arXiv:2603.02649v1 [cs.LG] 3 Mar 2026

HomeAdam：Adam 和 AdamW 算法有时回归本源，以获得更好的可证明泛化

黄飞虎 [1 2] 张冠一 [1] 陈松灿 [1 2]

摘要

Adam 和 AdamW 是机器学习中训练深度学习模型时常用的一类默认优化器。这些自适应算法的收敛速度比随机梯度下降更快，但泛化能力更差。事实上，它们经证明的泛化误差$O(\frac{1}{\sqrt{N}})$也大于随机梯度下降的 $O(\frac{1}{N})$，其中N 表示训练样本量。近期，尽管人们提出了一些 Adam 的变体来改善其泛化能力，但这些改进的泛化性能在理论上仍未得到探究。为填补这一空白，本文通过算法稳定性重新研究 Adam 和 AdamW 的泛化能力，并首次证明，不带平方根的 Adam 和 AdamW（即 Adam(W)-srf）的泛化误差为 $O(\frac{\hat{\rho}^{-2T}}{N})$，其中 T表示迭代次数，$\hat{\rho} > 0$ 表示二阶动量中最小的元素加上一个很小的正数。为改善泛化能力，我们提出了一类高效的巧妙 Adam（即 HomeAdam(W)）算法，通过间歇性地返回基于动量的随机梯度下降来实现这一点。此外，我们证明 HomeAdam(W) 的泛化误差 $O(\frac{1}{N})$小于 Adam(W)-srf 的 $O(\frac{\hat{\rho}^{-2T}}{N})$，因为$\hat{\rho}$ 通常非常小。尤其是，它也小于现有 Adam(W) 的$O(\frac{1}{\sqrt{N}})$。同时，我们证明，与 Adam(W)-srf 的 $O(\frac{\check{\rho}^{-1}}{T^{1/4}})$ 相比，我们的 HomeAdam(W) 具有更快的收敛速率 $O(\frac{1}{T^{1/4}})$，其中 $\check{\rho} \leq \hat{\rho}$也非常小。大量数值实验展示了 HomeAdam(W) 算法的效率。

1. 引言

深度学习模型已在许多机器学习应用中取得巨大成功，例如计算机视觉（LeCun 等人，2015年；He 等人，2016年）、自然语言处理（Vaswani 等人，2017年；Hu 等人，2021年）以及强化学习（Schulman 等人，2017年；Tang 等人，2025年）。事实上，这些深度学习模型的训练主要依赖高效的优化算法（Bottou 等人，2018年）。例如，随机梯度下降（Robbins 与 Monro，1951年；Ghadimi 与 Lan，2013年）是一种基础优化算法，它仅需查询小批量样本即可训练这些大型模型，但容易受到较大方差以及对学习率敏感等问题的影响。为缓解这些缺陷，Kingma 与 Ba（2014年）提出了 Adam，它结合了动量技术和自适应学习率，在训练许多深度学习模型时展现出卓越的性能。尤其是在训练 Transformer 等特定模型时，Adam 的表现显著优于随机梯度下降（Zhang 等人，2024年；2020年）。

由于效率高且对超参数具有鲁棒性，Adam 已成为深度学习中默认优化器之一。因此，其收敛性质在机器学习领域受到了广泛关注。近期，许多研究（Chen 等人，2018b；Zhou 等人，2018；Reddi 等人，2019；Zou 等人，2019；Guo 等人，2021；Jin 等人，2024；D´efossez 等人，2020；Zhang 等人，2022；Wang 等人，2023；Xie 等人，2024；Taniguchi 等人，2024；Peng 等人，2025）研究了 Adam 及其变体的收敛性质。例如，Reddi 等人（2019）发现了 Adam 不收敛的情形，并对 Adam 的一种变体（即 AMSGrad）进行了收敛分析。同时，Chen 等人（2018b）研究了一类用于非凸优化的 Adam 风格算法的收敛性质。随后，Guo 等人（2021）为一系列 Adam 风格方法（包括 Adam 和 AMSGrad）提供了通用的收敛分析，并证明 Adam 在非凸优化中的收敛速率为 $O(\frac{1}{T^{1/4}})$，其中 T表示总迭代次数。Wang 等人（2023）在仅假设光滑条件和有界方差假设的情况下，推导出了 Adam 的新收敛保证。最近，Jin 等人（2024）提出了一个用于分析 Adam 收敛性质的综合分析框架。同时，Jin 等人（2024）研究了 Adam 的收敛性质

[1]南京航空航天大学计算机科学与技术学院，中国南京 [2]工信部模式分析与机器智能重点实验室，中国南京。通讯作者：黄飞虎<huangfeihu2018@gmail.com>。

预印本。*2026*年*3*月*4*日。

表1. **泛化误差** 对我们的 HomeAdam(W) 算法与其他非凸优化算法的比较。此处 N 表示训练样本量。

算法	参考文献	泛化误差	自适应学习率
SGD	(Hardt等人，2016年)	$O(\frac{1}{N})$	
SGDM	(Ramezani-Kebrya等人，2024年)	$O(\frac{1}{N})$	
Adam	(Zhou 等人，2024年)	$O(\frac{1}{\sqrt{N}})$	✓
AdamW	(Zhou 等人，2024年)	$O(\frac{1}{\sqrt{N}})$	✓
HomeAdam	Ours	$O(\frac{1}{N})$	✓
HomeAdamW	Ours	$O(\frac{1}{N})$	✓

及其在广义光滑性假设下的方差缩减变体。

Adam 具有一些良好的经验表现，但在某些深度学习任务上，其泛化性能仍不如随机梯度下降（Wilson et al., 2017）。与此同时，已证明的 Adam 的泛化误差 $O(\frac{1}{\sqrt{N}})$（Zhou et al., 2024）大于随机梯度下降的 $O(\frac{1}{N})$（Hardt et al., 2016）以及基于动量的随机梯度下降（带动量的随机梯度下降）（Ramezani-Kebrya et al., 2024）。（Zou et al., 2023）也给出了一种解释，即 Adam 较差的泛化性能从根本上与深度学习的非凸景观相关。显然，Adam 这种不足的泛化能力阻碍了其更广泛的应用。因此，近期已提出一些自适应梯度方法（Loshchilov & Hutter, 2017; Keskar & Socher, 2017; Chen et al., 2018a; Zhuang et al., 2020; Jin et al., 2025），以改进 Adam 的泛化。例如，AdamW（Loshchilov & Hutter, 2017）通过权重衰减改进了 Adam 的泛化，它也是默认优化器之一。近期，AdamW 的收敛性质已在（Zhou et al., 2024; Xie & Li, 2024; Li et al., 2025）中得到研究，并且也证明了 AdamW 在非凸优化中具有 $O(\frac{1}{T^{1/4}})$ 的收敛速率。此外，（Keskar & Socher, 2017）提出了一种简单策略，即通过触发条件从 Adam 切换到随机梯度下降，以提升泛化性能。（Zhuang et al., 2020）提出了 AdaBelief 优化器，通过利用对当前梯度方向的“信念”来自适应步长，从而改进 Adam 的泛化。更近地，（Jin et al., 2025）提出了多重积分 Adam（即 MIAdam），通过引导优化器走向更平坦的区域来增强泛化能力。

尽管这些提出的方法在一些实证表现中提升了 Adam 的泛化能力，但很少有研究从理论上证明这些方法能够改善泛化。迄今为止，基于 PAC 贝叶斯框架，(Zhou 等人，2024年) 仅证明 AdamW 的泛化误差与 $O(\frac{1}{\sqrt{N}})$ 中的 Adam 相同，但在非凸优化中，这仍大于 SGD（Hardt等人，2016年）和带动量的随机梯度下降（Ramezani-Kebrya等人，2024年）的 $O(\frac{1}{N})$。此外，基于扩散理论框架，(Jin等人，2025年) 仅证明 MiAdam 比 Adam 更有可能逃离尖锐极小值，从而收敛到平坦极小值。显然，Adam 泛化误差得到改善这一点仍缺乏证明。为弥补这一空白，本文通过（Hardt 等人，2016年）所采用的算法稳定性方法，重新研究 Adam 和 AdamW 的泛化能力。我们的主要贡献如下：

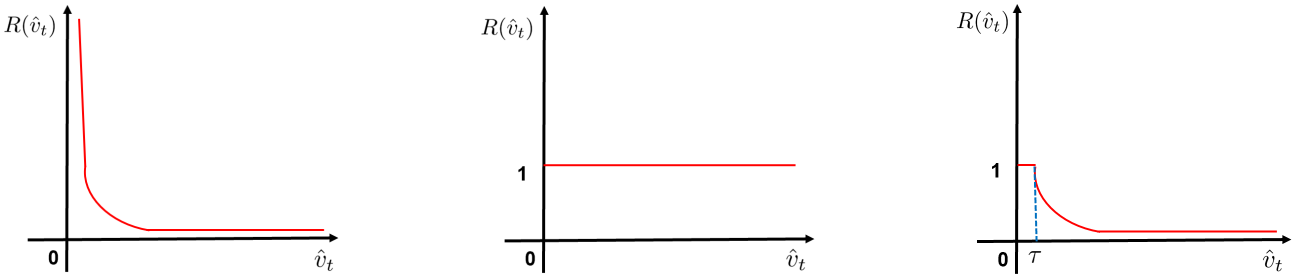
(1) 我们首先通过移除 Lin 等人、2024年；Choudhury 等人，2024年所采用的平方根，提出一类无平方根 Adam（Adam-srf 和 AdamW-srf，即 Adam(W)-srf）算法。同时，我们基于数学归纳法提供了一个实用的泛化分析框架，并证明 Adam(W)-srf 的泛化误差为 $O(\frac{\hat{\rho}^{-2T}}{N})$，其中 T 表示迭代次数，而 $\hat{\rho} > 0$ 表示二阶动量加上一个很小的正数后所得的最小元素。

(2) 为了提升泛化能力，我们进一步提出一类高效的巧妙 Adam（HomeAdam 和 HomeAdamW，即 HomeAdam(W)）算法，通过有时返回带动量的随机梯度下降（SGDM）来实现。这里，SGDM 可以看作 Adam 的一种非自适应变体，类似于 Adam 的归宿。此外，我们证明，HomeAdam(W) 的泛化误差 $O(\frac{1}{N})$ 小于上述误差 $O(\frac{\hat{\rho}^{-2T}}{N})$，因为 $\hat{\rho}$ 通常非常小。特别是，它也小于 Adam 和 AdamW 的现有误差$O(\frac{1}{\sqrt{N}})$（Zhou 等人，2024年）。

(3) 我们还对所提出的方法进行了收敛分析，并证明我们的 HomeAdam(W) 方法具有更快的收敛速率 $O(\frac{1}{T^{1/4}})$，相比于$O(\frac{\breve{\rho}^{-1}}{T^{1/4}})$ 的 Adam(W)-srf，其中 $\breve{\rho} \leq \hat{\rho}$ 也非常小。这与非凸优化中 Adam 和 AdamW 的现有 $O(\frac{1}{T^{1/4}})$ 相同（Guo 等人，2021年；Zhou 等人，2024年）。

(4) 大量数值实验表明，我们的 HomeAdam(W) 算法是有效的。

据我们所知，我们首次证明了自适应梯度方法与随机梯度下降和带动量的随机梯度下降方法具有相同的泛化误差。



(a) Adam(W)-srf/Adam(W) (b) SGD(M) (c) HomeAdam(W)

图 1. 不同步长函数 $R(\hat{v}_t)$ 的示意图：**(a)** Adam(W)-srf 使用 $R(\hat{v}_t) = {}^{1}\hat{v}_t$，Adam(W) 使用 $R(\hat{v}_t) = \frac{1}{\sqrt{\hat{v}_t}}$；**(b)** SGD(M) 使用 $R(\hat{v}_t) = 1$；**(c)** 我们的 HomeAdam(W) 在 $\min_{1\le j\le d}(\hat{v}_t)_j \ge \tau > 0$ 时使用 $R(\hat{v}_t) = {}^{1}\hat{v}_t$，否则使用 $R(\hat{v}_t) = 1$。

符号说明

$[N]$ 表示 $\{1, 2, \cdots, N\}$。$\mathbb{R}^+$ 表示正实数集。$\|\cdot\|$ 表示向量的 ℓ_2 范数和矩阵的谱范数。$\|\cdot\|_1$ 和 $\|\cdot\|_\infty$ 分别表示 ℓ_1范数和ℓ_∞范数。$\langle x,\ y\rangle$ 表示两个向量 $x \in \mathbb{R}^d$ 和 $y \in \mathbb{R}^d$的内积。令 x^γ $(\gamma > 0)$表示逐元素幂运算，x/y 表示逐元素除法。I_d 表示一个 d维单位矩阵。$a_t = O(b_t)$表示对于某个常数 $c > 0$，$a_t \le cb_t$。

2. 预备知识

2.1. 问题设定

在机器学习中，学习的最终目标是最小化总体风险，其定义为

$$\min_{\theta\in\mathbb{R}^d} F(\theta) = \mathbb{E}_{z\sim\mathcal{D}}[f(\theta; z)], \tag{1}$$

