， β 是（凸）动量因子，其取值为 $0 \leq \beta < 1$。Nesterov [19] 提出了另一种广受欢迎的加速技术，称为Nesterov加速梯度（NAG），其形式为 [25]

$$\text{SNAG: } \begin{cases} m_t = \beta m_{t-1} - \alpha\nabla f\left(x_t + \beta m_{t-1}, \xi_t\right) \\ x_{t+1} = x_t + m_t \end{cases}, \tag{3}$$

其中， $\alpha > 0$ 是步长， $\beta \in [0,1)$ 是动量因子。从物理意义上讲，NAG 会从 x_t 沿历史梯度 m_{t-1}的方向迈出一小步，并利用包含“前瞻梯度”的指数移动平均来更新参数。值得注意的是，Nesterov 加速法的收敛速度比重球法更快 [15, 20, 25]。

步长的选择对随机梯度下降的收敛性能至关重要，通常需要使用固定步长或逐渐减小的步长序列[1, 24]。因此，自适应学习率算法已成为一种流行的替代方案，例如 AdaGrad [7], 、RMSProp [26, 33], 和 Adam [13], ，它们根据历史梯度的二阶矩自适应地调整学习率。例如，AdaGrad [7]会逐元素累积所有过去的梯度平方，以调整步长，从而为梯度较小的维度分配较大的学习率。然而，在迭代过程中，累积操作可能导致步长过小，进而产生梯度消失现象。为了解决这一问题，Tieleman 和 Hinton [26] 提出了 RMSProp 算法，用指数移动平均取代 AdaGrad 中的累积操作。在此基础上，Adam [13] 将动量策略与 RMSProp 相结合，成为深度学习中最流行的自适应方法。

尽管 Adam 在实践中表现出色，Reddi $et\ al$。 [22] 指出，即使对于简单的凸优化问题，Adam 仍可能发散。为此，研究人员提出了 Adam 的各种变体，例如 AMSGrad [22], 、AdaBound [16, 18], 、AdaFom [3], 和 Yogi [34], ，这些算法的差异仅在于二阶矩。然而，这些进展大多基于启发式方法。因此，我们提出一个用于处理自适应动量算法的统一框架，其中 Adam、AdaFom、AMSGrad、AdaBound 及许多其他算法都是该框架的特例。最近，Zhang $et\ al$。 [35] 声称 Adam 无需修改更新规则即可收敛，并指出 Reddi $et\ al$。 [22] 提出的发散问题存在缺陷，即先确定参数，再选择问题。换言之，为了使 Adam 发散，会针对不同的一阶矩和二阶矩参数组合（β_1、 β_2）选取依赖于 β 的示例。然而，在实际应用中，需要针对给定的优化问题调整参数对（β_1、 β_2）。因此，若先给定问题，Adam 便可通过适当选择超参数来保证收敛。[35]中的收敛结果要求二阶矩参数 β_2 足够大。相比之下，我们的分析不对二阶矩参数 β_2施加任何限制，并证明 UAdam 配合各种形式的二阶矩时，只需选择适当的一阶动量参数 β_1即可收敛，这与实践观察一致。

1.1. 相关工作

尽管 Adam 及其变体在深度神经网络训练中取得了显著成功，但其理论分析 [7, 13] 主要针对在线凸设置，因此无法阐明非凸设置下的收敛性，而非凸设置通常见于实际应用。在证明 Adam 及其变体在非凸设置下收敛的诸多尝试中，Li 和 Orabona [14]给出了非凸情形下广义全局 AdaGrad 步长的收敛速度和高概率界。此外，Ward *et al*。 [31] 证明，AdaGrad 的范数版本（AdaGrad-Norm）在随机设置下以$\mathcal{O}(\log(T)/\sqrt{T})$的速度收敛到一个驻点。Chen *et al*。 [3] 和 Guo *etal*。 [9] 分析了一类 Adam 算法的收敛性能；他们的分析具有模块化特性，并可扩展用于求解其他优化问题，例如组合优化问题和极大极小问题 [9]。Zaheer *et al*。 [34] 研究了小批量大小对 Adam 收敛性能的影响，表明增大小批量大小有助于收敛。他们还提出了一种新颖的自适应优化方法（称为 Yogi），能够控制有效学习率的增长，从而取得更好的性能。Zou *et al*。 [38] 提出了易于检验的充分条件，以确保 Adam 及其变体全局收敛，并对 Adam 的发散给出了新的解释：发散可能是由二阶矩参数设置不当造成的。D´efossez *et al*。 [4] 为 AdaGrad 和 Adam 给出了任意小的上界，表明在适当的超参数设置下，这些算法可以按 $\mathcal{O}(d\log(T)/\sqrt{T})$ 的速度收敛。此外，Zhou *et al*。 [36] 证明，在梯度有界假设下，AMSGrad、改进的 RMSProp 和 AdaGrad 均以 $\mathcal{O}(1/\sqrt{T})$ 的速度收敛。此外，Zhang *et al*。 [35] 指出，Adam 无需修改更新规则即可收敛，且不需要梯度有界或二阶矩有界假设；而 Wang *et al*。 [28] 则进一步分析了 Adam 在（L_0、L_1）光滑性条件下的收敛性。值得注意的是，上述 Adam 型算法依赖重球法来估计随机梯度的一阶矩。

近年来，Nesterov 加速法已被应用于自适应学习算法的一阶矩和二阶矩中。一个例子是 Dozat [6] 提出的名为 NAdam 的自适应算法，该算法结合了 Nesterov 加速法和 RMSProp。尽管 NAdam 在实际实验中表现良好，但目前尚无充分的理论依据.....

为保证收敛性，需要提供理论分析。为解决这一问题，Zou $et\ al$ [37] 提出了一种新的自适应随机动量算法，将加权逐坐标 AdaGrad 与统一动量相结合。在非凸情形下，他们建立了 $\mathcal{O}(\log(T)/\sqrt{T})$的非渐近收敛速率，从而为理解 Adam 和 RMSProp 的收敛性提供了新的视角。此外，Xie $et\ al$ [32] 提出了一种自适应 Nesterov 动量算法（Adan），用于估计梯度的一阶矩和二阶矩，并证明在非凸情形下，Adan 需要 $\mathcal{O}(\epsilon^{-4})$ 的随机梯度复杂度才能找到一个 ϵ-驻点。

表1：不同自适应算法的比较。符号 $\uparrow$ 表示随迭代次数增加或接近 1，$\downarrow$ 表示随迭代次数减少，“-”表示没有任何限制，其中“常数”表示 $[0,1$ 中的任意常数。T 表示迭代次数。

优化器	设置	一阶矩参数	二阶矩参数	收敛 rate
Adam [13]	凸	$\downarrow$	常数	no
AMSGrad [22]	凸	非-↑	常数	$\mathcal{O}(1/\sqrt{T})$
Adam [3]	非凸	非-↑	常数	no
Adam [4, 38]	非凸	常数	↑	$\mathcal{O}(\ln T/\sqrt{T})$
Adam [35]	非凸	常数	↑	$\mathcal{O}(\ln T/\sqrt{T}) + \mathcal{O}(\sqrt{D_0})^{**}$
Adam/Yogi [34]	非凸	0	↑	$\mathcal{O}(1/T + 1/b)^{*}$
Adam式 [9]	非凸	↑	-	$\mathcal{O}(1/T) + \mathcal{O}(D_0)^{**}$
AMSGrad [3]	非凸	非↑	↑	$\mathcal{O}(\ln T/\sqrt{T})$
Adan [32]	非凸	↑	↑	$\mathcal{O}(1/T) + \mathcal{O}(D_0)^{**}$
UAdam（我们的方法）	非凸	↑	-	$\mathcal{O}(1/T) + \mathcal{O}(D_0)^{**}$

$^{*}b$ 表示小批量大小。$^{**}D_0$ 来自弱增长假设，即 $\mathbb{E}_t\left[\|g_t - \nabla f(x_t)\|^2\right] \leq D_0 + D_1\|\nabla f(x_t)\|^2$。

然而，以上所有结果都是分别研究 Adam、NAdam 或其变体的收敛性，因此难以进行比较，且不易看出差异。为此，我们进一步提出一个统一框架，用于研究自适应随机动量算法，将 Adam、NAdam 及其变体都作为特例纳入其中。通过这种方式，所提出的统一 Adam（UAdam）建立了

一个统一平台，既可用于分析现有的深度学习随机梯度下降算法，也可用于开发未来的算法。最后，表1总结了一些已有结果，并彰显了 UAdam 的优势与潜力。

1.2. 贡献

本工作的主要贡献可概括如下

- 我们提出了一个用于处理自适应随机动量算法的统一框架，称为 UAdam。该框架结合了自适应学习率方法和统一动量方法，因此将 Adam、AMSGrad、NAdam 和 Adan 等现有深度学习优化器纳入其中，作为其特例。