其中，$f(\theta; z)$：$\mathbb{R}^d \to \mathbb{R}^+$ 表示样本 $z \sim \mathcal{D}$上的损失函数，该函数可能是非凸的。这里，z 是从某个固定但未知的分布 $\mathcal{D}$中抽取的随机变量。这里，$F(\theta) = \mathbb{E}_{z\sim\mathcal{D}}[f(\theta; z)]$ 表示机器学习任务（例如训练深度学习模型）的总体损失（风险）。由于固定分布 $\mathcal{D}$ 未知，我们只能访问从 $\mathcal{D}$中独立同分布抽取的有限组训练样本$S = \{z_1, z_2, \cdots, z_N\}$。上述总体风险 $F(\theta)$可以用以下经验风险近似

$$F_S(\theta) = \frac{1}{N}\sum_{i=1}^{N} f(\theta; z_i). \tag{2}$$

2.2. 泛化差距

最优模型 θ^* 定义为

$$\theta^* = \arg\min_{\theta\in\mathbb{R}^d} F(\theta).$$

基于训练数据集 $S = \{z_1, z_2, \cdots, z_N\}$，我们运行随机算法 A 来最小化经验风险，以便得到输出模型 $A(S)$。事实上，我们关注超额总体风险 $F(A(S)) - F(\theta^*)$，它衡量的是输出模型相对于最优模型的表现。

在此，我们将超额总体风险分解为以下形式

$$\begin{aligned} F(A(S)) - F(\theta^*) = &\underbrace{F(A(S)) - F_S(A(S))}_{(a)} \\ &+ \underbrace{F_S(A(S)) - F_S(\theta^*)}_{(b)} + F_S(\theta^*) - F(\theta^*), \end{aligned} \tag{3}$$

其中，项(a) $F(A(S)) - F_S(A(S))$ 是泛化误差（泛化差距），用于衡量训练损失与总体损失之间的差距；而项$(b)$$F_S(A(S)) - F_S(\theta^*)$ 是优化误差，用于量化算法最小化经验风险的效果。对随机算法A和训练数据集S取等式（S）的期望，可得

$$\begin{aligned} \mathbb{E}_{A,S}[F(A(S)) - F(\theta^*)] = &\mathbb{E}_{A,S}[F(A(S)) - F_S(A(S))] \\ &+ \mathbb{E}_{A,S}[F_S(A(S)) - F_S(\theta^*)], \end{aligned}$$

其中 $\mathbb{E}_{A,\ S}[F_S(\theta^*) - F(\theta^*)] = 0$，即 θ^* 与 A 和 S相互独立。

2.3. 算法稳定性

本小节将介绍算法稳定性的一些概念。为简化符号，令 $S = \{z_1, z_2, \cdots, z_N\}$和 $S = \{\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_N\}$为从分布 $\mathcal{D}$中独立抽取的两个数据集。对于任意 $i \in [N]$，我们将数据集 $S^{(i)} = \{z_1, z_2, \cdots, \tilde{z}_i, \cdots, z_N\}$ 定义为：用一个独立样本 $\tilde{z}_i$替换第 i个样本 z_i。

定义 2.1.（稳定性）设 A 为一个随机算法，且$A(S)$表示算法 A 在数据集 S上运行的输出。我们称随机算法 A 是 ϵ-一致稳定的，如果对于

任意 S 和 $S^{(i)}$,

$$\sup_z \mathbb{E}_A[f(A(S), z) - f(A(S^{(i)}), z)] \leq \epsilon. \quad (4)$$

引理 2.2。（通过稳定性实现泛化）（*Shalev-Shwartz* 等人，*2010*年；*Hardt*等人，*2016*年）。设算法 A 是 ϵ-一致稳定的函数值算法，则有

$$|\mathbb{E}_{A,S}[F(A(S)) - F_S(A(S))]| \leq \epsilon.$$

算法 1 Adam(W)-srf 算法

输入： $\eta > 0$、β_1，$\beta_2 \in (0, 1)$、$\varepsilon \geq 0$和 $\lambda \geq 0$;
初始化： $\theta_0 \in \mathbb{R}^d$，$m_0 = 0$ 和 $v_0 = 0$;
对于 $t = 1, 2, \ldots, T$**执行**
  抽取一个样本 $z_t \sim \mathcal{D}$;
  $g_t = \nabla f(\theta_{t-1}; z_t)$;
  $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$;
  $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$;
  $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$;
  $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$;
  $\theta_t = \theta_{t-1} - \eta(\frac{\hat{m}_t}{\hat{v}_t+\varepsilon} + \lambda\theta_{t-1})$。
结束循环
输出： θ_T。

3. 我们的 HomeAdam(W)算法

本节首先提出一类无平方根 Adam 和 AdamW 算法（Adam(W)-srf），如算法 1所示。当 $\lambda = 0$时，算法 1为 Adam- srf 算法；否则，算法 1 展示的是 AdamW-srf 算法。事实上，算法1 与 AdamW 算法类似，不同之处在于移除了自适应学习率二阶动量中的平方根。

基于算法1，我们通过有时返回基于动量的随机梯度下降，提出了一类高效的巧妙 Adam（HomeAdam(W)）算法，如算法2所示。当 $\lambda = 0$时，算法2展示 HomeAdam 算法；否则，算法 2 为 HomeAdamW 算法。在算法 2中，我们同样移除了自适应学习率二阶动量中的平方根。特别地，当 $\min_{1\leq j\leq d}(\hat{v}_t)_j \geq \tau$（即二阶动量 $\hat{v}_t$ 中的最小元素大于阈值 $\tau > 0$）时，我们使用自适应随机梯度更新变量 θ_t，如下所示：

$$\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta\frac{\hat{m}_t}{\hat{v}_t + \varepsilon}, \quad (5)$$

否则，我们使用随机梯度更新 θ_t，如下所示：

$$\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta\hat{m}_t. \quad (6)$$

这里我们定义一个作用于二阶动量 $R(\cdot)$ 上的步长函数 $\hat{v}_t$，那么我们可以将算法 2 中的第 11 行和第 13 行统一改写如下：

$$\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta R(\hat{v}_t)\hat{m}_t, \quad (7)$$

其中 $R(\hat{v}_t) = \frac{1}{\hat{v}_t+\varepsilon}$ 在 $\min_{1\leq j\leq d}(\hat{v}_t)_j \geq \tau > 0$时，否则为 $R(\hat{v}_t) = 1$。实际上，我们的 Adam(W)-srf 使用$R(\hat{v}_t) = \frac{1}{\hat{v}_t+\varepsilon}$，Adam(W) 使用 $R(\hat{v}_t) = \frac{1}{\sqrt{\hat{v}_t+\varepsilon}}$，~~而~~ SGD(M) 使用 $R(\hat{v}_t) = 1$。

图 1展示了不同的步长函数 $R(\hat{v}_t)$，不失一般性，设 $\hat{v}_t \in \mathbb{R}$ 为一个标量，并令 ε =0。从图 1可以看出，我们的 HomeAdam(W) 不仅使用自适应学习率，还能避免学习率过大，从而保护其泛化能力和稳定性。然而，当 $\hat{v}_t$ 非常小时，Adam 及其某些变体（例如 Adam(W) 和 Adam(W)-srf）会使用过大的学习率，从而影响其泛化能力和稳定性。

算法 2 HomeAdam(W) 算法

输入： $\eta > 0, \beta_1$，$\beta_2 \in (0, 1)$, $\varepsilon \geq 0, \lambda \geq 0$和 $\tau > 0$;
初始化： $\theta_0 \in \mathbb{R}^d$，$m_0 = 0$ 和 $v_0 = 0$;
对于 $t = 1, 2, \ldots, T$ **do**
  抽取一个样本 $z_t \sim \mathcal{D}$;
  $g_t = \nabla f(\theta_{t-1}; z_t)$;
  $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$;
  $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$;
  $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$;
  $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$;
  若最小值$\min_{1\leq j\leq d}(\hat{v}_t)_j \geq \tau$ **则**
    $\theta_t = \theta_{t-1} - \eta(\frac{\hat{m}_t}{\hat{v}_t+\varepsilon} + \lambda\theta_{t-1})$;
  else
    $\theta_t = \theta_{t-1} - \eta(\hat{m}_t + \lambda\theta_{t-1})$。
  结束 if
结束 for
输出： θ_T。

请注意，SWATS（Keskar 与 Socher，2017年）通过一个切换点将优化过程分为两个阶段：第一阶段使用 Adam 训练模型，第二阶段则使用随机梯度下降训练模型。而我们的 HomeAdam(W) 算法会根据条件$\min_{1\leq j\leq d}(\hat{v}_t)_j < \tau$（或其他情况），在整个优化过程中随时从 Adam(W)-srf 切换到带动量的随机梯度下降（或从带动量的随机梯度下降切换到 Adam(W)-srf）。

在附录F中，我们还提供了 HomeAdam(W) 算法的逐元素变体。由于该变体与反向传播框架（Sampson，1987）相匹配，因此更适合用于训练深度学习模型。

4. 泛化分析

本节在一些温和的假设下，对我们提出的方法（即 Adam(W)-srf 和 HomeAdam(W)）进行泛化分析。所有相关证明均见以下附录A 和B。我们首先给出一些温和的条件。

假设 4.1. 假设每个分量函数 $f(\theta;\ z)$ 都是 L-利普希茨光滑的，则对任意 $z_1, z_2 \in \mathcal{D}$,

$$\|\nabla f(\theta_1; z) - \nabla f(\theta_2; z)\| \leq L\|\theta_1 - \theta_2\|. \tag{8}$$

假设 4.2. 假设每个分量函数 $f(\theta;\ z)$ 对任意 $z \sim \mathcal{D}$ 都是 Lipschitz 连续的，使得

$$|f(\theta_1; z) - f(\theta_2; z)| \leq G\|\theta_1 - \theta_2\|,\ \theta_1, \theta_2 \in \mathbb{R}^d \tag{9}$$

其中 $G > 0$。

假设 4.3。 $\nabla f(\theta;\ z)$ 是全梯度 $\nabla F(\theta)$ 的无偏随机估计量，并且具有有界方差，因此我们有

$$\mathbb{E}[\nabla f(\theta; z)] = \nabla F(\theta),\ \mathbb{E}\|\nabla f(\theta; z) - \nabla F(\theta)\|^2] \leq \sigma^2.$$

假设 4.1 假设每个分量函数 $f(\theta;\ z)$ 都是光滑的，这一假设常用于泛化分析（Zhou et al., 2024; Hardt et al., 2016）。假设 4.2 假设每个分量函数都是 Lipschitz连续的，这一假设常用于泛化分析（Hardt et al., 2016; Lei 和Ying, 2020; Lei, 2023）。假设 4.2 意味着 $f(\theta;\ z)$ 的梯度有界，即对于任意 $z \in \mathcal{D}$, $\|\nabla f(\theta;\ z)\| \leq G$。假设 4.3 给出了随机优化中使用的标准有界方差假设（ Bottou 等人, 2018; Ghadimi 与 Lan, 2013）。根据假设 4.1 和 4.3，我们有$\|\nabla F(\theta_1) - \nabla F(\theta_2)\| = \|\mathbb{E}[\nabla f(\theta_1;\ z) - \nabla f(\theta_2;\ z)]\| \leq \mathbb{E}\|\nabla f(\theta_1;\ z) - \nabla f(\theta_2;\ z)\| \leq L\|\theta_1 - \theta_2\|$。因此，我们可以利用每个分量函数 $f(\theta;\ z)$ 的光滑性，得到函数 $F(\theta)$ 的光滑性。

4.1. Adam(W)-srf 算法的泛化误差

引理4.4。 假设序列 $\{\hat{m}_t、\hat{v}_t、m_t、v_t\}_{t=1}^T$和 $\{\hat{m}_t^{(i)}、\hat{v}_t^{(i)}、m_t^{(i)}、v_t^{(i)}\}_{t=1}^T$ 均由算法 *1* 分别基于数据集 S 和 $S^{(i)}$生成，则有

$$\Big\|\frac{\hat{m}_t}{\hat{v}_t+\varepsilon} - \frac{\hat{m}_t^{(i)}}{\hat{v}_t^{(i)}+\varepsilon}\Big\| \leq \frac{\sqrt{d}}{(\rho_t+\varepsilon)(1-\beta_1^t)}\|m_t - m_t^{(i)}\| + \frac{G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)(\rho_t+\varepsilon)^2}\|v_t - v_t^{(i)}\|, \tag{10}$$

其中 $\rho_t = \min^{j\in[d]}(\hat{v}_t)_j$。

定理 4.5. 假设序列 $\{\theta_t, \hat{v}_t\}_{t=1}^T$ 由算法 *1* 在数据集 $S = \{z_1, z_2, \cdots, z_N\}$上生成，

在假设 *4.1*、*4.2*、*4.3*下，令 $\eta = O(\frac{1}{\sqrt{d}})$,$\lambda \in [0, \frac{1}{\eta})$, $\beta_1 = O(1)$ 其中 $\beta_1 \in (0, 1)$, $\beta_2 = O(1)$ 其中 $\beta_2 \in (0,1)$, $\sigma = O(1)$, $G = O(1)$ 和 $L = O(1)$，我们有

$$|\mathbb{E}[F(\theta_T) - F_S(\theta_T)]| \leq O(\frac{\hat{\rho}^{-2T}}{N}), \tag{11}$$