- 无需对二阶动量参数β_2施加任何限制，只需使一阶动量参数 β_1 足够接近 1，即可确保 UAdam 收敛，这与实际的超参数设置一致。

- 我们证明，在光滑非凸情形下，UAdam 能以 $\mathcal{O}(1/T)$ 的速率收敛到驻点邻域，且该邻域的大小会随着 β 的增大而减小。此外，在附加条件（强增长条件）下，我们可以证明 Adam 收敛到驻点。进一步地，通过选择插值因子 λ，我们可以直接得到 Adam 型算法和 NAdam 型算法的收敛性。

本文其余部分组织如下：第2节介绍相关符号和假设。第3节提出一种称为 UAdam 的自适应随机动量算法统一框架。技术引理、主要收敛结果及其严格证明将在第4节中给出。最后，第5节对本文进行总结。

2. 预备知识

记号。设 $[T]$ 为集合 $\{1, 2, \ldots, T\}$，若无另行说明，$\|\cdot\|$ 表示向量的 ℓ_2 范数或矩阵的谱范数。对于任意$t \in [T]$，我们用 g_t 表示目标函数 f 在第 t次迭代 x_t处的随机梯度。符号 $\mathbb{E}_t[\cdot]$ 表示在给定过去的 $g_1, g_2, \ldots, g_{t-1}$ 的条件下，相对于 g_t 的条件期望，而 $\mathbb{E}[\cdot]$ 表示

相对于底层概率空间的期望。对于任意$x_t \in \mathbb{R}^d$，x_t 的第 i个元素记为 $x_{t,i}$。向量上的所有运算均按逐坐标方式进行，因此对于任意 $x, y \in \mathbb{R}^d$，$p > 0$,$x/y = (x_1/y_1, x_2/y_2, \ldots, x_d/y_d)^T$和 $x^p = (x_1^p, x_2^p, \ldots, x_d^p)^T$。

为了分析 UAdam 的收敛性能，我们引入一些必要的假设。

假设 2.1. 目标函数 f 有下界，且其梯度 ∇f 是 L-Lipschitz 连续的，即

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \ \forall x, y \in \mathbb{R}^d. \tag{4}$$

假设 2.2. 随机梯度 g_t 是真实梯度 $\nabla f(x_t)$ 的无偏估计，即 $\mathbb{E}_t [g_t] = \nabla f(x_t)$。

假设 2.3. 随机梯度 g_t 的方差满足弱增长条件（WGC），即对于某些 D_0, $D_1 > 0$,

$$\mathbb{E}_t \left[\|g_t - \nabla f(x_t)\|^2 \right] \leq D_0 + D_1 \|\nabla f(x_t)\|^2 . \tag{5}$$

备注 2.1. 值得注意的是，前两个假设是标准假设，且在 [3, 4, 32, 38] 中经常使用。当 $D_1 = 0$ 时，假设 2.3 变为标准的有界方差条件。因此，假设 2.3 弱于有界方差条件 [2, 34]。当 $D_1 \neq 0$ 时，基于梯度的算法仅收敛到驻点的一个有界邻域，且邻域大小与 D_0 [12, 35] 成正比。当 $D_0 = 0$ 时，假设 2.3 称为强增长条件（SGC）。

3. 统一自适应随机动量算法

3.1. 随机统一动量算法

假设给定 $x \in \mathbb{R}^d$ 后，它会返回目标函数 f 的随机梯度 $\nabla f(x, \xi)$，该目标函数由问题 (1) 定义，其中 ξ 是一个服从未知分布 $\mathbb{P}$ 的随机变量。我们用 g_t 表示第 t 次迭代 x_t 时的随机梯度 $\nabla f(x_t, \xi_t)$。

通过引入 $\eta = \alpha/(1-\beta)$ 和 $m_t = (x_t - x_{t+1})/\eta$，其中 $m_0 = 0$，随机Heavy Ball（SHB）（2）的更新变为

$$\text{SHB:} \begin{cases} m_t = \beta m_{t-1} + (1-\beta) g_t \\ x_{t+1} = x_t - \eta m_t \end{cases}, \tag{6}$$

此外，我们考虑如下形式的随机 NAG（SNAG）：

$$\text{SNAG: } \begin{cases} m_t = \beta m_{t-1} + (1-\beta) g_t \\ x_{t+1} = x_t - \eta\beta m_t - \eta(1-\beta) g_t \end{cases}, \tag{7}$$

这与 SNAG (3) 等价，但更易于分析（证明见命题 A.1）。因此，(6) 中 SHB 和 (7) 中 SNAG 的更新可以表示为以下随机统一动量（SUM）的形式。

$$\text{SUM}_1\text{: } \begin{cases} m_t = \beta m_{t-1} + (1-\beta) g_t \\ \bar{m}_t = m_t - \tilde{\lambda}(m_t - g_t) \\ x_{t+1} = x_t - \eta \bar{m}_t \end{cases}, \tag{8}$$

其中 $\lambda = (1-\beta)\lambda \in [0, 1]$、 $\lambda \in [0, 1/(1-\beta)]$ 是插值因子，且 β 是动量参数。注意，当 $\lambda = 0$、 $\lambda = 0$时，(8) 式中的 SUM 变为 (6) 式中的 SHB；当 $\lambda = 1$、 $\lambda = 1-\beta$时，(8) 式中的 SUM 变为 (7) 式中的 SNAG；当 $\lambda = 1/(1-\beta)$、 $\lambda = 1$时，(8) 式中的 SUM 变为 SGD。

注释 3.1。Xie *et al*。 [32] 提出了一种 Nesterov动量估计（NME）方法，用于估计随机梯度的一阶矩，其更新公式为

$$\text{NME: } \begin{cases} \bar{m}_t = \beta \bar{m}_{t-1} + (1-\beta)\left(g_t + \beta\left(g_t - g_{t-1}\right)\right) \\ x_{t+1} = x_t - \eta \bar{m}_t \end{cases}, \tag{9}$$

值得一提的是，上述 NME 与式（7）中的 SNAG 在本质上等价。关于这种等价性的详细证明见命题 A.2。

注 3.2. Liu *et al.* [17] 将 SHB 和 SNAG 统一为以下形式

$$\text{SUM}_2\text{: } \begin{cases} m_t = \mu m_{t-1} - \eta_t g_t \\ x_{t+1} = x_t - \lambda \eta_t g_t + (1-\tilde{\lambda}) m_t \end{cases}, \tag{10}$$

其中 $\lambda := (1-\mu)\lambda$。请注意，式 (10) 中的 SUM 在功能上等同于式 (8) 中的 SUM，只是参数有所变化。命题 B.1 中提供了关于这一观察的更详细信息。

3.2. 自适应学习率

接下来，我们研究具有有界性假设的自适应学习率，该假设涵盖了广泛的一类自适应梯度算法，如表 2 所示。

假设 3.1. 自适应学习率 η_t 有上界和下界，即存在 $0 < \eta_l < \eta_u$，使得 $\forall i \in [d], \eta_l \leq \eta_{t,i} \leq \eta_u$，其中 $\eta_{t,i}$ 表示 η_t的第 i 个分量。

表 2：学习率 η_t 的形式及其是否满足假设 3.1

优化器	学习率 η_t	额外假设	η_l 和 η_u
SUM [17]	$\eta_t = \eta$	-	$\eta_l = \eta, \quad \eta_u = \eta$
Adam [13]	$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$, $\eta_t = \eta/\sqrt{v_t} + \epsilon$	$\|g_t\|_\infty \leq G$	$\eta_l = \frac{\eta}{G+\varepsilon}$ $\eta_u = \frac{\eta}{\varepsilon}$
AMSGrad [22]	$\bar{v}_t = \beta_2 \bar{v}_{t-1} + (1-\beta_2) g_t^2$, $v_t = \max(v_{t-1}, \bar{v}_t)$ $\eta_t = \eta/\sqrt{v_t} + \epsilon$	$\|g_t\|_\infty \leq G$	$\eta_l = \frac{\eta}{G+\varepsilon}, \quad \eta_u = \frac{\eta}{\varepsilon}$
AdaFom [3]	$v_t = \frac{1}{t}\sum_{i=1}^{t} g_i^2, \quad \eta_t = \eta/\sqrt{v_t} + \epsilon$	$\|g_t\|_\infty \leq G$	$\eta_l = \frac{\eta}{G+\varepsilon}$ $\eta_u = \frac{\eta}{\varepsilon}$
AdaBound [18]	$\bar{v}_t = \beta_2 \bar{v}_{t-1} + (1-\beta_2) g_t^2$, $v_t = \text{裁剪}(\bar{v}_t, 1/c_u^2, 1/c_l^2)$ $\eta_t = \eta/\sqrt{v_t}$	-	$\eta_l = \eta c_l, \eta_u = \eta c_u$
Yogi [34]	$v_t = v_{t-1} - (1-\beta_2)\,\text{sign}(v_{t-1} - g_t^2) g_t^2$, $\eta_t = \eta/\sqrt{v_t} + \epsilon$	$\|g_t\|_\infty \leq G$	$\eta_l = \frac{\eta}{\sqrt{2}G+\varepsilon}$ $\eta_u = \frac{\eta}{\varepsilon}$
AdaEMA [38]	$v_t = \frac{1}{W_t}\sum_{i=1}^{t} \omega_i g_i^2, W_t = \sum_{i=1}^{t} \omega_i$, $\eta_t = \eta/\sqrt{v_t} + \epsilon$	$\|g_t\|_\infty \leq G$	$\eta_l = \frac{\eta}{G+\varepsilon}$ $\eta_u = \frac{\eta}{\varepsilon}$
Adan [32]	$v_t = (1-\beta_2) v_{t-1}$ $+\beta_2 (g_t + (1-\beta_1)(g_t - g_{t-1}))^2$, $\eta_t = \eta/\sqrt{v_t} + \epsilon$	$\|g_t\|_\infty \leq G/3$	$\eta_l = \frac{\eta}{G+\varepsilon}$ $\eta_u = \frac{\eta}{\varepsilon}$
SAdam [27]	$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$, 软加函数$(x) = \log(1+\exp(\theta x))/\theta$, $\eta_t = \eta/$软加函数$(\sqrt{v_t})$	$\|g_t\|_\infty \leq G$	$\eta_l = \frac{\eta\theta}{\log(1+\exp(\theta G))}$, $\eta_u = \frac{\eta\theta}{\log 2}$

备注 3.3. 在有界随机梯度条件下，Adam 及其变体（如 AMSGrad [22], 、AdaEMA [38], 和 Adan [32] ）可以满足假设 3.1。即使不满足有界性条件，我们仍可以使用梯度裁剪技术 [18] 使假设 3.1 成立。我们

强调，在 Adam 类算法 [3, 4, 34, 38]的收敛分析中，通常需要额外的假设 $\|g_t\|_\infty \le G$。

注释 3.4。现有的收敛分析 [4, 35, 38] 要求二阶动量因子 β_2接近 1，以保证 Adam 收敛。相比之下，我们不对 β_2施加任何限制，只需随机梯度有界即可满足假设 3.1。

3.3. UAdam：统一自适应随机动量算法

本节介绍一种自适应随机动量算法的统一框架，称为 UAdam。该框架有效地将式 (8) 中的 SUM 与一类满足假设 3.1 的自适应学习率方法结合起来。UAdam 算法的伪代码见算法 1。

算法 1 UAdam：统一自适应随机动量算法

参数：一阶矩因子 $\beta \in [0,1)$，插值因子 $\lambda \in [0, 1/(1-\beta)]$, $\tilde{\lambda} = (1-\beta)\lambda$。

初始化：$x_1 \in \mathbb{R}^d$, $m_0 = 0$

对于 $t = 1, 2, \ldots, T$ **do**

  采样一个无偏随机梯度估计量：$g_t = \nabla f(x_t, \xi_t)$

  $m_t = \beta m_{t-1} + (1-\beta) g_t$

  $\bar{m}_t = m_t - \tilde{\lambda}(m_t - g_t)$

  $\eta_t = h_t(g_1, g_2, \ldots, g_t)$（参见表 2 中 η_t 的不同形式）

  $x_{t+1} = x_t - \eta_t \bar{m}_t$

结束循环。

注 3.5. 注意，当插值因子 λ 和自适应学习率 η_t 取不同形式时，UAdam 对应于不同的深度学习算法。例如，若学习率取为 $\eta_t = \eta/\sqrt{v_t} + \epsilon$，其中 $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$，则对于 $\lambda = 0$，UAdam 退化为原始 Adam；而当 $\lambda = 1$ 时，UAdam 则变为 NAdam [6]。另外，若学习率取为 $\eta_t = \eta/\sqrt{v_t} + \epsilon$，其中$\bar{v}_t = \beta_2 \bar{v}_{t-1} + (1-\beta_2) g_t^2$、$v_t = \max(v_{t-1}, \bar{v}_t)$ 且 $\lambda = 0$，则 UAdam 变为 AMSGrad [22]。类似地，当学习率取为 $\eta_t = \eta/\sqrt{v_t} + \epsilon$，其中 $v_t = (1-\beta_2) v_{t-1} + \beta_2 (g_t + (1-\beta_1)(g_t - g_{t-1}))^2$ 且 $\lambda = 1$ 时，由命题 A.2 可知，UAdam 变为 Adan [32]。

4. 主要结果

4.1. 技术引理

接下来，我们给出一些引理，它们在 UAdam 的收敛分析中起着至关重要的作用。收敛分析的关键在于估计随机指数移动平均序列的方差，如引理 4.1 所示。

引理 4.1. 考虑一个随机指数移动平均序列，$m_t = \beta_t m_{t-1} + (1-\beta_t)\, g_t$，其中 $0 \leq \beta_t < 1$。假设假设2.1和2.2成立，则

$$\begin{aligned}\mathbb{E}_t\left[\|m_t - \nabla f(x_t)\|^2\right] \leq & \beta_t \|m_{t-1} - \nabla f(x_{t-1})\|^2 + \frac{\beta_t^2}{1-\beta_t} L^2 \|x_t - x_{t-1}\|^2 \\ & + (1-\beta_t)^2 \mathbb{E}_t\left[\|g_t - \nabla f(x_t)\|^2\right].\end{aligned} \tag{11}$$

证明. 根据 m_t的定义，我们有

$$\begin{aligned}m_t - \nabla f(x_t) &= \beta_t m_{t-1} + (1-\beta_t)\, g_t - \nabla f(x_t) \\ &= \beta_t \left(m_{t-1} - \nabla f(x_{t-1})\right) + (1-\beta_t)\left(g_t - \nabla f(x_t)\right) \\ &\quad + \beta_t \left(\nabla f(x_{t-1}) - \nabla f(x_t)\right).\end{aligned} \tag{12}$$

对（12）取平方范数，可得

$$\begin{aligned}\|m_t - \nabla f(x_t)\|^2 = & \beta_t^2 \|m_{t-1} - \nabla f(x_{t-1})\|^2 \\ & + \beta_t^2 \|\nabla f(x_{t-1}) - \nabla f(x_t)\|^2 + (1-\beta_t)^2 \|g_t - \nabla f(x_t)\|^2 \\ & + 2\beta_t (1-\beta_t) \underbrace{\left\langle m_{t-1} - \nabla f(x_t), g_t - \nabla f(x_t)\right\rangle}_{\spadesuit} \\ & + 2\beta_t^2 \underbrace{\left\langle m_{t-1} - \nabla f(x_{t-1}), \nabla f(x_{t-1}) - \nabla f(x_t)\right\rangle}_{\clubsuit}.\end{aligned} \tag{13}$$

对于 (13) 中的项 $\spadesuit$，在取条件期望时，假设 g_1、……、g_{t-1} 已知，则 x_t 和 m_{t-1} 均为可测的。由假设 2.2，因 $\mathbb{E}_t[g_t] = \nabla f(x_t)$，我们进一步得到

$$\mathbb{E}_t[\spadesuit] = \left\langle m_{t-1} - \nabla f(x_t), \mathbb{E}_t[g_t] - \nabla f(x_t)\right\rangle = 0. \tag{14}$$

对于 (13) 中的项 ♣，由 $\langle a、\ b\rangle \leq \epsilon/2\left\|a\right\|^2+1/2\epsilon\left\|b\right\|^2、\ \forall\epsilon>0$ 这一事实，令 $a=m_{t-1}-\nabla f(x_{t-1})\ b=\nabla f(x_{t-1})-\nabla f(x_t)$，且 $\epsilon=(1-\beta_t)/\beta_t$。于是可得

$$\clubsuit \leq \frac{1-\beta_t}{2\beta_t}\left\|m_{t-1}-\nabla f(x_{t-1})\right\|^2+\frac{\beta_t}{2(1-\beta_t)}\left\|\nabla f(x_{t-1})-\nabla f(x_t)\right\|^2. \quad (15)$$

对式 (13) 取条件期望，并将式 (14) 和 (15) 代入式 (13)，在已知 $g_1、\ldots、g_{t-1}$，且 $x_t、x_{t-1}$ 和 m_{t-1} 可测的条件下，我们有

$$\begin{aligned}\mathbb{E}_t\left[\left\|m_t-\nabla f(x_t)\right\|^2\right] \leq & \beta_t\left\|m_{t-1}-\nabla f(x_{t-1})\right\|^2 \\ & +\frac{\beta_t^2}{1-\beta_t}\left\|\nabla f(x_{t-1})-\nabla f(x_t)\right\|^2 \\ & +(1-\beta_t)^2\mathbb{E}_t\left[\left\|g_t-\nabla f(x_t)\right\|^2\right].\end{aligned} \quad (16)$$