其中 $\hat{\rho} = \rho + \varepsilon$，且 $\rho = \min^{t\geq 1}\min^{j\in[d]}(\hat{v}_t)_j$。

备注 4.6。 根据上述定理 4.5，当 $\lambda = 0$在算法1中时，我们的 Adam-srf 算法的泛化误差为 $O(\frac{1}{(\rho+\varepsilon)^{2T}N})$；当 $\lambda \in (0, \frac{1}{\eta})$ 在算法 1中时，我们的 AdamW-srf 算法的泛化误差也为$O(\frac{1}{(\rho+\varepsilon)^{2T}N})$。根据定理 4.5的证明，对所有 $t \geq 1$

$$\mathbb{E}\|\theta_{t+1} - \theta_{t+1}^{(i)}\| \leq \frac{\varphi_{t+1}}{N} = O(\frac{1}{(\rho+\varepsilon)^{2(t+1)}N}),$$

$$\varphi_{t+1} = \underbrace{(1-\eta\lambda)\varphi_t}_{(i)} + \frac{\eta\sqrt{d}\phi_{t+1}}{(1-\beta_1^{t+1})(\rho+\varepsilon)} + \frac{\eta G\sqrt{d}\psi_{t+1}}{(1-\beta_1^{t+1})(1-\beta_2^{t+1})(\rho+\varepsilon)^2},$$

其中 $\phi_{t+1} = \beta_1\phi_t + 2(1-\beta_1)\sigma + (1-\beta_1)L\varphi_t$ 且 $\psi_{t+1} = \beta_2\psi_t + 4(1-\beta_2)G\sigma + 2(1-\beta_2)GL\varphi_t$。为简化符号，对于我们的 Adam-srf 算法（即采用 $\lambda = 0$ 的算法 1），令 ϕ_{t+1}^a、ψ_{t+1}^a 和 φ_{t+1}^a；对于我们的 AdamW-srf 算法（即采用 $\lambda \in (0, \frac{1}{\eta})$ 的算法 1），令 ϕ_{t+1}^w、ψ_{t+1}^w和 φ_{t+1}^w。从上面的项（i）中，我们可以发现

$$\phi_{t+1}^w < \phi_{t+1}^a,\ \psi_{t+1}^w < \psi_{t+1}^a,\ \varphi_{t+1}^w < \varphi_{t+1}^a,\ t \geq 1.$$

由于对所有 $t \geq 1$ 都有 $\varphi_{t+1}^w < \varphi_{t+1}^a$，显然，得益于采用解耦权重衰减正则化，我们的 AdamW-srf 的泛化误差小于我们的 Adam-srf。

根据上述结果，Adam(W)-srf 算法的泛化误差取决于二阶动量中的最小元素$\min^{j\in[d]}(\hat{v}_t)_j$。当二阶动量中的元素$(\hat{v}_t)_j$ 非常小时，自适应学习率会变得过大（见图1 (a)），这会影响算法的泛化能力和稳定性。

4.2. 我们的 HomeAdam(W) 算法的泛化误差

定理 *4.7.* 假设序列 $\{\theta_t\}_{t=1}^T$ 是在数据集 $S = \{z_1, z_2, \cdots, z_N\}$上通过算法 *2*生成的，并满足假设 *4.1*、*4.2*、*4.3*。不妨一般性，令$\tau \geq 1$、 $\lambda \in [0, \frac{1}{\eta})$、 $\beta_1 = O(1)$满足 $\beta_1 \in (0, 1)$， $\beta_2 = O(1)$满足 $\beta_2 \in (0, 1)$， $\sigma = O(1)$、$G = O(1)$以及 $L = O(1)$。如果迭代次数较小（即 $T = O(1)$），则设为 $\eta = \frac{1}{\sqrt{d}}$；否则设为$\eta = \frac{1}{\sqrt{dT}}$，则有

$$|\mathbb{E}[F(\theta_T) - F_S(\theta_T)]| \leq O(\frac{1}{N}). \tag{12}$$

备注 4.8： 根据上述定理4.7，当 $\lambda = 0$在算法 2中时，我们的 HomeAdam 算法具有 $O(\frac{1}{N})$的泛化误差。当 $\lambda \in (0, \frac{1}{\eta})$ 在算法 2中时，我们的 HomeAdamW 算法同样具有 $O(\frac{1}{N})$的泛化误差。根据定理 4.7的证明，对于所有$t \geq 1$

$$\mathbb{E}\|\theta_{t+1} - \theta_{t+1}^{(i)}\| \leq \frac{\varphi_{t+1}}{N} = O(\frac{1}{N}),$$

$$\varphi_{t+1} = \underbrace{(1-\eta\lambda)\varphi_t}_{(i)} + \frac{\eta\sqrt{d}\phi_{t+1}}{1-\beta_1^{t+1}} + \frac{\eta G\sqrt{d}\psi_{t+1}}{(1-\beta_1^{t+1})(1-\beta_2^{t+1})},$$

其中 $\phi_{t+1} = \beta_1\phi_t + 2(1-\beta_1)\sigma + (1-\beta_1)L\varphi_t$ 和 $\psi_{t+1} = \beta_2\psi_t + 4(1-\beta_2)G\sigma + 2(1-\beta_2)GL\varphi_t$。类似地，对于我们的 HomeAdam 算法（即使用 $\lambda = 0$ 的算法 2），令 ϕ_{t+1}^a、ψ_{t+1}^a 和 φ_{t+1}^a；对于我们的 HomeAdamW 算法（即使用 $\lambda \in (0, \frac{1}{\eta})$ 的算法 2），令 ϕ_{t+1}^w、ψ_{t+1}^w 和 φ_{t+1}^w。根据上面的项（i），我们可以发现

$$\phi_{t+1}^w < \phi_{t+1}^a,\ \psi_{t+1}^w < \psi_{t+1}^a,\ \varphi_{t+1}^w < \varphi_{t+1}^a,\ t \geq 1.$$

由于对于所有 $t \geq 1$ 都有 $\varphi_{t+1}^w < \varphi_{t+1}^a$，我们的 HomeAdamW 使用解耦权重衰减正则化，因此其泛化误差小于我们的 HomeAdam。

从图 1 (c) 可以看出，我们的 HomeAdam(W) 算法能避免学习率变得过大，从而保护算法的泛化能力和稳定性。因此，与 Adam(W)-srf 算法的泛化误差 $O(\frac{1}{N})$ 相比，我们的 HomeAdam(W) 算法的泛化误差更小；$O(\frac{1}{(\rho+\varepsilon)^{2T}N})$ 同时也小于 Adam 和 AdamW 算法的泛化误差 $O(\frac{1}{\sqrt{N}})$ (Zhou 等人，2024年)。

5. 收敛分析

本节将在一些温和条件下，对我们的 Adam(W)-srf 和 HomeAdam(W) 算法进行收敛分析。所有相关证明均见附录C 和 D。首先给出一些温和的假设。

假设 5.1。 目标函数 $F(\theta)$ 是 L-利普希茨光滑的，满足

$$\|\nabla F(\theta_1) - \nabla F(\theta_2)\| \leq L\|\theta_1 - \theta_2\|,\ \theta_1, \theta_2 \in \mathbb{R}^d. \quad (13)$$

假设 5.2。 目标函数 $F(\theta)$ 的下界有限，即 $F^* = \inf_{\theta\in\mathbb{R}^d} F(\theta) > -\infty$。

假设5.1 表明目标函数具有光滑性$F(\theta)$，这比上述假设 4.1 更宽松。假设 5.2 保证上述问题（ 1 ）具有可行性，这也是优化中常用的假设（Bottou 等人，2018年；Ghadimi 与 Lan，2013）。

5.1. Adam(W)-srf 算法的收敛性质

引理 5.3. 假设序列 $\{m_t\}_{t=0}^T$ 由算法 *1*生成，令 $\beta_1 = 1 - c\eta \in (0,\ 1)$，$\beta_2 \in (0,\ 1)$，则有

$$\mathbb{E}\|\nabla F(\theta_t) - m_{t+1}\|^2 \leq (1-c\eta)\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{2}{c\eta}L^2\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + c^2\eta^2\sigma^2, \quad (14)$$

其中 $c > 0$。

定理 5.4。 假设序列 $\{\theta_t\}_{t=0}^T$ 由算法 *1*生成。在假设 *5.1*、*4.2*、*4.3*、*5.2*下，并令 $0 \leq \lambda < \min(\frac{1}{\eta}, \frac{1}{\eta T^\gamma \hat{G}})$、$\|\theta_0\| \leq \eta\bar{G}$、$c \geq \frac{16L}{\breve{\rho}}$、$\beta_1 = 1 - c\eta \in (0,\ 1)$、$\beta_2 \in (0,\ 1)$ 以及 $0 < \eta \leq \frac{\rho}{4L}$，则有

$$\frac{1}{T+1}\sum_{t=0}^T \mathbb{E}\|\nabla F(\theta_t)\| \leq \frac{4\sqrt{2\Delta}\hat{G}}{\sqrt{T\eta\breve{\rho}}} + \frac{4\sqrt{2}\hat{G}}{T^{\gamma-1}\breve{\rho}} + \frac{4c\sigma\sqrt{\eta}\hat{G}}{\sqrt{L\breve{\rho}}} + \frac{1}{T^{\gamma-1}}, \quad (15)$$

其中 $\bar{G} = \frac{G}{(1-\beta_1)(\rho+\varepsilon)}$，$\hat{G} = \frac{G^2+\varepsilon}{1-\beta_2}$，$\breve{\rho} = (1-\beta_1)(\rho + (1-\beta_2)\varepsilon)$ 以及 $\rho = \min_{t\geq 1}\min_{j\in[d]}(\hat{v}_t)_j$。

注 5.5. 基于上述定理 5.4，令 $\eta = \frac{1}{\sqrt{T}}$和 $\gamma = \frac{3}{4}$，我们可以得到

$$\frac{1}{T+1}\sum_{t=0}^T \mathbb{E}\|\nabla F(\theta_t)\| \leq \frac{4\sqrt{2\Delta}\hat{G}}{\sqrt{\breve{\rho}}T^{1/4}} + \frac{4\sqrt{2}\hat{G}}{\breve{\rho}T^{1/4}} + \frac{4c\sigma\hat{G}}{\sqrt{L\breve{\rho}}T^{1/4}} + \frac{1}{T^{1/4}}.$$

令 $G = O(1)$、$L = O(1)$、$\sigma = O(1)$、$\beta_1 = O(1)$，其中 $\beta_1 \in (0,1)$，令 $\beta_2 = O(1)$，其中 $\beta_2 \in (0,1)$，以及 $c = O(1)$，我们有 $\hat{G} = \frac{G^2+\varepsilon}{1-\beta_2} = O(1)$ 和 $\Delta = F(\theta_0) + \frac{1}{L}(\sigma^2 + \beta_1^2 G^2) - F^* = O(1)$。由于 $\breve{\rho} = (1-\beta_1)(\rho + (1-\beta_2)\varepsilon) \leq \rho + \varepsilon$且 ε 非常小，$\breve{\rho}$也非常小。于是我们有

$$\frac{1}{T+1}\sum_{t=0}^T \mathbb{E}\|\nabla F(\theta_t)\| \leq O(\frac{\breve{\rho}^{-1}}{T^{1/4}}). \quad (16)$$

5.2. 我们的 HomeAdam(W) 算法的收敛性质

定理 5.6。 假设序列 $\{\theta_t\}_{t=0}^T$ 由算法 *2*生成。在假设 *5.1*、*4.2*、*4.3*和*5.2*下，并令 $0 \leq \lambda < \min(\frac{1}{\eta}, \frac{1}{\eta T^\gamma \tilde{G}\hat{G}})$、$\|\theta_0\| \leq \eta\tilde{G}$、$c \geq \frac{32L}{\breve{\rho}}$、$\beta_1 = 1 - c\eta \in (0, 1)$、$\beta_2 \in (0, 1)$、$0 < \eta \leq \frac{\tau}{4L}$和 $\tau > 0$，