利用 ∇f 是 L-Lipschitz 连续这一事实，我们得到

$$\begin{aligned}\mathbb{E}_t\left[\left\|m_t-\nabla f(x_t)\right\|^2\right] \leq & \beta_t\left\|m_{t-1}-\nabla f(x_{t-1})\right\|^2 \\ & +\frac{\beta_t^2}{1-\beta_t}L^2\left\|x_t-x_{t-1}\right\|^2 \\ & +(1-\beta_t)^2\mathbb{E}_t\left[\left\|g_t-\nabla f(x_t)\right\|^2\right].\end{aligned} \quad (17)$$

证明至此完成。 □

注释 4.1. 将 β_t 替换为 $1-\beta_t$，即可得到引理 4.1 的等价形式。

$$\begin{aligned}\mathbb{E}_t\left[\left\|m_t-\nabla f(x_t)\right\|^2\right] \leq & (1-\beta_t)\left\|m_{t-1}-\nabla f(x_{t-1})\right\|^2 \\ & +\frac{(1-\beta_t)^2}{\beta_t}L^2\left\|x_t-x_{t-1}\right\|^2+\beta_t^2\mathbb{E}_t\left[\left\|g_t-\nabla f(x_t)\right\|^2\right].\end{aligned} \quad (18)$$

由 $\beta_t\in[0,1)$ 可知，我们的估计式 (18) 比 [29] 中的引理 2 更紧。

Le引理 4.2。设 x_t 为 UAda 生成的迭代序列 m

算法。假设 假设 2.1、2.2、2.3 和 3.1 成立。则，

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\Delta_t\right] \leq \mathbb{E}\left[\frac{\Delta_1-\Delta_{T+1}}{1-\beta}\right] + 2\left(1-\beta\right)D_1\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2\right] \\ + \left(1-\beta\right)D_0T + \tilde{\lambda}\left(2\left(1-\beta\right)D_1L^2\eta_u^2 + \frac{\beta^2L^2\eta_u^2}{(1-\beta)^2}\right)\mathbb{E}\left[\sum_{t=1}^{T}\left\|g_t\right\|^2\right] \\ + \left(1-\tilde{\lambda}\right)\left(2\left(1-\beta\right)D_1L^2\eta_u^2 + \frac{\beta^2L^2\eta_u^2}{(1-\beta)^2}\right)\mathbb{E}\left[\sum_{t=1}^{T}\left\|m_t\right\|^2\right].\end{aligned} \tag{19}$$

其中 $\tilde{\lambda} = (1-\beta)\lambda \in [0,1]$、$m_t = \beta m_{t-1} + (1-\beta)\, g_t$、$0 \leq \beta < 1$和 $\Delta_t = \left\|m_t - \nabla f(x_t)\right\|^2$。

证明。令 $\Delta_t = \left\|m_t - \nabla f(x_t)\right\|^2$。然后，应用引理 4.1 和$\beta_t = \beta \in [0,1)$，并取全期望，可得

$$\begin{aligned}\mathbb{E}\left[\Delta_{t+1}\right] \leq & \beta\mathbb{E}\left[\Delta_t\right] + \frac{\beta^2L^2}{1-\beta}\mathbb{E}\left[\left\|x_{t+1}-x_t\right\|^2\right] \\ & + \left(1-\beta\right)^2\mathbb{E}\left[\left\|g_{t+1}-\nabla f(x_{t+1})\right\|^2\right].\end{aligned} \tag{20}$$

重新整理（20）中的各项后，我们得到

$$\begin{aligned}\mathbb{E}\left[\Delta_t\right] \leq & \frac{\mathbb{E}\left[\Delta_t-\Delta_{t+1}\right]}{1-\beta} + \frac{\beta^2L^2}{(1-\beta)^2}\mathbb{E}\left[\left\|x_{t+1}-x_t\right\|^2\right] \\ & + \left(1-\beta\right)\mathbb{E}\left[\left\|g_{t+1}-\nabla f(x_{t+1})\right\|^2\right].\end{aligned} \tag{21}$$

对于 (21) 右侧的第二项，根据算法 1 中的迭代，$x_{t+1} = x_t - \eta_t\bar{m}_t$，$\bar{m}_t = m_t - \lambda(m_t - g_t)$，我们有

$$\begin{aligned}\left\|x_{t+1}-x_t\right\|^2 = \left\|\eta_t\bar{m}_t\right\|^2 &= \left\|\tilde{\lambda}\eta_t g_t + (1-\tilde{\lambda})\eta_t m_t\right\|^2 \\ &\overset{(i)}{\leq} \tilde{\lambda}\|\eta_t g_t\|^2 + (1-\tilde{\lambda})\|\eta_t m_t\|^2 \\ &\overset{(ii)}{\leq} \tilde{\lambda}\eta_u^2\left\|g_t\right\|^2 + (1-\tilde{\lambda})\eta_u^2\left\|m_t\right\|^2,\end{aligned} \tag{22}$$

其中，(i) 利用了 $\|\cdot\|^2$ 的凸性，而 (ii) 则由假设 3.1 得出$\eta_l \leq \eta_{t,i} \leq \eta_u$。

对于 (21) 右侧的第三项，根据假设 2.3，我们有

$$\mathbb{E}\left[\left\|g_{t+1}-\nabla f(x_{t+1})\right\|^2\right] \leq D_0 + D_1\mathbb{E}\left[\left\|\nabla f(x_{t+1})\right\|^2\right]. \tag{23}$$

对于 (23) 右侧的第二项，由于 ∇f 是 L-Lipschitz 连续的，我们得到

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(x_{t+1})\|^2\right] &= \mathbb{E}\left[\|\nabla f(x_{t+1}) - \nabla f(x_t) + \nabla f(x_t)\|^2\right] \\
&\leq 2\mathbb{E}\left[\|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2\right] + 2\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\
&\leq 2L^2\mathbb{E}\left[\|x_{t+1} - x_t\|^2\right] + 2\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\
&\overset{(22)}{\leq} 2\tilde{\lambda}L^2\eta_u^2\mathbb{E}\left[\|g_t\|^2\right] + 2(1-\tilde{\lambda})L^2\eta_u^2\mathbb{E}\left[\|m_t\|^2\right] + 2\mathbb{E}\left[\|\nabla f(x_t)\|^2\right].
\end{aligned} \tag{24}$$

结合（23）和（24），我们得到

$$\begin{aligned}
\mathbb{E}\left[\|g_{t+1} - \nabla f(x_{t+1})\|^2\right] \leq\ & D_0 + 2D_1\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\
& + 2\tilde{\lambda}D_1L^2\eta_u^2\mathbb{E}\left[\|g_t\|^2\right] + 2(1-\tilde{\lambda})D_1L^2\eta_u^2\mathbb{E}\left[\|m_t\|^2\right].
\end{aligned} \tag{25}$$

将（22）和（25）代入（21）可得

$$\begin{aligned}
\mathbb{E}\left[\Delta_t\right] \leq & \frac{\mathbb{E}\left[\Delta_t - \Delta_{t+1}\right]}{1-\beta} + (1-\beta)D_0 + 2(1-\beta)D_1\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \\
& + \tilde{\lambda}\left(2(1-\beta)D_1L^2\eta_u^2 + \frac{\beta^2L^2\eta_u^2}{(1-\beta)^2}\right)\mathbb{E}\left[\|g_t\|^2\right] \\
& + (1-\tilde{\lambda})\left(2(1-\beta)D_1L^2\eta_u^2 + \frac{\beta^2L^2\eta_u^2}{(1-\beta)^2}\right)\mathbb{E}\left[\|m_t\|^2\right].
\end{aligned} \tag{26}$$

对所有迭代 $t \in [T]$ 将上述不等式求和，我们得到 n

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\Delta_t\right] \leq\ & \mathbb{E}\left[\frac{\Delta_1 - \Delta_{T+1}}{1-\beta}\right] + 2(1-\beta)D_1\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\right] \\
& + (1-\beta)D_0T + \tilde{\lambda}\left(2(1-\beta)D_1L^2\eta_u^2 + \frac{\beta^2L^2\eta_u^2}{(1-\beta)^2}\right)\mathbb{E}\left[\sum_{t=1}^{T}\|g_t\|^2\right] \\
& + (1-\tilde{\lambda})\left(2(1-\beta)D_1L^2\eta_u^2 + \frac{\beta^2L^2\eta_u^2}{(1-\beta)^2}\right)\mathbb{E}\left[\sum_{t=1}^{T}\|m_t\|^2\right].
\end{aligned} \tag{27}$$

证明至此完成。 □

引理 4.3. 设 x_t 为 UAdam 算法生成的迭代序列。假设满足假设 2.1，则，

$$\begin{aligned} f(x_{t+1}) \leq & f(x_t) + \frac{\tilde{\lambda}}{2} \left\| \sqrt{\eta_t} \left(g_t - \nabla f(x_t) \right) \right\|^2 + \frac{1-\tilde{\lambda}}{2} \left\| \sqrt{\eta_t} \left(m_t - \nabla f(x_t) \right) \right\|^2 \\ & - \frac{1}{2} \left\| \sqrt{\eta_t} \nabla f(x_t) \right\|^2 - \frac{\tilde{\lambda}}{2} \left\| \sqrt{\eta_t} g_t \right\|^2 + \frac{\tilde{\lambda} L}{2} \left\| \eta_t g_t \right\|^2 \\ & - \frac{1-\tilde{\lambda}}{2} \left\| \sqrt{\eta_t} m_t \right\|^2 + \frac{(1-\tilde{\lambda}) L}{2} \left\| \eta_t m_t \right\|^2, \end{aligned} \tag{28}$$

其中 $\lambda = (1-\beta)\lambda \in [0,1]$ 且 $\beta \in [0,1)$。

证明。由于梯度 ∇f 是 L-Lipschitz 连续的，根据下降引理 [20], 以及 UAdam 算法的迭代，$x_{t+1} = x_t - \eta_t \bar{m}_t$，$\bar{m}_t = m_t - \lambda(m_t - g_t)$，我们有

$$\begin{aligned} f(x_{t+1}) \leq & f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \left\| x_{t+1} - x_t \right\|^2 \\ = & f(x_t) - \langle \nabla f(x_t), \eta_t \bar{m}_t \rangle + \frac{L}{2} \left\| \eta_t \bar{m}_t \right\|^2 \\ = & f(x_t) - \left\langle \nabla f(x_t), \eta_t \left(m_t - \tilde{\lambda}(m_t - g_t) \right) \right\rangle \\ & + \frac{L}{2} \left\| \eta_t \left(m_t - \tilde{\lambda}(m_t - g_t) \right) \right\|^2 \\ = & f(x_t) - \tilde{\lambda} \langle \nabla f(x_t), \eta_t g_t \rangle - (1-\tilde{\lambda}) \langle \nabla f(x_t), \eta_t m_t \rangle \\ & + \frac{L}{2} \left\| \tilde{\lambda} \eta_t g_t + (1-\tilde{\lambda}) \eta_t m_t \right\|^2. \end{aligned} \tag{29}$$

由 $\|\cdot\|^2$ 的凸性：$\|\tilde{\lambda} x + (1-\tilde{\lambda}) y\|^2 \leq \tilde{\lambda} \|x\|^2 + (1-\tilde{\lambda}) \|y\|^2$ 且

事实上， $-2\langle a、b\rangle = \|a-b\|^2 - \|a\|^2 - \|b\|^2$，我们有

$$
\begin{aligned}
f(x_{t+1}) \leq & f(x_t) + \frac{\tilde{\lambda}}{2}\left\|\sqrt{\eta_t}\left(g_t - \nabla f(x_t)\right)\right\|^2 - \frac{\tilde{\lambda}}{2}\left\|\sqrt{\eta_t}\nabla f(x_t)\right\|^2 - \frac{\tilde{\lambda}}{2}\left\|\sqrt{\eta_t}g_t\right\|^2 \\
& + \frac{1-\tilde{\lambda}}{2}\left\|\sqrt{\eta_t}\left(m_t - \nabla f(x_t)\right)\right\|^2 - \frac{1-\tilde{\lambda}}{2}\left\|\sqrt{\eta_t}\nabla f(x_t)\right\|^2 \\
& - \frac{1-\tilde{\lambda}}{2}\left\|\sqrt{\eta_t}m_t\right\|^2 + \frac{\tilde{\lambda}L}{2}\left\|\eta_t g_t\right\|^2 + \frac{(1-\tilde{\lambda})L}{2}\left\|\eta_t m_t\right\|^2 \\
= & f(x_t) + \frac{\tilde{\lambda}}{2}\left\|\sqrt{\eta_t}\left(g_t - \nabla f(x_t)\right)\right\|^2 + \frac{1-\tilde{\lambda}}{2}\left\|\sqrt{\eta_t}\left(m_t - \nabla f(x_t)\right)\right\|^2 \\
& - \frac{1}{2}\left\|\sqrt{\eta_t}\nabla f(x_t)\right\|^2 - \frac{\tilde{\lambda}}{2}\left\|\sqrt{\eta_t}g_t\right\|^2 + \frac{\tilde{\lambda}L}{2}\left\|\eta_t g_t\right\|^2 \\
& - \frac{1-\tilde{\lambda}}{2}\left\|\sqrt{\eta_t}m_t\right\|^2 + \frac{(1-\tilde{\lambda})L}{2}\left\|\eta_t m_t\right\|^2 .
\end{aligned}
\tag{30}
$$

这就完成了证明。 □

4.2. UAdam 在非凸优化中的收敛分析

借助第4.1节中的引理，我们现在着手建立 UAdam 算法的收敛分析。

定理 4.1. 设 x_t 为由 UAdam 生成的迭代序列。假设假设 2.1、2.2、2.3 和 3.1 均成立。令 $0 < 1-\beta \leq \min\left\{\frac{\eta_l}{2\ 2+\lambda}()D_1\eta_u、1\right\}$ 和 $\eta_u \leq \min\left\{\frac{\sqrt[3]{\eta_l}}{2\sqrt[3]{D_1L^2}}、\sqrt[3]{(1-\beta)_{4L^2}}^{2}\eta_l、\sqrt{\frac{\eta_l}{2L}}\right\}$，则有

$$
\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2\right] \leq \mathcal{O}\left(\frac{1}{T}\right) + \mathcal{O}\big((1-\beta)D_0\big). \tag{31}
$$

Proof。根据引理 4.3，利用假设 3.1， $\eta_l \leq \eta_{t,\ i} \leq \eta_u$，以及

$\eta_u^2 \leq \eta_l/(2L)$，我们有

$$\begin{aligned} f(x_{t+1}) \leq & f(x_t) + \frac{\tilde{\lambda}\eta_u}{2} \left\| g_t - \nabla f(x_t) \right\|^2 + \frac{(1-\tilde{\lambda})\eta_u}{2} \left\| m_t - \nabla f(x_t) \right\|^2 \\ & - \frac{\eta_l}{2} \left\| \nabla f(x_t) \right\|^2 + \frac{\tilde{\lambda}(L\eta_u^2 - \eta_l)}{2} \left\| g_t \right\|^2 + \frac{(1-\tilde{\lambda})(L\eta_u^2 - \eta_l)}{2} \left\| m_t \right\|^2 \\ \leq & f(x_t) + \frac{\tilde{\lambda}\eta_u}{2} \left\| g_t - \nabla f(x_t) \right\|^2 + \frac{(1-\tilde{\lambda})\eta_u}{2} \left\| m_t - \nabla f(x_t) \right\|^2 \\ & - \frac{\eta_l}{2} \left\| \nabla f(x_t) \right\|^2 - \frac{\tilde{\lambda}\eta_l}{4} \left\| g_t \right\|^2 - \frac{(1-\tilde{\lambda})\eta_l}{4} \left\| m_t \right\|^2 . \end{aligned} \tag{32}$$

重新整理（32）中的各项，对 $t \in [T]$ 求和，并取全期望，得到

$$\begin{aligned} \frac{\eta_l}{2} \mathbb{E} \left[\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2 \right] \leq & f(x_1) - f_* + \frac{\tilde{\lambda}\eta_u}{2} \mathbb{E} \left[\sum_{t=1}^{T} \left\| g_t - \nabla f(x_t) \right\|^2 \right] \\ & + \frac{(1-\tilde{\lambda})\eta_u}{2} \mathbb{E} \left[\sum_{t=1}^{T} \left\| m_t - \nabla f(x_t) \right\|^2 \right] \\ & - \frac{\tilde{\lambda}\eta_l}{4} \mathbb{E} \left[\sum_{t=1}^{T} \left\| g_t \right\|^2 \right] - \frac{(1-\tilde{\lambda})\eta_l}{4} \mathbb{E} \left[\sum_{t=1}^{T} \left\| m_t \right\|^2 \right] , \end{aligned} \tag{33}$$

其中，f_* 是根据假设 2.1 得到的 f 的下界。令 $\Delta_t = \left\| m_t - \nabla f(x_t) \right\|^2$。然后，根据引理 4.2，我们有

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^{T} \Delta_t \right] \leq & \mathbb{E} \left[\frac{\Delta_1}{1-\beta} \right] + (1-\beta) D_0 T + 2(1-\beta) D_1 \mathbb{E} \left[\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2 \right] \\ & + \tilde{\lambda} \left(2(1-\beta) D_1 L^2 \eta_u^2 + \frac{\beta^2 L^2 \eta_u^2}{(1-\beta)^2} \right) \mathbb{E} \left[\sum_{t=1}^{T} \left\| g_t \right\|^2 \right] \\ & + (1-\tilde{\lambda}) \left(2(1-\beta) D_1 L^2 \eta_u^2 + \frac{\beta^2 L^2 \eta_u^2}{(1-\beta)^2} \right) \mathbb{E} \left[\sum_{t=1}^{T} \left\| m_t \right\|^2 \right] . \end{aligned} \tag{34}$$