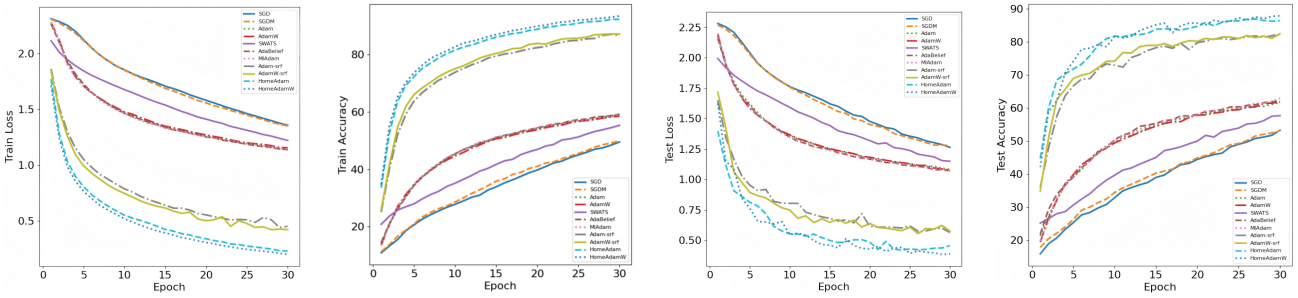


图 2. *CIFAR-10* 数据集上的图像分类结果　　集。

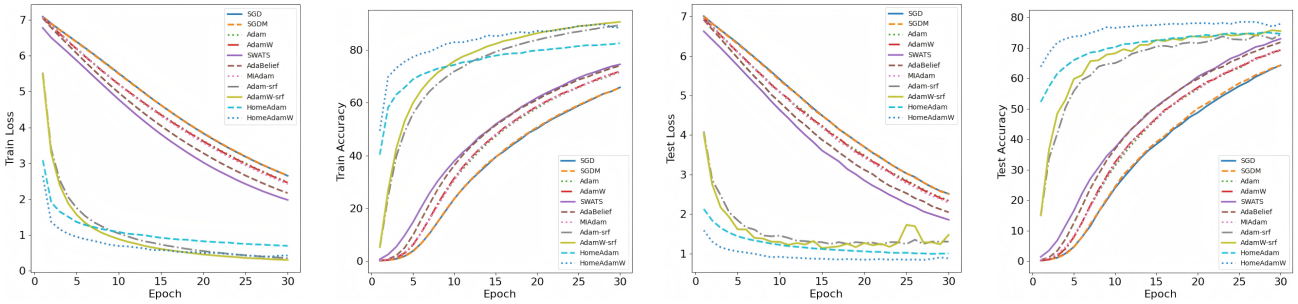


图 3. *Tiny-ImageNet* 数据集上的图像分类结果。

我们有

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| \le \frac{8\sqrt{\Delta}\breve{G}}{\sqrt{T\eta\breve{\tau}}} + \frac{8\breve{G}}{\breve{\tau}T^{\gamma-1}} + \frac{4\sqrt{2\eta}c\sigma\breve{G}}{\sqrt{L\breve{\tau}}} + \frac{\breve{G}}{\hat{G}T^{\gamma-1}}, \quad (17)$$

其中 $\tilde{G} = \max(\frac{G}{(1-\beta_1)(\tau+(1-\beta_2)\varepsilon)}, \frac{G}{1-\beta_1})$，$\hat{G} = \frac{G^2+\varepsilon}{1-\beta_2}$，$\breve{G} = \max(1, \frac{G^2+\varepsilon}{1-\beta_2})$，$\hat{\tau} = (1-\beta_1)(\tau+(1-\beta_2)\varepsilon)$ 和 $\breve{\tau} = \min(1-\beta_1, \hat{\tau})$。

注释 5.7：令 $\eta = \frac{1}{\sqrt{T}}$ 和 $\gamma = \frac{3}{4}$，我们有

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| \le \breve{G}\Big(\frac{8\sqrt{\Delta}}{\sqrt{\breve{\tau}}T^{1/4}} + \frac{8}{\breve{\tau}T^{1/4}} + \frac{4\sqrt{2}c\sigma}{\sqrt{L\breve{\tau}}T^{1/4}} + \frac{1}{\hat{G}T^{1/4}}\Big). \quad (18)$$

设 $G = O(1)$、$L = O(1)$、$\tau = O(1)$ 和 $c = O(1)$，由于 $\eta = \frac{1}{\sqrt{T}}$，我们有 $\beta_1 = 1 - c\eta = O(1)$、$\hat{\tau} = (1-\beta_1)(\tau+(1-\beta_2)\varepsilon) = O(1)$，且 $\breve{\tau} = \min(1-\beta_1, \hat{\tau}) = \min(1-\beta_1, (1-\beta_1)(\tau+(1-\beta_2)\varepsilon)) = O(1)$。于是我们有

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| \le O(\frac{1}{T^{1/4}}). \quad (19)$$

这里，我们的 HomeAadm(W) 算法的收敛速度为 $O(\frac{1}{T^{1/4}})$，快于 Aadm(W)-srf 算法的 $O(\frac{\breve{\rho}^{-1}}{T^{1/4}})$，因为 $\breve{\rho}$ 通常非常小。

6. 数值实验

本节将在计算机视觉（CV）和自然语言处理（NLP）任务上开展一些数值实验，以展示我们算法的有效性。在实验中，我们将 Adam(W)-srf 和 HomeAdam(W) 方法与一些典型方法进行比较，例如 SGD、SGDM、Adam（Kingma &Ba，2014年）、AdamW（Loshchilov&Hutter，2017年）、SWATS（Keskar& Socher，2017年）、AdaBelief（Zhuang 等，2020年）和 MI-Adam（Jin 等，2025年）。所有实验均在一台配备 24 个 vCPU、第 13 代英特尔 Core(TM) i9-13900KF CPU 和 2 块英伟达 RTX 4090 GPU 的机器上运行。

6.1. 计算机视觉任务

在本实验中，我们在 CIFAR-10（Krizhevsky 等人，2009年）和 Tiny-ImageNet（Le 和 Yang，2015年）数据集上实现图像分类任务。具体而言，我们分别在 CIFAR-10 和 Tiny-ImageNet 数据集上训练 VGG16（Simonyan 和 Zisserman，2014年）和 ResNet34（He 等人，2016年）模型。所有算法使用的超参数见附录E。

图 2 和 3 展示了图像分类任务中随训练轮次变化的训练损失、训练准确率、测试损失和测试准确率

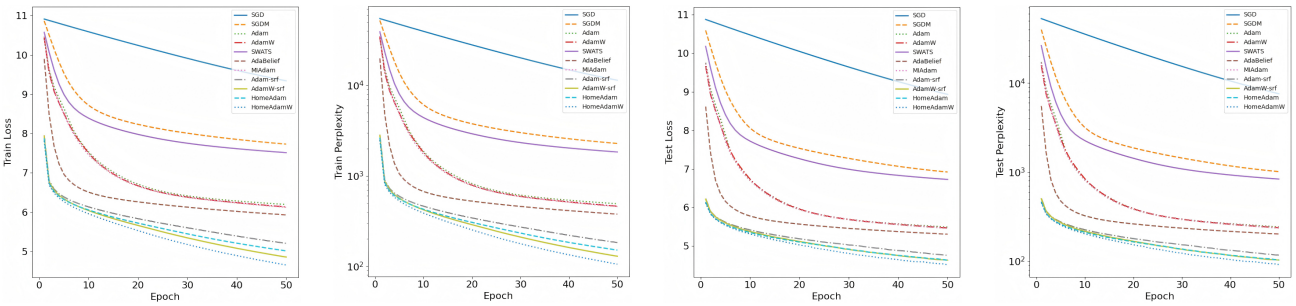


图4. 在*Wikitext-2*数据集上的语言建模任务结果。

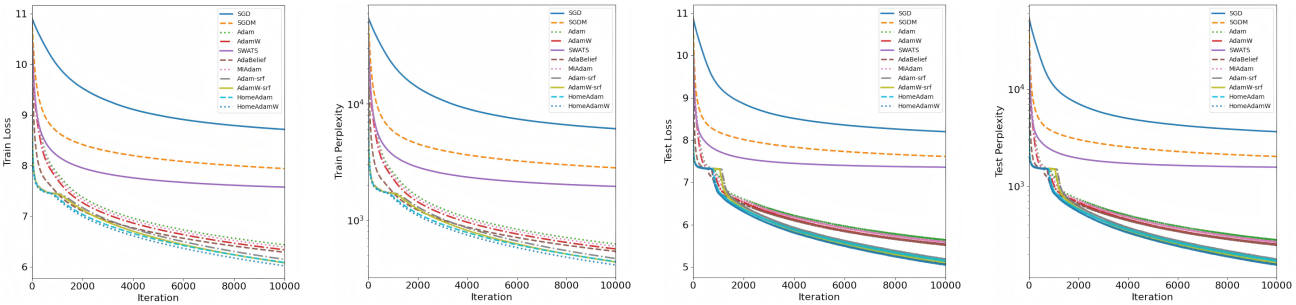


图5. 在*Wikitext-103* 数据集上的语言建模任务结果。

任务，其中每个训练轮次的迭代次数等于训练样本数。从这些结果可以看出，我们的 Adam(W)-srf 和 HomeAdam(W) 算法优于其他对比算法，这验证了在 Adam(W) 中采用无平方根方法的有效性。同时，我们的 HomeAdam(W) 算法的测试准确率高于 Adam(W)-srf 算法，这验证了 HomeAdam(W) 算法具有更好的泛化能力，这一点也得到了上述泛化结果的支持。此外，我们的 HomeAdamW 算法的测试准确率高于 HomeAdam，这验证了通过使用权重衰减，HomeAdamW 算法具有更好的泛化能力，这一点也得到了备注 4.8。

6.2. NLP任务

在本实验中，我们在WikiText2（Merity等人，2016年）和WikiText-103（Merity等人，2016年）数据集上进行语言建模任务。具体而言，我们训练8层Transformer（Vaswani 等人，2017年）和24层Transformer模型，分别在WikiText2和WikiText-103数据集上。8层Transformer和24层Transformer模型描述于附录E。所有算法使用的超参数亦提供于附录E。

图 4 和5 展示了语言建模任务中按训练轮次或迭代次数统计的训练损失、训练困惑度、测试损失和测试困惑度。从这些结果可以看出，我们的 Adam(W)-srf 和 HomeAdam(W) 算法优于其他对比方法，这证明了无平方根在 Adam(W) 中的有效性。同时，我们的 HomeAdam(W) 算法的测试困惑度低于我们的 Adam(W)-srf 算法，这进一步验证了我们的 HomeAdam(W) 算法具有更好的泛化能力，这一点也得到了上述泛化结果的支持。此外，我们的 HomeAdamW 算法的测试困惑度低于我们的 HomeAdam，这验证了通过使用权重衰减，我们的 HomeAdamW 算法具有更好的泛化能力，备注4.8中也对此提供了理论支持。

7. 结论

在本文中，我们重新研究了广受欢迎的 Adam 和 AdamW 算法的泛化性质。我们首先引入了一类无平方根 Adam（即 Adam(W)-srf）算法，并证明我们的 Adam(W)-srf 算法具有 $O(\hat{\rho}_N^{-2T})$ 的泛化误差。为改善其泛化性能，我们进一步设计了一类高效的~~Adam~~（即 HomeAdam(W)）算法，通过在某些情况下返回基于动量的随机梯度下降来实现这一点。此外，我们证明，与 Adam 和 AdamW 算法现有的 $O(\frac{1}{\sqrt{N}})$ 相比，我们的 HomeAdam(W) 方法具有更小的 $O(\frac{1}{N})$ 泛化误差。通过泛化分析，我们还发现，由于在 HomeAdamW 中使用了权重衰减，我们的 HomeAdamW 比我们的 HomeAdam 具有更好的泛化性能。在附录 F 中，我们还提供了 HomeAdam(W) 算法的一个逐元素变体。由于该变体与反向传播框架相匹配，因此更适合训练深度学习模型。

影响声明

本文介绍的工作旨在推动机器学习领域的发展。我们的工作可能对社会产生许多潜在影响，但我们认为其中没有任何影响需要在此特别强调。

参考文献

Bottou, L., Curtis, F. E., 和 Nocedal, J. 面向大规模机器学习的优化方法。*SIAM*评论, 60(2):223–311, 2018年。

Chen, J., Zhou, D., Tang, Y., Yang, Z., Cao, Y., 和 Gu, Q. 缩小训练深度神经网络时自适应梯度方法的泛化差距。*arXiv*预印本 *arXiv:1806.06763*, 2018年a。

Chen, X., Liu, S., Sun, R., and Hong, M. 关于一类用于非凸优化的 Adam 类算法收敛性的研究。*arXiv* 预印本 *arXiv:1808.02941*，2018b。

Choudhury, S., Tupitsa, N., Loizou, N., Horv´ath, S., Takac, M., and Gorbunov, E. 去掉平方根：一种新型高效的尺度不变版Adagrad。神经信息处理系统进展，37:47400– 47431，2024年。

D´efossez, A., Bottou, L., Bach, F., and Usunier, N. Adam 和 Adagrad 的一种简单收敛性证明。*arXiv* 预印本 *arXiv:2003.02395*，2020年。

Ghadimi, S. and Lan, G. 用于非凸随机规划的随机一阶和零阶方法。*SIAM*优化期刊，23(4):2341–2368，2013年。

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. Adam系列算法的一种新型收敛分析。*arXiv* 预印本 *arXiv:2112.03459*，2021年。

Hardt, M., Recht, B., and Singer, Y. 训练更快，泛化更好：随机梯度下降的稳定性。载于国际机器学习会议，第 1225–1234 页。PMLR，2016年。

He, K., Zhang, X., Ren, S., and Sun, J. 用于图像识别的深度残差学习。载于*IEEE* 计算机视觉与模式识别会议论文集，第 770–778 页，2016年。

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA：大语言模型的低秩适配。*arXiv* 预印本 *arXiv:2106.09685*，2021年。

Jin, L.、Nong, H.、Chen, L. 和 Su, Z. 通过多重积分增强 Adam 泛化能力的方法。载于*AAAI* 人工智能会议论文集，第 39 卷，第 4147–4155 页，2025年。

Jin, R.、Li, X.、Yu, Y. 和 Wang, B. 一种全面分析 Adam 收敛性的框架：弥合与随机梯度下降之间的差距。*arXiv* 预印本 *arXiv:2410.04458*，2024年。

Keskar, N. S. 和 Socher, R. 从 Adam 切换到随机梯度下降以提升泛化性能。*arXiv* 预印本 *arXiv:1712.07628*，2017年。Kingma, D. P. 和 Ba, J. Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014年。Krizhevsky, A.、Hinton, G. 等。学习微型图像的多层特征。2009年。

Le, Y. 和 Yang, X. Tiny ImageNet 视觉识别挑战。*CS231N*，7(7):3，2015年。

LeCun, Y.、Bengio, Y.与Hinton, G. 深度学习。《自然》，521(7553):436–444，2015年。

Lei, Y. 非凸与非光滑问题下随机优化的稳定性与泛化。载于第三十六届年度学习理论会议，第191–227页。PMLR，2023年。

Lei, Y.与Ying, Y. 随机梯度下降稳定性与泛化的精细分析。载于国际机器学习会议，第5809–5819页。PMLR，2020年。

Li, H.、Dong, Y.与Lin, Z. 关于以 ℓ_1 范数度量的AdamW的 $o(\frac{\sqrt{d}}{T^{1/4}})$收敛速率。*arXiv*预印本 *arXiv:2505.11840*，2025年。

Lin, W.、Dangel, F.、Eschenhagen, R.、Bae, J.、Turner, R. E.与Makhzani, A. 我们能否去除自适应梯度方法中的平方根?一种二阶视角。*arXiv*预印本 *arXiv:2402.03496*，2024年。

Loshchilov, I. 与 Hutter, F. 解耦权重衰减正则化。*arXiv*预印本 *arXiv:1711.05101*，2017年。

Merity, S.、Xiong, C.、Bradbury, J. 与 Socher, R. 指针哨兵混合模型。*arXiv*预印本 *arXiv:1609.07843*，2016年。

Peng, H.、Qin, S.、Yu, Y.、Jiang, F.、Wang, H. 与 Lin, Z. 从类符号下降视角对 Adam 的简单收敛性证明。*arXiv*预印本 *arXiv:2507.05966*，2025年。

Ramezani-Kebrya, A.、Antonakopoulos, K.、Cevher, V.、Khisti, A. 与 Liang, B. 论带动量的随机梯度下降的泛化。《机器学习研究杂志》，25(22):1–56，2024年。

Reddi, S. J., Kale, S. 和 Kumar, S. 论 Adam 及其后续方法的收敛性。*arXiv* 预印本 *arXiv:1904.09237*，2019年。

Robbins, H. 和 Monro, S. 一种随机逼近方法。《数学统计年刊》，第 400– 407 页，1951年。Sampson, G. 并行分布式处理：认知微观结构探索，1987。Schulman, J., Wolski, F., Dhariwal, P., Radford, A. 和 Klimov, O. 近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017年。

Shalev-Shwartz, S., Shamir, O., Srebro, N., 和 Sridharan, K. 可学习性、稳定性与一致收敛。《机器学习研究杂志》，11:2635–2670，2010年。

Simonyan, K. 和 Zisserman, A. 用于大规模图像识别的超深卷积网络。*arXiv*预印本 *arXiv:1409.1556*，2014年。

Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Mart´ın- Mart´ın, R., 和 Stone, P. 面向机器人学的深度强化学习：真实世界成功案例综述。《控制、机器人与自主系统年度评论》，8 (1):153–188，2025年。

Taniguchi, S., Harada, K., Minegishi, G., Oshima, Y., J eong, S. C., Nagahara, G., Iiyama, T., Suzuki, M., I wasawa, Y., 和 Matsuo, Y. Adopt：改进后的 Adam 可以在任意 β_2 下以最优速率收敛。神经信息处理系统进展，37:72438–72474，2024年。

Vaswani, A.、Shazeer, N.、Parmar, N.、Uszkoreit, J.、Jones, L.、Gomez, A. N.、Kaiser, Ł.与Polosukhin, I. 注意力就是你所需要的一切。神经信息处理系统进展，30，2017年。

Wang, B.、Fu, J.、Zhang, H.、Zheng, N.与Chen, W. 缩小Adam迭代复杂度上界与下界之间的差距。神经信息处理系统进展，36：39006–39032，2023年。

Wilson, A.C.、Roelofs, R.、Stern, M.、Srebro, N.与Recht, B. 自适应梯度方法在机器学习中的边际价值。神经信息处理系统进展，30，2017年。

Xie, S.与Li, Z. AdamW的隐式偏差：ℓ_∞-范数约束优化。收录于国际机器学习会议，第54488–54510页。PMLR，2024年。

Xie, X.、Zhou, P.、Li, H.、Lin, Z. 和 Yan, S. Adan：用于更快优化深度模型的自适应 Nesterov 动量算法。*IEEE* 模式分析与机器智能汇刊，46(12)：9508–9520，2024年。

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S. 和 Sra, S. 为什么自适应方法适用于注意力模型？神经信息处理系统进展，33:15383–15393，2020年。Zhang, Y., Chen, C., Shi, N., Sun, R. 和 Luo, Z.-Q. Adam 无需修改更新规则即可收敛。神经信息处理系统进展，35: 28386–28399，2022年。

Zhang, Y.、Chen, C.、Ding, T.、Li, Z.、Sun, R. 和 Luo, Z. 为什么 Transformer 需要 Adam：海森矩阵视角。神经信息处理系统进展，37：131786–131823，2024年。

Zhou, D.、Chen, J.、Cao, Y.、Yang, Z. 和 Gu, Q. 关于非凸优化中自适应梯度方法的收敛性。*arXiv* 预印本 *arXiv:1808.05671*，2018年。

Zhou, P.、Xie, X.、Lin, Z. 和 Yan, S. 探索 Ada mW 的收敛性与泛化。*IEEE* 模式分析与机器智能汇刊，46(9)：6486–6493，2024年。

Zhuang, J., Tang, T., Ding, Y., Tatikonda, S. C., Dvornek, N., Papademetris, X., and Duncan, J. AdaBelief 优化器：通过对观测梯度的信念来调整步长。神经信息处理系统进展，33：18795–18806，2020年。

Zou, D., Cao, Y., Li, Y., and Gu, Q. 通过适当的正则化理解 Adam 在神经网络学习中的泛化。发表于国际学习表征会议 *(ICLR)*，2023年。

Zou, F., Shen, L., Jie, Z., Zhang, W., 和 Liu, W. Adam 和 RMSprop 收敛的充分条件。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 11127–11135 页，2019年。

A. Adam(W)-srf 算法的泛化分析

本节将详细分析我们提出的 Adam(W)-srf 算法的泛化性能

引理 A.1. （引理 4.4的重述）假设序列 $\{\hat{}m_t,\ \hat{}v_t,\ m_t,\ v_t\}_{t=1}^T$ 和 $\{\hat{}m_t^{(i)},\ \hat{}v_t^{(i)},\ m_t^{(i)},\ v_t^{(i)}\}_{t=1}^T$ 由算法 1 基于数据集 S 和 $S^{(i)}$分别生成，则有

$$\Big\|\frac{\hat{m}_t}{\hat{v}_t+\varepsilon}-\frac{\hat{m}_t^{(i)}}{\hat{v}_t^{(i)}+\varepsilon}\Big\|\le\frac{\sqrt{d}}{(1-\beta_1^t)(\rho_t+\varepsilon)}\big\|m_t-m_t^{(i)}\big\|+\frac{G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)(\rho_t+\varepsilon)^2}\big\|v_t-v_t^{(i)}\big\|,\tag{20}$$

其中 $\rho_t=\min_{j\in[d]}(\hat{}v_t)_j$。

证明。由算法 1，由于 S、$S^{(i)}\sim\mathcal{D}$，我们有 $m_t=\beta_1 m_{t-1}+(1-\beta_1)g_t$、$v_t=\beta_2 v_{t-1}+(1-\beta_2)g_t^2$、$m_t^{(i)}=\beta_1 m_{t-1}^{(i)}+(1-\beta_1)g_t^{(i)}$ 和 $v_t^{(i)}=\beta_2 v_{t-1}^{(i)}+(1-\beta_2)(g_t^{(i)})^2$。同时，我们有 $\hat{}m_t=\frac{m_t}{1-\beta_1^t}$、$\hat{}v_t=\frac{v_t}{1-\beta_2^t}$、$\hat{}m_t^{(i)}=\frac{m_t^{(i)}}{1-\beta_1^t}$ 和 $\hat{}v_t^{(i)}=\frac{v_t^{(i)}}{1-\beta_2^t}$。

令 $(\hat{}m_t)_j$、$(\hat{}v_t)_j$、$(\hat{}m_t^{(i)})_j$ 和 $(\hat{}v_t^{(i)})_j$ 分别表示向量 $\hat{}m_t$、$\hat{}v_t$、$\hat{}m_t^{(i)}$ 和 $\hat{}v_t^{(i)}$ 的第 j 个元素。

由于 $\rho_t=\min_{j\in[d]}(\hat{}v_t)_j$，其中 $\hat{}v_t$ 由算法 1 对任意数据集 $S\sim\mathcal{D}$ 生成，我们有 $\rho_t=\min_{j\in[d]}(\hat{}v_t^{(i)})_j$，其中 $\hat{}v_t^{(i)}$ 由算法 1 基于数据集 $S^{(i)}\sim\mathcal{D}$ 生成。由于 $m_t^{(i)}$ 是 $g_t^{(i)}$ 的指数移动平均，利用假设4.2，我们有 $\|m_t^{(i)}\|\le G$。于是我们有

$$\begin{aligned}
&\Big|\frac{(\hat{m}_t)_j}{(\hat{v}_t)_j+\varepsilon}-\frac{(\hat{m}_t^{(i)})_j}{(\hat{v}_t^{(i)})_j+\varepsilon}\Big|\\
&=\frac{1}{((\hat{v}_t)_j+\varepsilon)((\hat{v}_t^{(i)})_j+\varepsilon)}\big|((\hat{v}_t^{(i)})_j+\varepsilon)(\hat{m}_t)_j-(\hat{m}_t^{(i)})_j((\hat{v}_t)_j+\varepsilon)\big|\\
&\le\frac{1}{((\hat{v}_t)_j+\varepsilon)((\hat{v}_t^{(i)})_j+\varepsilon)}\Big(\big|((\hat{v}_t^{(i)})_j+\varepsilon)(\hat{m}_t)_j-(\hat{m}_t^{(i)})_j((\hat{v}_t^{(i)})_j+\varepsilon)\big|\\
&\qquad+\big|(\hat{m}_t^{(i)})_j((\hat{v}_t^{(i)})_j+\varepsilon)-(\hat{m}_t^{(i)})_j((\hat{v}_t)_j+\varepsilon)\big|\Big)\\
&=\frac{1}{(\hat{v}_t)_j+\varepsilon}\big|(\hat{m}_t)_j-(\hat{m}_t^{(i)})_j\big|+\frac{(\hat{m}_t^{(i)})_j}{((\hat{v}_t)_j+\varepsilon)((\hat{v}_t^{(i)})_j+\varepsilon)}\big|(\hat{v}_t^{(i)})_j-(\hat{v}_t)_j\big|\\
&\le\frac{1}{\rho_t+\varepsilon}\big|(\hat{m}_t)_j-(\hat{m}_t^{(i)})_j\big|+\frac{G}{(1-\beta_1^t)(\rho_t+\varepsilon)^2}\big|(\hat{v}_t^{(i)})_j-(\hat{v}_t)_j\big|\\
&=\frac{1}{(1-\beta_1^t)(\rho_t+\varepsilon)}\big|(m_t)_j-(m_t^{(i)})_j\big|+\frac{G}{(1-\beta_1^t)(1-\beta_2^t)(\rho_t+\varepsilon)^2}\big|(v_t^{(i)})_j-(v_t)_j\big|,
\end{aligned}\tag{21}$$

其中最后一个不等式是由于 $(\hat{}m_t^{(i)})_j\le\|\hat{}m_t^{(i)}\|\le\frac{G}{1-\beta_1^t}$

因此，我们可以得到

$$\Big\|\frac{\hat{m}_t}{\hat{v}_t+\varepsilon}-\frac{\hat{m}_t^{(i)}}{\hat{v}_t^{(i)}+\varepsilon}\Big\|_1\le\frac{1}{(\rho_t+\varepsilon)(1-\beta_1^t)}\big\|m_t-m_t^{(i)}\big\|_1+\frac{G}{(1-\beta_1^t)(1-\beta_2^t)(\rho_t+\varepsilon)^2}\big\|v_t-v_t^{(i)}\big\|_1.\tag{22}$$

由于 $\|\cdot\|\le\|\cdot\|_1\le\sqrt{d}\|\cdot\|$，我们有

$$\begin{aligned}
\Big\|\frac{\hat{m}_t}{\hat{v}_t+\varepsilon}-\frac{\hat{m}_t^{(i)}}{\hat{v}_t^{(i)}+\varepsilon}\Big\|&\le\Big\|\frac{\hat{m}_t}{\hat{v}_t+\varepsilon}-\frac{\hat{m}_t^{(i)}}{\hat{v}_t^{(i)}+\varepsilon}\Big\|_1\\
&\le\frac{1}{(\rho_t+\varepsilon)(1-\beta_1^t)}\big\|m_t-m_t^{(i)}\big\|_1+\frac{G}{(1-\beta_1^t)(1-\beta_2^t)(\rho_t+\varepsilon)^2}\big\|v_t-v_t^{(i)}\big\|_1\\
&\le\frac{\sqrt{d}}{(\rho_t+\varepsilon)(1-\beta_1^t)}\big\|m_t-m_t^{(i)}\big\|+\frac{G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)(\rho_t+\varepsilon)^2}\big\|v_t-v_t^{(i)}\big\|.
\end{aligned}\tag{23}$$