将（34）代入（33）可得

$$\begin{aligned}\frac{\eta_l}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2\right] &\leq f(x_1)-f_*+\frac{(1-\tilde{\lambda})\eta_u\Delta_1}{2(1-\beta)}+\frac{(1-\tilde{\lambda})\left(1-\beta\right)D_0\eta_u}{2}T\\ &+\left(1-\tilde{\lambda}\right)\left(1-\beta\right)D_1\eta_u\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2\right]+\frac{\tilde{\lambda}\eta_u}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|g_t-\nabla f(x_t)\right\|^2\right]\\ &+\tilde{\lambda}\left(\left(1-\tilde{\lambda}\right)\left(1-\beta\right)D_1L^2\eta_u^3+\frac{(1-\tilde{\lambda})\beta^2L^2\eta_u^3}{2\left(1-\beta\right)^2}-\frac{\eta_l}{4}\right)\mathbb{E}\left[\sum_{t=1}^{T}\left\|g_t\right\|^2\right]\\ &+\left(1-\tilde{\lambda}\right)\left(\left(1-\tilde{\lambda}\right)\left(1-\beta\right)D_1L^2\eta_u^3+\frac{(1-\tilde{\lambda})\beta^2L^2\eta_u^3}{2\left(1-\beta\right)^2}-\frac{\eta_l}{4}\right)\mathbb{E}\left[\sum_{t=1}^{T}\left\|m_t\right\|^2\right].\end{aligned} \tag{35}$$

记 $\Psi=\left(1-\tilde{\lambda}\right)\left(1-\beta\right)D_1L^2\eta_u^3+\left(1-\tilde{\lambda}\right)2\beta^2L^2\eta_u^3\left(1-\beta\right)_2-\frac{\eta_l}{4}$。则不等式 (35) 可重新整理为

$$\begin{aligned}\frac{\eta_l}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2\right] &\leq f(x_1)-f_*+\frac{(1-\tilde{\lambda})\eta_u\Delta_1}{2\left(1-\beta\right)}+\frac{(1-\tilde{\lambda})\left(1-\beta\right)D_0\eta_u}{2}T\\ &+\left(1-\tilde{\lambda}\right)\left(1-\beta\right)D_1\eta_u\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2\right]+\frac{\tilde{\lambda}\eta_u}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|g_t-\nabla f(x_t)\right\|^2\right]\\ &+\tilde{\lambda}\Psi\mathbb{E}\left[\sum_{t=1}^{T}\left\|g_t\right\|^2\right]+(1-\tilde{\lambda})\Psi\mathbb{E}\left[\sum_{t=1}^{T}\left\|m_t\right\|^2\right].\end{aligned} \tag{36}$$

根据假设 2.3, $\mathbb{E}_t\left[\left\|g_t-\nabla f(x_t)\right\|^2\right]\leq D_0+D_1\left\|\nabla f(x_t)\right\|^2$, 我们有

$$\begin{aligned}\frac{\eta_l}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2\right] &\leq f(x_1)-f_*+\left(\frac{(1-\tilde{\lambda})\left(1-\beta\right)D_0\eta_u}{2}+\frac{\tilde{\lambda}D_0\eta_u}{2}\right)T\\ &+\frac{(1-\tilde{\lambda})\eta_u\Delta_1}{2\left(1-\beta\right)}+\left(\left(1-\tilde{\lambda}\right)\left(1-\beta\right)D_1\eta_u+\frac{\tilde{\lambda}D_1\eta_u}{2}\right)\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla f(x_t)\right\|^2\right]\\ &+\tilde{\lambda}\Psi\mathbb{E}\left[\sum_{t=1}^{T}\left\|g_t\right\|^2\right]+(1-\tilde{\lambda})\Psi\mathbb{E}\left[\sum_{t=1}^{T}\left\|m_t\right\|^2\right].\end{aligned} \tag{37}$$

由于

$$\tilde{\lambda} \in [0,1], \quad \beta \in [0,1), \quad \eta_u^3 \leq \min\left\{\frac{\eta_l}{8D_1L^2}, \frac{(1-\beta)^2\eta_l}{4L^2}\right\}, \tag{38}$$

我们得到

$$(1-\tilde{\lambda})(1-\beta)D_1L^2\eta_u^3 \leq D_1L^2\eta_u^3 \leq D_1L^2\frac{\eta_l}{8D_1L^2} = \frac{\eta_l}{8}, \tag{39}$$

$$\frac{(1-\tilde{\lambda})\beta^2L^2\eta_u^3}{2(1-\beta)^2} \leq \frac{L^2\eta_u^3}{2(1-\beta)^2} \leq \frac{L^2}{2(1-\beta)^2}\frac{(1-\beta)^2\eta_l}{4L^2} = \frac{\eta_l}{8}. \tag{40}$$

结合（39）和（40），不难看出

$$\begin{aligned}\Psi &= (1-\tilde{\lambda})(1-\beta)D_1L^2\eta_u^3 + \frac{(1-\tilde{\lambda})\beta^2L^2\eta_u^3}{2(1-\beta)^2} - \frac{\eta_l}{4} \\ &\leq \frac{\eta_l}{8} + \frac{\eta_l}{8} - \frac{\eta_l}{4} = 0.\end{aligned} \tag{41}$$

这样，（37）可化简为

$$\begin{aligned}\frac{\eta_l}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\right] &\leq f(x_1) - f_* \\ &+ \left(\frac{(1-\tilde{\lambda})(1-\beta)D_0\eta_u}{2} + \frac{\tilde{\lambda}D_0\eta_u}{2}\right)T + \frac{(1-\tilde{\lambda})\eta_u\Delta_1}{2(1-\beta)} \\ &+ \left((1-\tilde{\lambda})(1-\beta)D_1\eta_u + \frac{\tilde{\lambda}D_1\eta_u}{2}\right)\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\right].\end{aligned} \tag{42}$$

由于

$$\tilde{\lambda} = (1-\beta)\lambda \in [0,1], \quad 1-\beta \leq \min\left\{\frac{\eta_l}{2(2+\lambda)D_1\eta_u}, 1\right\}, \tag{43}$$

我们有

$$\begin{aligned}(1-\tilde{\lambda})(1-\beta)D_1\eta_u + \frac{\tilde{\lambda}D_1\eta_u}{2} &\leq \frac{(1-\beta)(2+\lambda)D_1\eta_u}{2} \\ &\leq \frac{(2+\lambda)D_1\eta_u}{2}\frac{\eta_l}{2(2+\lambda)D_1\eta_u} = \frac{\eta_l}{4},\end{aligned} \tag{44}$$

and

$$\begin{aligned}\frac{(1-\tilde{\lambda})(1-\beta)D_0\eta_u}{2}+\frac{\tilde{\lambda}D_0\eta_u}{2}&\leq\frac{D_0\eta_u}{2}\left((1-\beta)+\tilde{\lambda}\right)\\&=\frac{(1-\beta)(1+\lambda)D_0\eta_u}{2}.\end{aligned}\tag{45}$$

最后，将 (44) 和 (45) 代回 (42)，并重新整理各项，得到

$$\begin{aligned}\frac{\eta_l}{4}\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\right]\leq&f(x_1)-f_*+\frac{(1-\tilde{\lambda})\eta_u\Delta_1}{2(1-\beta)}\\&+\frac{(1-\beta)(1+\lambda)D_0\eta_u}{2}T.\end{aligned}\tag{46}$$

将（46）两边同时乘以 $\frac{4}{\eta_l T}$，得到

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\right]\leq&\frac{4(f(x_1)-f_*)}{\eta_l T}+\frac{2(1-\tilde{\lambda})\eta_u\Delta_1}{(1-\beta)\eta_l T}\\&+\frac{2(1-\beta)(1+\lambda)D_0\eta_u}{\eta_l}.\end{aligned}\tag{47}$$

证明至此完成。 □

由定理 4.1 可知，UAdam 收敛到驻点的邻域，且随着 β 增大，邻域的大小会减小。特别地，当满足强增长条件（$D_0 = 0$）和 $\eta_u = K\eta_l$（$K > 1$）时，我们得到以下推论。

推论 4.1. 假设定理 4.1 中的条件对 UAdam 成立（参见算法 1）。当 $\eta_u \leq \min\left\{\frac{1}{2L\sqrt{2KD_1}}, \frac{1-\beta}{2L\sqrt{K}}, \frac{1}{2KL}\right\}$ and $0 < 1-\beta \leq \min\left\{\frac{1}{2}\frac{1}{2+\lambda}()D_1K, 1\right\}$ 时，我们有 ————

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\right]\leq\mathcal{O}\left(\frac{1}{T}\right).\tag{48}$$

证明. 根据定理 4.1，由于 $\eta_u = K\eta_l, K > 1$，因此对于 η_u，我们有

$$\begin{aligned}
&\eta_u^3 \leq \frac{\eta_l}{8D_1L^2} = \frac{\eta_u}{8KD_1L^2} \quad \Leftrightarrow \quad \eta_u \leq \frac{1}{2L\sqrt{2KD_1}},\\
&\eta_u^3 \leq \frac{(1-\beta)^2\eta_l}{4L^2} = \frac{(1-\beta)^2\eta_u}{4KL^2} \quad \Leftrightarrow \quad \eta_u \leq \frac{1-\beta}{2L\sqrt{K}},\\
&\eta_u^2 \leq \frac{\eta_l}{2L} = \frac{\eta_u}{2KL} \quad \Leftrightarrow \quad \eta_u \leq \frac{1}{2KL}.
\end{aligned} \tag{49}$$