□

定理 *A.2*。（定理 *4.5* 的重述）假设序列 $\{\theta_t, \hat{}v_t\}_{t=1}^T$ 是由算法 *1* 在数据集 $S = \{z_1, z_2, \cdots, z_N\}$ 上生成的。在假设 *4.1*、*4.2*、*4.3* 下，令 $\eta = O(\frac{1}{\sqrt{d}})$、$\lambda \in [0, \frac{1}{\eta}]$、$\beta_1 = O(1)$，其中 $\beta_1 \in [0, 1)$，$\beta_2 = O(1)$，其中 $\beta_2 \in (0,1)$，$\sigma = O(1)$、$G = O(1)$ 和 $L = O(1)$，我们有

$$|\mathbb{E}[F(\theta_T) - F_S(\theta_T)]| \leq O(\frac{1}{(\rho+\varepsilon)^{2T}N}), \tag{24}$$

其中 $\rho = \min_{t\geq 1}\min_{j\in[d]}(\hat{}v_t)_j$。

证明。 在数据集上运行算法 1 S 和 $S^{(i)}$，并使用相同的随机索引序列 $\{j_t\}_{t=1}^T$，并令 $\{\theta_t\}_{t=1}^T$和 $\{\theta_t^{(i)}\}_{t=1}^T$ 由算法1 在 S 和 $S^{(i)}$生成。

根据算法 1, $\theta_t = \theta_{t-1} - \eta(\frac{m_t}{\hat{}v_t+\varepsilon} + \underline{\lambda\theta_{t-1}})$ 和 $\theta_t^{(i)} = \theta_{t-1}^{(i)} - \eta(\frac{\hat{}m_t^{(i)}}{\hat{v}_t^{(i)}+\varepsilon} + \underline{\lambda\theta_{t-1}^{(i)}})$，我们有

$$\theta_t - \theta_t^{(i)} = (1-\eta\lambda)(\theta_{t-1} - \theta_{t-1}^{(i)}) - \eta\big(\frac{\hat{m}_t}{\hat{v}_t+\varepsilon} - \frac{\hat{m}_t^{(i)}}{\hat{v}_t^{(i)}+\varepsilon}\big). \tag{25}$$

于是有

$$\begin{aligned}
\|\theta_t - \theta_t^{(i)}\| &= \|(1-\eta\lambda)(\theta_{t-1} - \theta_{t-1}^{(i)}) - \eta\big(\frac{\hat{m}_t}{\hat{v}_t+\varepsilon} - \frac{\hat{m}_t^{(i)}}{\hat{v}_t^{(i)}+\varepsilon}\big)\| \\
&\leq (1-\eta\lambda)\|\theta_{t-1} - \theta_{t-1}^{(i)}\| + \eta\|\frac{\hat{m}_t}{\hat{v}_t+\varepsilon} - \frac{\hat{m}_t^{(i)}}{\hat{v}_t^{(i)}+\varepsilon}\| \\
&\leq (1-\eta\lambda)\|\theta_{t-1} - \theta_{t-1}^{(i)}\| + \frac{\eta\sqrt{d}}{(1-\beta_1^t)(\rho_t+\varepsilon)}\big\|m_t - m_t^{(i)}\big\| \\
&\quad + \frac{\eta G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)(\rho_t+\varepsilon)^2}\big\|v_t - v_t^{(i)}\big\| \\
&\leq (1-\eta\lambda)\|\theta_{t-1} - \theta_{t-1}^{(i)}\| + \frac{\eta\sqrt{d}}{(1-\beta_1^t)(\rho+\varepsilon)}\big\|m_t - m_t^{(i)}\big\| \\
&\quad + \frac{\eta G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)(\rho+\varepsilon)^2}\big\|v_t - v_t^{(i)}\big\|,
\end{aligned} \tag{26}$$

其中，第一个不等式由 $0 \leq \lambda < \frac{1}{\eta}$ 得出,- 倒数第二个不等式根据引理 4.4 成立，最后一个不等式由 $\rho =$ 最小值 $_{t\geq 1}$最小值$_{j\in[d]}(\hat{}v_t)_j$。

若 $j_1 \neq i$ 以概率 $1 - \frac{1}{N}$ 成立,- 由于 $m_1 = (1-\beta_1)\nabla f(\theta_0; z_{j_1})$，$m_1^{(i)} = (1-\beta_1)\nabla f(\theta_0^{(i)}; z_{j_1})$，$v_1 = (1-\beta_2)(\nabla f(\theta_0; z_{j_1}))^2$、$v_1^{(i)} = (1-\beta_2)(\nabla f(\theta_0^{(i)}; z_{j_1}))^2$ 以及 $\theta_0 = \theta_0^{(i)}$，我们有 $m_1 = m_1^{(i)}$和 $v_1 = v_1^{(i)}$。

如果以 $\frac{1}{N}$ 的概率满足 $j_1 = i$，则有

$$\begin{aligned}
&\mathbb{E}\|m_1 - m_1^{(i)}\| \\
&= \frac{1}{N}\mathbb{E}\|(1-\beta_1)\nabla f(\theta_0; z_i) - (1-\beta_1)\nabla f(\theta_0^{(i)}; \tilde{z}_i)\| \\
&= \frac{(1-\beta_1)}{N}\mathbb{E}\|\nabla f(\theta_0; z_i) - \nabla F(\theta_0) + \nabla F(\theta_0) - \nabla F(\theta_0^{(i)}) + \nabla F(\theta_0^{(i)}) - \nabla f(\theta_0^{(i)}; \tilde{z}_i)\| \\
&\leq \frac{2(1-\beta_1)\sigma}{N} + \frac{(1-\beta_1)}{N}\mathbb{E}\|\nabla F(\theta_0) - \nabla F(\theta_0^{(i)})\| \\
&= \frac{2(1-\beta_1)\sigma}{N},
\end{aligned} \tag{27}$$

其中，最后一个等式是由于 $\theta_0 = \theta_0^{(i)}$。我们还有

$$\begin{aligned}\mathbb{E}\|v_1 - v_1^{(i)}\| &= \frac{1}{N}\mathbb{E}\|(1-\beta_2)(\nabla f(\theta_0; z_i))^2 - (1-\beta_2)\nabla f(\theta_0^{(i)}; \tilde{z}_i)\| \\ &= \frac{(1-\beta_2)}{N}\mathbb{E}\|(\nabla f(\theta_0; z_i))^2 - (\nabla f(\theta_0^{(i)}; \tilde{z}_i))^2\| \\ &\leq \frac{(1-\beta_2)}{N}\big(\mathbb{E}\|(\nabla f(\theta_0; z_i))^2\| + \mathbb{E}\|(\nabla f(\theta_0^{(i)}; \tilde{z}_i))^2\|\big) \\ &\leq \frac{2(1-\beta_2)G^2}{N}, \end{aligned} \tag{28}$$

令 $\phi_1 = 2(1-\beta_1)\sigma$ 和 $\psi_1 = 2(1-\beta_2)G^2$，我们有

$$\mathbb{E}\|m_1 - m_1^{(i)}\| \leq \frac{\phi_1}{N}, \quad \mathbb{E}\|v_1 - v_1^{(i)}\| \leq \frac{\psi_1}{N}. \tag{29}$$

令 $\varphi_1 = \frac{2\eta\sqrt{d}\sigma}{\rho+\varepsilon} + \frac{2\eta\sqrt{d}G^3}{(1-\beta_1)(\rho+\varepsilon)^2}$，由于 $\theta_0 = \theta_0^{(i)}$，我们有

$$\begin{aligned}\mathbb{E}\|\theta_1 - \theta_1^{(i)}\| &\leq (1-\eta\lambda)\mathbb{E}\|\theta_0 - \theta_0^{(i)}\| + \frac{\eta\sqrt{d}}{(1-\beta_1)(\rho+\varepsilon)}\mathbb{E}\big\|m_1 - m_1^{(i)}\big\| \\ &\quad + \frac{\eta G\sqrt{d}}{(1-\beta_1)(1-\beta_2)(\rho+\varepsilon)^2}\mathbb{E}\big\|v_1 - v_1^{(i)}\big\| \\ &\leq \frac{\eta\sqrt{d}}{(1-\beta_1)(\rho+\varepsilon)}\frac{2(1-\beta_1)\sigma}{N} + \frac{\eta G\sqrt{d}}{(1-\beta_1)(1-\beta_2)(\rho+\varepsilon)^2}\frac{2(1-\beta_2)G^2}{N} \\ &= \frac{1}{N}\big(\frac{2\eta\sqrt{d}\sigma}{\rho+\varepsilon} + \frac{2\eta\sqrt{d}G^3}{(1-\beta_1)(\rho+\varepsilon)^2}\big) \\ &= \frac{\varphi_1}{N}. \end{aligned} \tag{30}$$

令 $\eta = O(\frac{1}{\sqrt{d}})$、$\beta_1 = O(1)$、$\beta_2 = O(1)$、$\sigma = O(1)$、$G = O(1)$ 和 $L = O(1)$，我们有

$$\begin{aligned}&\phi_1 = 2(1-\beta_1)\sigma = O(1), \\ &\psi_1 = 2(1-\beta_2)G^2 = O(1), \\ &\varphi_1 = \frac{2\eta\sqrt{d}\sigma}{\rho+\varepsilon} + \frac{2\eta\sqrt{d}G^3}{(1-\beta_1)(\rho+\varepsilon)^2} = O(\frac{1}{(\rho+\varepsilon)^2}), \end{aligned} \tag{31}$$

其中参数 $\rho = \min_{t\geq 1}\min_{j\in[d]}(\hat{v}_t)_j \geq 0$ 和 $\varepsilon > 0$ 通常非常小。因此，我们有

$$\mathbb{E}\|\theta_1 - \theta_1^{(i)}\| \leq \frac{\varphi_1}{N} = O(\frac{1}{(\rho+\varepsilon)^2 N}). \tag{32}$$

若 $j_2 \neq i$ 以概率 $1 - \frac{1}{N}$ 成立，由于 $m_2 = \beta_1 m_1 + (1-\beta_1)\nabla f(\theta_1; z_{j_2})$ 且 $m_2^{(i)} = \beta_1 m_1^{(i)} + (1-\beta_1)\nabla f(\theta_1^{(i)}; z_{j_2})$，我们有

$$\begin{aligned}\mathbb{E}\|m_2 - m_2^{(i)}\| &= (1-\frac{1}{N})\mathbb{E}\|\beta_1(m_1 - m_1^{(i)}) + (1-\beta_1)(\nabla f(\theta_1; z_{j_2}) - \nabla f(\theta_1^{(i)}; z_{j_2}))\| \\ &\leq (1-\frac{1}{N})\big(\beta_1\mathbb{E}\|m_1 - m_1^{(i)}\| + (1-\beta_1)\mathbb{E}\|\nabla f(\theta_1; z_{j_2}) - \nabla f(\theta_1^{(i)}; z_{j_2})\|\big) \\ &\leq (1-\frac{1}{N})\beta_1\frac{\phi_1}{N} + (1-\frac{1}{N})(1-\beta_1)L\mathbb{E}\|\theta_1 - \theta_1^{(i)}\| \\ &\leq (1-\frac{1}{N})\beta_1\frac{\phi_1}{N} + (1-\frac{1}{N})(1-\beta_1)L\frac{\varphi_1}{N}, \end{aligned} \tag{33}$$

其中倒数第二个不等式由假设 4.1 得出。由于 $v_2 = \beta_2 v_1 + (1-\beta_2)(\nabla f(\theta_1; z_{j_2}))^2$ 且 $v_2^{(i)} = \beta_2 v_1^{(i)} + (1-\beta_2)(\nabla f(\theta_1^{(i)}; z_{j_2}))^2$, 我们有

$$\begin{aligned}
\mathbb{E}\|v_2 - v_2^{(i)}\| &= (1-\frac{1}{N})\mathbb{E}\big\|\beta_2(v_1 - v_1^{(i)}) + (1-\beta_2)((\nabla f(\theta_1; z_{j_2}))^2 - (\nabla f(\theta_1^{(i)}; z_{j_2}))^2\big\| \\
&\le (1-\frac{1}{N})\big(\beta_2\mathbb{E}\|v_1 - v_1^{(i)}\| + (1-\beta_2)\mathbb{E}\|(\nabla f(\theta_1; z_{j_2}))^2 - \nabla f(\theta_1; z_{j_2})\nabla f(\theta_1^{(i)}; z_{j_2}) \\
&\quad + \nabla f(\theta_1; z_{j_2})\nabla f(\theta_1^{(i)}; z_{j_2}) - (\nabla f(\theta_1^{(i)}; z_{j_2}))^2\|\big) \\
&\overset{(i)}{\le} (1-\frac{1}{N})\beta_2\frac{\psi_1}{N} + (1-\frac{1}{N})(1-\beta_2)2G\mathbb{E}\|\nabla f(\theta_1; z_{j_2}) - \nabla f(\theta_1^{(i)}; z_{j_2})\| \\
&\le (1-\frac{1}{N})\beta_2\frac{\psi_1}{N} + (1-\frac{1}{N})(1-\beta_2)2GL\mathbb{E}\|\theta_1 - \theta_1^{(i)}\| \\
&\le (1-\frac{1}{N})\beta_2\frac{\psi_1}{N} + (1-\frac{1}{N})(1-\beta_2)2GL\frac{\varphi_1}{N},
\end{aligned} \tag{34}$$

其中上述不等式(i) 由假设 4.2 成立，即 $\|\nabla f(\theta;\ z)\| \le G$ 对所有 $\theta \in \mathbb{R}^d,\ \ z \sim \mathcal{D}$。如果 $j_2 = i$ 以概率 $\frac{1}{N}$，我们有

$$\begin{aligned}
\mathbb{E}\|m_2 - m_2^{(i)}\| &= \frac{1}{N}\mathbb{E}\big\|\beta_1(m_1 - m_1^{(i)}) + (1-\beta_1)(\nabla f(\theta_1; z_i) - \nabla f(\theta_1^{(i)}; \tilde{z}_i))\big\| \\
&\le \frac{1}{N}\beta_1\mathbb{E}\|m_1 - m_1^{(i)}\| + \frac{1}{N}(1-\beta_1)\mathbb{E}\|\nabla f(\theta_1; z_i) - \nabla f(\theta_1^{(i)}; \tilde{z}_i)\| \\
&\le \frac{1}{N}\beta_1\frac{\phi_1}{N} + \frac{1-\beta_1}{N}\mathbb{E}\big\|\nabla f(\theta_1; z_i) - \nabla F(\theta_1) + \nabla F(\theta_1) - \nabla F(\theta_1^{(i)}) \\
&\quad + \nabla F(\theta_1^{(i)}) - \nabla f(\theta_1^{(i)}; \tilde{z}_i)\big\| \\
&\le \frac{1}{N}\beta_1\frac{\phi_1}{N} + \frac{1-\beta_1}{N}\big(\mathbb{E}\|\nabla f(\theta_1; z_i) - \nabla F(\theta_1)\| + \mathbb{E}\|\nabla F(\theta_1) - \nabla F(\theta_1^{(i)})\| \\
&\quad + \mathbb{E}\|\nabla F(\theta_1^{(i)}) - \nabla f(\theta_1^{(i)}; \tilde{z}_i)\|\big) \\
&\overset{(i)}{\le} \frac{1}{N}\frac{\beta_1\phi_1}{N} + \frac{2(1-\beta_1)\sigma}{N} + \frac{(1-\beta_1)L}{N}\mathbb{E}\|\theta_1 - \theta_1^{(i)}\| \\
&\le \frac{1}{N}\frac{\beta_1\phi_1}{N} + \frac{2(1-\beta_1)\sigma}{N} + \frac{(1-\beta_1)L}{N}\frac{\varphi_1}{N},
\end{aligned} \tag{35}$$

其中上述不等式(i)由假设 4.3 成立。我们还有

$$\begin{aligned}
\mathbb{E}\|v_2 - v_2^{(i)}\| &= \frac{1}{N}\mathbb{E}\big\|\beta_2(v_1 - v_1^{(i)}) + (1-\beta_2)((\nabla f(\theta_1; z_i))^2 - (\nabla f(\theta_1^{(i)}; \tilde{z}_i))^2\big\| \\
&\le \frac{1}{N}\big(\beta_2\mathbb{E}\|v_1 - v_1^{(i)}\| + (1-\beta_2)\mathbb{E}\|(\nabla f(\theta_1; z_i))^2 - \nabla f(\theta_1; z_i)\nabla f(\theta_1^{(i)}; \tilde{z}_i) \\
&\quad + \nabla f(\theta_1; z_i)\nabla f(\theta_1^{(i)}; \tilde{z}_i) - (\nabla f(\theta_1^{(i)}; \tilde{z}_i))^2\|\big) \\
&\le \frac{1}{N}\beta_2\frac{\psi_1}{N} + \frac{1}{N}(1-\beta_2)2G\mathbb{E}\|\nabla f(\theta_1; z_i) - \nabla f(\theta_1^{(i)}; \tilde{z}_i)\| \\
&\le \frac{1}{N}\beta_2\frac{\psi_1}{N} + \frac{1}{N}(1-\beta_2)2G\mathbb{E}\|\nabla f(\theta_1; z_i) - \nabla F(\theta_1) + \nabla F(\theta_1) - \nabla F(\theta_1^{(i)}) \\
&\quad + \nabla F(\theta_1^{(i)}) - \nabla f(\theta_1^{(i)}; \tilde{z}_i)\| \\
&\le \frac{1}{N}\beta_2\frac{\psi_1}{N} + \frac{1}{N}(1-\beta_2)2G\big(2\sigma + L\mathbb{E}\|\theta_1 - \theta_1^{(i)}\|\big) \\
&\le \frac{1}{N}\beta_2\frac{\psi_1}{N} + \frac{1}{N}(1-\beta_2)4G\sigma + \frac{1}{N}(1-\beta_2)2GL\frac{\varphi_1}{N},
\end{aligned} \tag{36}$$

其中最后一个不等式由（32）成立。

令 $\phi_2 = \beta_1\phi_1 + 2(1-\beta_1)\sigma + (1-\beta_1)L\varphi_1$，利用上述不等式（33）和（35），我们有

$$\begin{aligned}\mathbb{E}\|m_2 - m_2^{(i)}\| &\le (1-\frac{1}{N})\frac{\beta_1\phi_1}{N} + (1-\frac{1}{N})(1-\beta_1)L\frac{\varphi_1}{N} \\ &\quad + \frac{1}{N}\frac{\beta_1\phi_1}{N} + \frac{2(1-\beta_1)\sigma}{N} + \frac{(1-\beta_1)L}{N}\frac{\varphi_1}{N} \\ &= \frac{\beta_1\phi_1}{N} + \frac{2(1-\beta_1)\sigma}{N} + \frac{(1-\beta_1)L\varphi_1}{N} \\ &= \frac{\phi_2}{N}. \end{aligned} \tag{37}$$

令 $\psi_2 = \beta_2\psi_1 + 4(1-\beta_2)G\sigma + 2(1-\beta_2)GL\varphi_1$，利用上述不等式（34）和（36），可得

$$\begin{aligned}\mathbb{E}\|v_2 - v_2^{(i)}\| &\le (1-\frac{1}{N})\beta_2\frac{\psi_1}{N} + (1-\frac{1}{N})(1-\beta_2)2GL\frac{\varphi_1}{N} \\ &\quad + \frac{1}{N}\beta_2\frac{\psi_1}{N} + \frac{1}{N}(1-\beta_2)4G\sigma + \frac{1}{N}(1-\beta_2)2GL\frac{\varphi_1}{N} \\ &= \frac{\beta_2\psi_1}{N} + \frac{4(1-\beta_2)G\sigma}{N} + \frac{2(1-\beta_2)GL\varphi_1}{N} \\ &= \frac{\psi_2}{N}. \end{aligned} \tag{38}$$

根据上述不等式（26），可得

$$\begin{aligned}\mathbb{E}\|\theta_2 - \theta_2^{(i)}\| &\le (1-\eta\lambda)\mathbb{E}\|\theta_1 - \theta_1^{(i)}\| + \frac{\eta\sqrt{d}}{(1-\beta_1^2)(\rho+\varepsilon)}\mathbb{E}\big\|m_2 - m_2^{(i)}\big\| \\ &\quad + \frac{\eta G\sqrt{d}}{(1-\beta_1^2)(1-\beta_2^2)(\rho+\varepsilon)^2}\mathbb{E}\big\|v_2 - v_2^{(i)}\big\| \\ &\le (1-\eta\lambda)\frac{\varphi_1}{N} + \frac{\eta\sqrt{d}}{(1-\beta_1^2)(\rho+\varepsilon)}\frac{\phi_2}{N} + \frac{\eta G\sqrt{d}}{(1-\beta_1^2)(1-\beta_2^2)(\rho+\varepsilon)^2}\frac{\psi_2}{N} \\ &= \frac{\varphi_2}{N}, \end{aligned} \tag{39}$$

其中 $\varphi_2 = (1-\eta\lambda)\varphi_1 + \frac{\eta\sqrt{d}\phi_2}{(\rho+\varepsilon)(1-\beta_1^2)} + \frac{\eta G\sqrt{d}\psi_2}{(1-\beta_1^2)(1-\beta_2^2)(\rho+\varepsilon)^2}$。

令 $\eta = O(\frac{1}{\sqrt{d}})$, $\lambda \in [0, \frac{1}{\eta})$, $G = O(1)$, $L = O(1)$, $\beta_1 = O(1)$ 和 $\beta_2 = O(1)$，其中 $\beta_1, \beta_2 \in (0,1)$，由于 $\varphi_1 = O(\frac{1}{(\rho+\varepsilon)^2})$,$\phi_1 = O(1)$ 和 $\psi_1 = O(1)$，我们有

$$\begin{aligned}\phi_2 &= \beta_1\phi_1 + 2(1-\beta_1)\sigma + (1-\beta_1)L\varphi_1 = O(\frac{1}{(\rho+\varepsilon)^2}), \\ \psi_2 &= \beta_2\psi_1 + 4(1-\beta_2)G\sigma + 2(1-\beta_2)GL\varphi_1 = O(\frac{1}{(\rho+\varepsilon)^2}), \\ \varphi_2 &= (1-\eta\lambda)\varphi_1 + \frac{\eta\sqrt{d}\phi_2}{(1-\beta_1^2)(\rho+\varepsilon)} + \frac{\eta G\sqrt{d}\psi_2}{(1-\beta_1^2)(1-\beta_2^2)(\rho+\varepsilon)^2} = O(\frac{1}{(\rho+\varepsilon)^4}). \end{aligned} \tag{40}$$

因此，我们有

$$\mathbb{E}\|\theta_2 - \theta_2^{(i)}\| \le \frac{\varphi_2}{N} = O(\frac{1}{(\rho+\varepsilon)^4 N}). \tag{41}$$

根据数学归纳法，我们假设 $\mathbb{E}\|\theta_t - \theta_t^{(i)}\| \le \frac{\varphi_t}{N}$ 满足 $\varphi_t = O(\frac{1}{(\rho+\varepsilon)^{2t}})$，且 $\mathbb{E}\|m_t - m_t^{(i)}\| \le \frac{\phi_t}{N}$ 满足 $\phi_t = O(\frac{1}{(\rho+\varepsilon)^{2(t-1)}})$，以及 $\mathbb{E}\|v_t - v_t^{(i)}\| \le \frac{\psi_t}{N}$ 满足 $\psi_t = O(\frac{1}{(\rho+\varepsilon)^{2(t-1)}})$。

如果 $j_{t+1} \neq i$ 以概率 $1-\frac{1}{N}$，由于 $m_{t+1} = \beta_1 m_t + (1-\beta_1)\nabla f(\theta_t;\ z_{j_{t+1}})$ 和 $m_{t+1}^{(i)} = \beta_1 m_t^{(i)} + (1-\beta_1)\nabla f(\theta_t^{(i)};\ z_{j_{t+1}})$，我们有

$$\begin{aligned}
\mathbb{E}\|m_{t+1} - m_{t+1}^{(i)}\| &= (1-\frac{1}{N})\mathbb{E}\|\beta_1(m_t - m_t^{(i)}) + (1-\beta_1)(\nabla f(\theta_t; z_{j_{t+1}}) - \nabla f(\theta_t^{(i)}; z_{j_{t+1}}))\| \\
&\leq (1-\frac{1}{N})\big(\beta_1\mathbb{E}\|m_t - m_t^{(i)}\| + (1-\beta_1)\mathbb{E}\|\nabla f(\theta_t; z_{j_{t+1}}) - \nabla f(\theta_t^{(i)}; z_{j_{t+1}})\|\big) \\
&\leq (1-\frac{1}{N})\beta_1\frac{\phi_t}{N} + (1-\frac{1}{N})(1-\beta_1)L\mathbb{E}\|\theta_t - \theta_t^{(i)}\| \\
&\leq (1-\frac{1}{N})\beta_1\frac{\phi_t}{N} + (1-\frac{1}{N})(1-\beta_1)L\frac{\varphi_t}{N},
\end{aligned} \tag{42}$$