此外，由于 $\frac{\eta_l}{\eta_u} = \frac{1}{K}$， β 的取值范围变为

$$0 < 1-\beta \leq \min\left\{\frac{1}{2(2+\lambda)D_1K}, 1\right\}. \tag{50}$$

这就完成了证明。 □

注 4.2. 从算法 1 可知，当 $\lambda = 0$ 时，UAdam 退化为 Adam 型算法；而当 $\lambda = 1$ 时，UAdam 则退化为 NAdam 型算法。由此，根据推论 4.1，我们可以直接得出 Adam 型算法和 NAdam 型算法的收敛性。这充分体现了推论 4.1 的强大功能与通用性，使我们能够立即获得许多热门深度学习算法的收敛结果，例如 AMSGrad、AdaBound、AdaFom 和 Adan 等。这与 [9, 32] 中的收敛结果一致。最后但同样重要的是，我们能够获得比 [3, 4, 35, 38]，中现有收敛结果更快的收敛速度，这归功于假设 3.1 的设定。

5. 结论

我们提出了一种用于设计和分析深度学习中自适应动量优化器的新型统一框架。这个统一平台称为统一Adam（UAdam），它将统一动量方法（包括 SHB 和 SNAG）与一类满足有界性条件的自适应学习率算法相结合。通过利用随机梯度移动平均估计器的方差递推，我们证明了在光滑非凸情形下，UAdam 能以 $\mathcal{O}(1/T)$ 的速率收敛到驻点邻域，并且该邻域的大小会随着 β 的增大而减小。在额外条件（强增长条件）下，我们进一步证明了 Adam 会收敛到驻点。

这些结果表明，对于手头给定的问题，经过适当的超参数选择，Adam 无需修改其更新规则即可收敛。此外，我们对 UAdam 的分析不对二阶矩参数 β_2 施加任何限制，只要求一阶动量参数足够大（接近 1），这与实践中的超参数设置一致。该分析为 Adam 和 NAdam 的收敛提供了新的见解，也为这一领域中新算法的开发提供了一个统一平台。未来的工作将研究偏置梯度条件下 Adam 的收敛性。

附录 A SNAG 的等价形式

命题 A.1. 令 $\bar{x}_t$ 和 m_t 分别表示原始 SNAG 的迭代和动量，其更新公式为

$$\text{SNAG}_1\text{:}\ \begin{cases} \bar{m}_t = \beta \bar{m}_{t-1} - \alpha \nabla f\left(\bar{x}_t + \beta \bar{m}_{t-1}, \xi_t\right) \\ \bar{x}_{t+1} = \bar{x}_t + \bar{m}_t \end{cases}. \tag{51}$$

然后，当 $\alpha = \eta(1-\beta)$ 且 $\bar{m}_t = -\eta m_t$ 时，SNAG_1 等价于

$$\text{SNAG}_2\text{:}\ \begin{cases} m_t = \beta m_{t-1} + (1-\beta)\, g_t \\ x_{t+1} = x_t - \eta\beta m_t - \eta\,(1-\beta)\, g_t \end{cases}. \tag{52}$$

证明。定义 $x_t = \bar{x}_t + \beta \bar{m}_{t-1}$ 和 $g_t = \nabla f\ (x_t,\ \xi_t)$。那么，(51) 中的第一个恒等式变为

$$\bar{m}_t = \beta \bar{m}_{t-1} - \alpha \nabla f\left(x_t, \xi_t\right) = \beta \bar{m}_{t-1} - \alpha g_t. \tag{53}$$

由于 $\alpha = \eta(1-\beta)$ 且 $\bar{m}_t = -\eta m_t$，则 (53) 变为

$$m_t = \frac{-\beta \bar{m}_{t-1} + \alpha g_t}{\eta} = \frac{\beta \eta m_{t-1} + \eta(1-\beta) g_t}{\eta} = \beta m_{t-1} + (1-\beta) g_t. \tag{54}$$

回顾 $x_t = \bar{x}_t + \beta \bar{m}_{t-1}$，我们得到

$$\begin{aligned} x_{t+1} - x_t &= \bar{x}_{t+1} + \beta \bar{m}_t - \bar{x}_t - \beta \bar{m}_{t-1} \\ &\overset{(51)}{=} \bar{m}_t + \beta \bar{m}_t - \beta \bar{m}_{t-1} \\ &= -\eta m_t - \eta\beta m_t + \eta\beta m_{t-1} \\ &\overset{(54)}{=} -\eta m_t - \eta\beta m_t + \eta\left(m_t - (1-\beta)\, g_t\right) \\ &= -\eta\beta m_t - \eta\,(1-\beta)\, g_t, \end{aligned} \tag{55}$$

其中第三个等式由 $\bar{m}_t = -\eta m_t$ 得出。因此，原始 SNAG$_1$ 的最终等价形式为

$$\begin{cases} m_t = \beta m_{t-1} + (1-\beta)\, g_t \\ x_{t+1} = x_t - \eta\beta m_t - \eta\,(1-\beta)\, g_t \end{cases}. \tag{56}$$

证明至此完成。 □

命题 A.2. Xie $et\ al$。 [32] 提出了一种 Nesterov动量估计（NME）方法，如下所示

$$\text{NME:}\ \begin{cases} \bar{m}_t = \beta\bar{m}_{t-1} + (1-\beta)\,(g_t + \beta\,(g_t - g_{t-1})) \\ x_{t+1} = x_t - \eta\bar{m}_t \end{cases}. \tag{57}$$

于是，NME 等价于

$$\text{SNAG:}\ \begin{cases} m_t = \beta m_{t-1} + (1-\beta)\, g_t \\ x_{t+1} = x_t - \eta\beta m_t - \eta\,(1-\beta)\, g_t \end{cases}. \tag{58}$$

证明。根据 SNAG，令 $\bar{m}_t = \beta m_t + (1-\beta)\, g_t$，则 (58) 的第二个等式变为

$$x_{t+1} = x_t - \eta\bar{m}_t. \tag{59}$$

根据 $\bar{m}_t$ 的定义，我们有

$$\begin{aligned} \bar{m}_t - \beta\bar{m}_{t-1} &= \beta m_t + (1-\beta)\, g_t - \beta\,(\beta m_{t-1} + (1-\beta)\, g_{t-1}) \\ &= \beta\,(m_t - \beta m_{t-1}) + (1-\beta)\, g_t - \beta\,(1-\beta)\, g_{t-1} \\ &\overset{(58)}{=} \beta\,(1-\beta)\, g_t + (1-\beta)\, g_t - \beta\,(1-\beta)\, g_{t-1} \\ &= (1-\beta)\,(g_t + \beta\,(g_t - g_{t-1}))\,. \end{aligned} \tag{60}$$

因此，NME 等价于 SNAG。 □

附录B SUM的等价关系

命题 B.1。Liu $et\ al$。 [17] 统一 SHB 和 SNAG 如下

$$\text{SUM}_1\text{:}\ \begin{cases} m_t = \mu m_{t-1} - \eta_t g_t \\ x_{t+1} = x_t - \lambda\eta_t g_t + (1-\tilde{\lambda}) m_t \end{cases}, \tag{61}$$

˜其中 $\lambda := (1-\mu)\lambda \in [0,1]$。当 $\eta_t = \eta(1-\beta)$ 且 $\mu = \beta$ 时，SUM_1 等价于以下统一动量方法

$$\text{SUM}_2\text{:} \quad \begin{cases} m_t = \beta m_{t-1} + (1-\beta) g_t \\ \bar{m}_t = m_t - \tilde{\lambda}(m_t - g_t) \\ x_{t+1} = x_t - \eta \bar{m}_t \end{cases}, \tag{62}$$

˜其中 $\lambda = (1-\beta)\,\lambda \in [0,1]$ 和 $\beta \in [0,1)$。

证明。首先，SUM_1 可以写为

$$\begin{aligned} x_{t+1} &\overset{(61)}{=} x_t - \lambda\eta_t g_t + (1-\tilde{\lambda}) m_t \\ &\overset{(61)}{=} x_t - \lambda\eta_t g_t + (1-\tilde{\lambda}) \left(\mu m_{t-1} - \eta_t g_t\right) \\ &= x_t - \lambda\eta_t g_t - (1-\tilde{\lambda})\eta_t g_t + \mu(1-\tilde{\lambda}) m_{t-1} \\ &\overset{(61)}{=} x_t - \lambda\eta_t g_t - (1-\tilde{\lambda})\eta_t g_t + \mu\left(x_t - x_{t-1} + \lambda\eta_{t-1} g_{t-1}\right) \\ &= x_t - \lambda\eta_t g_t - (1-(1-\mu)\lambda)\eta_t g_t + \mu\left(x_t - x_{t-1} + \lambda\eta_{t-1} g_{t-1}\right). \end{aligned} \tag{63}$$

以类似的方式，SUM_2 可以写为

$$\begin{aligned} x_{t+1} &\overset{(62)}{=} x_t - \eta\bar{m}_t \\ &\overset{(62)}{=} x_t - \eta(m_t - \tilde{\lambda}(m_t - g_t)) \\ &= x_t - \eta\tilde{\lambda} g_t - \eta(1-\tilde{\lambda}) m_t. \end{aligned} \tag{64}$$

将（62）式中的第一个等式代入（64）式的最后一项，我们得到

$$\begin{aligned} x_{t+1} &= x_t - \eta\tilde{\lambda} g_t + \eta(1-\tilde{\lambda})(\beta m_{t-1} + (1-\beta) g_t) \\ &= x_t - \eta\tilde{\lambda} g_t - \eta(1-\tilde{\lambda})(1-\beta) g_t - \eta(1-\tilde{\lambda})\beta m_{t-1} \\ &\overset{(64)}{=} x_t - \eta\tilde{\lambda} g_t - \eta(1-\beta)(1-\tilde{\lambda}) g_t + \beta\left(x_t - x_{t-1} + \eta\tilde{\lambda} g_{t-1}\right) \\ &= x_t - \eta\left(1-\beta\right)\lambda g_t - \eta(1-\beta)(1-(1-\beta)\lambda) g_t \\ &\quad + \beta\left(x_t - x_{t-1} + \eta\left(1-\beta\right)\lambda g_{t-1}\right). \end{aligned} \tag{65}$$

通过比较 (63) 和 (65) 中的系数，可以直接看出，当 $\eta_t = \eta(1-\beta)$ 和 $\mu = \beta$ 时，SUM_2 和 SUM_1 等价。 □

参考文献

[1] Bottou, L., Curtis, F. E., and Nocedal, J.（2018年）。大规模机器学习的优化方法。《SIAM评论》，60(2):223–311。[2] Chen, C., Shen, L., Zou, F., and Liu, W.（2022年）。迈向实用 Adam：非凸性、收敛理论与小批量加速。《机器学习研究杂志》，23(229):1–47。[3] Chen, X., Liu, S., Sun, R., and Hong, M.（2019年）。关于一类用于非凸优化的 Adam 型算法的收敛性。载于《国际学习表征会议论文集》，第 1–30 页。[4] D´efossez, A., Bottou, L., Bach, F., and Usunier, N.（2022年）。Adam 和 Adagrad 的一个简单收敛证明。《机器学习研究汇刊》，第 1–30 页。[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Un- terthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszko- reit, J., and Houlsby, N.（2021年）。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。载于《国际学习表征会议论文集》，第 1–21 页。[6] Dozat, T.（2016年）。将 Nesterov 动量融入 Adam。载于《国际学习表征会议论文集》，第 1–4 页。[7] Duchi, J., Hazan, E., and Singer, Y.（2011年）。用于在线学习和随机优化的自适应次梯度方法。《机器学习研究杂志》，12:2121–2159。[8] Ghadimi, S. and Lan, G.（2013年）。用于非凸随机规划的随机一阶和零阶方法。《SIAM优化杂志》，23(4):2341–2368。[9] Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T.（2021年）。Adam系列算法的一种新颖收敛性分析。载于第 13 届机器学习优化年度研讨会论文集，第 1–13 页。

[10] He, K., Zhang, X., Ren, S., and Sun, J.（2016年）。用于图像识别的深度残差学习。载于 IEEE计算机视觉与模式识别会议（CVPR）论文集，第 770–778 页。[11] Howard, J. and Ruder, S.（2018年）。用于文本分类的通用语言模型微调。载于计算语言学协会第 56 届年会论文集，第 328–339 页。[12] Khaled, A. and Richt´arik, P.（2023年）。非凸领域中 SGD 的更优理论。《机器学习研究汇刊》，第 1–32 页。[13] Kingma, D. and Ba, J.（2015）。Adam：一种随机优化方法。载于第 3 届国际学习表征会议论文集，第 1–15 页。[14] Li, X. and Orabona, F.（2019年）。关于采用自适应步长的随机梯度下降的收敛性。载于第二十二届人工智能与统计国际会议论文集，机器学习研究论文集第 89 卷，第 983–992 页。[15] Lin, Z., Li, H., and Fang, C.（2020）。《机器学习加速优化》。施普林格。[16] Liu, J., Kong, J., Xu, D., Qi, M., and Lu, Y.（2022年）。利用松弛边界函数分析 AdaBound 在非凸优化中的收敛性。《神经网络》，145：300–307。[17] Liu, J., Xu, D., Lu, Y., Kong, J., and Mandic, D. P.（2023年）。神经网络随机动量方法的最后迭代收敛分析。《神经计算》，527：27–35。[18] Luo, L., Xiong, Y., Liu, Y., and Sun, X.（2019年）。具有学习率动态界的自适应梯度方法。载于第 7 届国际学习表征会议论文集，第 1–21 页。[19] Nesterov, Y.（1983）。一种求解凸规划问题的方法及其收敛速度 $\mathcal{O}(1/k^2)$。《苏联数学文献》，27：372–376。

[20] Nesterov, Y.（2013年）。《凸优化导论：基础课程》。施普林格科学与商业媒体。[21] Polyak, B. T.（1964）。若干加速迭代方法收敛的方法。《苏联计算数学与数学物理》，4(5)：1–17。[22] Reddi, S. J., Kale, S., and Kumar, S.（2018年）。关于 Adam 及其后续方法的收敛性。载于国际学习表征会议（ICLR）论文集，第 1–23 页。[23] Robbins, H. E.（1951）。一种随机逼近方法。《数理统计年刊》，22：400–407。[24] Sun, R. Y.（2020）。深度学习优化：综述。《中国运筹学会学报》，8(2)：249–294。[25] Sutskever, I., Martens, J., Dahl, G., and Hinton, G.（2013年）。初始化和动量在深度学习中的重要性。载于国际机器学习会议论文集，第 1139–1147 页。[26] Tieleman, T. and Hinton, G.（2012）。第 6.5 讲——RMSProp：用梯度近期幅度的运行平均值对梯度进行归一化。《Coursera：机器学习神经网络》。[27] Tong, Q., Liang, G., and Bi, J.（2022年）。校准自适应学习率以提升 Adam 的收敛性。《神经计算》，481：333–356。[28]Wang, B., Zhang, Y., Zhang, H., Meng, Q., Ma, Z., Liu, T., and Chen, W.（2022年）。Adam 中可证明的自适应性。arXiv 预印本 arXiv:2208.09900。[29] Wang, M., Fang, E., and Liu, H.（2017）。随机复合梯度下降：用于最小化期望值函数复合的算法。《数学规划》，161：419–449。[30] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Mansor, H., Kartiwi, M., and Ismail, N.（2020）。使用卷积神经网络和深度步幅卷积神经网络进行语音情感识别。载于第 6 届无线与远程通信国际会议（ICWT）论文集，第 1–6 页。

[31] Ward, R., Wu, X., and Bottou, L.（2020年）。Adagrad 步长：非凸景观上的快速收敛。《机器学习研究杂志》，21(219):1–30。[32] Xie, X., Zhou, P., Li, H., Lin, Z., and Yan, S.（2022年）。Adan：用于更快优化深度模型的自适应 Nesterov 动量算法。arXiv 预印本 arXiv:2208.06677。[33] Xu, D., Zhang, S., Zhang, H., and Mandic, D. P.（2021年）。带惩罚项的 RMSProp 深度学习方法在非凸优化中的收敛性。《神经网络》，139:17–23。[34] Zaheer, M., Reddi, S. J., Sachan, D. S., Kale, S., and Kumar, S.（2018年）。非凸优化的自适应方法。载于《神经信息处理系统进展》论文集，第31卷，第9815–9825页。[35] Zhang, Y., Chen, C., Shi, N., Sun, R., and Luo, Z.（2022年）。Adam 无需修改更新规则即可收敛。载于《神经信息处理系统进展》论文集，第1–14页。[36] Zhou, D., Tang, Y., Yang, Z., Cao, Y., and Gu, Q.（2020年）。关于自适应梯度方法在非凸优化中的收敛性。载于第12届机器学习优化年度研讨会论文集，第1–25页。[37] Zou, F., Shen, L., Jie, Z., Sun, J., and Liu, W.（2018年）。带统一动量的加权 Adagrad。arXiv 预印本 arXiv:1808.03408。[38] Zou, F., Shen, L., Jie, Z., Zhang, W., and Liu, W.（2019年）。Adam 和 RMSProp 收敛的一个充分条件。载于 IEEE/CVF 计算机视觉与模式识别会议（CVPR）论文集，第11119–11127页。