其中倒数第二个不等式由假设4.1成立。由于 $v_{t+1} = \beta_2 v_t + (1-\beta_2)(\nabla f(\theta_t; z_{j_{t+1}}))^2$ 且 $v_{t+1}^{(i)} = \beta_2 v_t^{(i)} + (1-\beta_2)(\nabla f(\theta_t^{(i)}; z_{j_{t+1}}))^2$，我们有

$$\begin{aligned}
\mathbb{E}\|v_{t+1} - v_{t+1}^{(i)}\| &= (1-\frac{1}{N})\mathbb{E}\big\|\beta_2(v_t - v_t^{(i)}) + (1-\beta_2)((\nabla f(\theta_t; z_{j_{t+1}}))^2 - (\nabla f(\theta_t^{(i)}; z_{j_{t+1}}))^2\big\| \\
&\leq (1-\frac{1}{N})\big(\beta_2\mathbb{E}\|v_t - v_t^{(i)}\| + (1-\beta_2)\mathbb{E}\|(\nabla f(\theta_t; z_{j_{t+1}}))^2 - \nabla f(\theta_t; z_{j_{t+1}})\nabla f(\theta_t^{(i)}; z_{j_{t+1}}) \\
&\quad + \nabla f(\theta_t; z_{j_{t+1}})\nabla f(\theta_t^{(i)}; z_{j_{t+1}}) - (\nabla f(\theta_t^{(i)}; z_{j_{t+1}}))^2\|\big) \\
&\leq (1-\frac{1}{N})\beta_2\frac{\psi_t}{N} + (1-\frac{1}{N})(1-\beta_2)2G\mathbb{E}\|\nabla f(\theta_t; z_{j_{t+1}}) - \nabla f(\theta_t^{(i)}; z_{j_{t+1}})\| \\
&\leq (1-\frac{1}{N})\beta_2\frac{\psi_t}{N} + (1-\frac{1}{N})(1-\beta_2)2GL\mathbb{E}\|\theta_t - \theta_t^{(i)}\| \\
&\leq (1-\frac{1}{N})\beta_2\frac{\psi_t}{N} + (1-\frac{1}{N})(1-\beta_2)2GL\frac{\varphi_t}{N},
\end{aligned} \tag{43}$$

其中第二个不等式由假设4.2成立。

如果以概率 $\frac{1}{N}$ 发生 $j_{t+1} = i$，则有

$$\begin{aligned}
\mathbb{E}\|m_{t+1} - m_{t+1}^{(i)}\|_{\mathrm{F}} &= \frac{1}{N}\mathbb{E}\big\|\beta_1(m_t - m_t^{(i)}) + (1-\beta_1)(\nabla f(\theta_t; z_i) - \nabla f(\theta_t^{(i)}; \tilde{z}_i))\big\| \\
&\leq \frac{1}{N}\beta_1\mathbb{E}\|m_t - m_t^{(i)}\| + \frac{1}{N}(1-\beta_1)\mathbb{E}\|\nabla f(\theta_t; z_i) - \nabla f(\theta_t^{(i)}; \tilde{z}_i)\| \\
&\leq \frac{1}{N}\beta_1\frac{\phi_t}{N} + \frac{1-\beta_1}{N}\mathbb{E}\big\|\nabla f(\theta_t; z_i) - \nabla F(\theta_t) + \nabla F(\theta_t) - \nabla F(\theta_t^{(i)}) \\
&\quad + \nabla F(\theta_t^{(i)}) - \nabla f(\theta_t^{(i)}; \tilde{z}_i)\big\| \\
&\leq \frac{1}{N}\beta_1\frac{\phi_t}{N} + \frac{1-\beta_1}{N}\big(\mathbb{E}\|\nabla f(\theta_t; z_i) - \nabla F(\theta_t)\| + \mathbb{E}\|\nabla F(\theta_t) - \nabla F(\theta_t^{(i)})\| \\
&\quad + \mathbb{E}\|\nabla F(\theta_t^{(i)}) - \nabla f(\theta_t^{(i)}; \tilde{z}_i)\|\big) \\
&\leq \frac{1}{N}\beta_1\frac{\phi_t}{N} + \frac{2(1-\beta_1)\sigma}{N} + \frac{(1-\beta_1)L}{N}\mathbb{E}\|\theta_t - \theta_t^{(i)}\| \\
&\leq \frac{1}{N}\beta_1\frac{\phi_t}{N} + \frac{2(1-\beta_1)\sigma}{N} + \frac{(1-\beta_1)L}{N}\frac{\varphi_t}{N}.
\end{aligned} \tag{44}$$

我们还可以得到

$$\begin{aligned}
\mathbb{E}\|v_{t+1}-v_{t+1}^{(i)}\| &= \frac{1}{N}\mathbb{E}\big\|\beta_2(v_t-v_t^{(i)})+(1-\beta_2)((\nabla f(\theta_t;z_i))^2-(\nabla f(\theta_t^{(i)};\tilde{z}_i))^2\big\| \\
&\le \frac{1}{N}\big(\beta_2\mathbb{E}\|v_t-v_t^{(i)}\|+(1-\beta_2)\mathbb{E}\|(\nabla f(\theta_t;z_i))^2-\nabla f(\theta_t;z_i)\nabla f(\theta_t^{(i)};\tilde{z}_i) \\
&\quad +\nabla f(\theta_t;z_i)\nabla f(\theta_t^{(i)};\tilde{z}_i)-(\nabla f(\theta_t^{(i)};\tilde{z}_i))^2\|\big) \\
&\le \frac{1}{N}\beta_2\frac{\psi_t}{N}+\frac{1}{N}(1-\beta_2)2G\mathbb{E}\|\nabla f(\theta_t;z_i)-\nabla f(\theta_t^{(i)};\tilde{z}_i)\| \\
&\le \frac{1}{N}\beta_2\frac{\psi_t}{N}+\frac{1}{N}(1-\beta_2)2G\mathbb{E}\|\nabla f(\theta_t;z_i)-\nabla F(\theta_t)+\nabla F(\theta_t)-\nabla F(\theta_t^{(i)}) \\
&\quad +\nabla F(\theta_t^{(i)})-\nabla f(\theta_t^{(i)};\tilde{z}_i)\| \\
&\le \frac{1}{N}\beta_2\frac{\psi_t}{N}+\frac{1}{N}(1-\beta_2)2G\big(2\sigma+L\mathbb{E}\|\theta_t-\theta_t^{(i)}\|\big) \\
&\le \frac{1}{N}\beta_2\frac{\psi_t}{N}+\frac{1}{N}(1-\beta_2)4G\sigma+\frac{1}{N}(1-\beta_2)2GL\frac{\varphi_t}{N}. \qquad (45)
\end{aligned}$$

令 $\phi_{t+1}=\beta_1\phi_t+2(1-\beta_1)\sigma+(1-\beta_1)L\varphi_t$，利用上述不等式（42）和（44），可得

$$\begin{aligned}
\mathbb{E}\|m_{t+1}-m_{t+1}^{(i)}\| &\le (1-\frac{1}{N})\frac{\beta_1\phi_t}{N}+(1-\frac{1}{N})(1-\beta_1)L\frac{\varphi_t}{N} \\
&\quad +\frac{1}{N}\frac{\beta_1\phi_t}{N}+\frac{2(1-\beta_1)\sigma}{N}+\frac{(1-\beta_1)L}{N}\frac{\varphi_t}{N} \\
&= \frac{\beta_1\phi_t}{N}+\frac{2(1-\beta_1)\sigma}{N}+\frac{(1-\beta_1)L\varphi_t}{N} \\
&= \frac{\phi_{t+1}}{N}. \qquad (46)
\end{aligned}$$

令 $\psi_{t+1}=\beta_2\psi_t+4(1-\beta_2)G\sigma+2(1-\beta_2)GL\varphi_t$，利用上述不等式（43）和（45），可得

$$\begin{aligned}
\mathbb{E}\|v_{t+1}-v_{t+1}^{(i)}\| &\le (1-\frac{1}{N})\beta_2\frac{\psi_t}{N}+(1-\frac{1}{N})(1-\beta_2)2GL\frac{\varphi_t}{N} \\
&\quad +\frac{1}{N}\beta_2\frac{\psi_t}{N}+\frac{1}{N}(1-\beta_2)4G\sigma+\frac{1}{N}(1-\beta_2)2GL\frac{\varphi_t}{N} \\
&= \frac{\beta_2\psi_t}{N}+\frac{4(1-\beta_2)G\sigma}{N}+\frac{2(1-\beta_2)GL\varphi_t}{N} \\
&= \frac{\psi_{t+1}}{N}. \qquad (47)
\end{aligned}$$

根据上述不等式（26），因此可得

$$\begin{aligned}
\mathbb{E}\|\theta_{t+1}-\theta_{t+1}^{(i)}\| &\le (1-\eta\lambda)\mathbb{E}\|\theta_t-\theta_t^{(i)}\|+\frac{\eta\sqrt{d}}{(1-\beta_1^{t+1})(\rho+\varepsilon)}\mathbb{E}\big\|m_{t+1}-m_{t+1}^{(i)}\big\| \\
&\quad +\frac{\eta G\sqrt{d}}{(1-\beta_1^{t+1})(1-\beta_2^{t+1})(\rho+\varepsilon)^2}\mathbb{E}\big\|v_{t+1}-v_{t+1}^{(i)}\big\| \\
&\le (1-\eta\lambda)\frac{\varphi_t}{N}+\frac{\eta\sqrt{d}}{(1-\beta_1^{t+1})(\rho+\varepsilon)}\frac{\phi_{t+1}}{N}+\frac{\eta G\sqrt{d}}{(1-\beta_1^{t+1})(1-\beta_2^{t+1})(\rho+\varepsilon)^2}\frac{\psi_{t+1}}{N} \\
&= \frac{\varphi_{t+1}}{N}, \qquad (48)
\end{aligned}$$

其中 $\varphi_{t+1}=(1-\eta\lambda)\varphi_t+\frac{\eta\sqrt{d}\phi_{t+1}}{(1-\beta_1^{t+1})(\rho+\varepsilon)}+\frac{\eta G\sqrt{d}\psi_{t+1}}{(1-\beta_1^{t+1})(1-\beta_2^{t+1})(\rho+\varepsilon)^2}$。

令 $\eta=O(\frac{1}{\sqrt{d}})$，$\lambda\in[0,\frac{1}{\eta})$，$G=O(1)$，$L=O(1)$，$\beta_1=O(1)$ 和 $\beta_2=O(1)$，其中 $\beta_1,\ \beta_2\in(0,\ 1)$，因为 $\varphi_t=$

$O(\frac{1}{(\rho+\varepsilon)^{2(t-1)}})$、 $\phi_t = O(\frac{1}{(\rho+\varepsilon)^{2(t-1)}})$ 和 $\psi_t = O(\frac{1}{(\rho+\varepsilon)^{2t}})$，我们有

$$\begin{aligned}
\phi_{t+1} &= \beta_1\phi_t + 2(1-\beta_1)\sigma + (1-\beta_1)L\varphi_t = O(\frac{1}{(\rho+\varepsilon)^{2t}}),\\
\psi_{t+1} &= \beta_2\psi_t + 4(1-\beta_2)G\sigma + 2(1-\beta_2)GL\varphi_t = O(\frac{1}{(\rho+\varepsilon)^{2t}}),\\
\varphi_{t+1} &= (1-\eta\lambda)\varphi_t + \frac{\eta\sqrt{d}\phi_{t+1}}{(1-\beta_1^{t+1})(\rho+\varepsilon)} + \frac{\eta G\sqrt{d}\psi_{t+1}}{(1-\beta_1^{t+1})(1-\beta_2^{t+1})(\rho+\varepsilon)^2} = O(\frac{1}{(\rho+\varepsilon)^{2(t+1)}}).
\end{aligned} \tag{49}$$

因此，我们有

$$\mathbb{E}\|\theta_{t+1} - \theta_{t+1}^{(i)}\| \le \frac{\varphi_{t+1}}{N} = O(\frac{1}{(\rho+\varepsilon)^{2(t+1)}N}). \tag{50}$$

通过数学归纳法，我们有

$$\mathbb{E}\|\theta_T - \theta_T^{(i)}\| = O\Big(\frac{1}{(\rho+\varepsilon)^{2T}N}\Big). \tag{51}$$

B利用假设 4.2，即满足 G-Lipschitz $f(\theta;\ z)$ 对任意 $z \in \mathcal{D}$，那么我们有

$$\mathbb{E}|f(\theta_T; z) - f(\theta_T^{(i)}; z)| \le G\mathbb{E}\|\theta_T - \theta_T^{(i)}\| = O\Big(\frac{1}{(\rho+\varepsilon)^{2T}N}\Big). \tag{52}$$

对 S、 $S^{(i)}$ 以及算法的随机性，对上述不等式（52）取期望，并根据上述引理2.2，我们可以得到

$$|\mathbb{E}[F(\theta_T) - F_S(\theta_T)]| \le O\Big(\frac{1}{(\rho+\varepsilon)^{2T}N}\Big). \tag{53}$$

□

B. 我们的 HomeAdam(W) 算法的泛化分析

定理 *B.1*。（定理 *4.7* 的重述）假设序列 $\{\theta_t\}_{t=1}^T$ 由算法 *2* 在数据集$S = \{z_1, z_2, \cdots, z_N\}$上生成。在假设 *4.1*、*4.2*、*4.3* 下，不失一般性，令 $\tau \ge 1, \lambda \in [0, \frac{1}{\eta})$, $\beta_1 = O(1)$ 满足 $\beta_1 \in (0, 1)$, $\beta_2 = O(1)$ 满足 $\beta_2 \in (0, 1)$, $\sigma = O(1)$、$G = O(1)$ 和 $L = O(1)$。若迭代次数较小（即$T = O(1)$），则设为 $\eta = \frac{1}{\sqrt{d}}$；否则设为 $\eta = \frac{1}{\sqrt{dT}}$，我们有

$$|\mathbb{E}[F(\theta_T) - F_S(\theta_T)]| \le O(\frac{1}{N}). \tag{54}$$

证明。在数据集2 上执行算法 S 和 $S^{(i)}$，并使用相同的随机索引序列 $\{j_t\}_{t=1}^T$，并令 $\{\theta_t\}_{t=1}^T$和 $\{\theta_t^{(i)}\}_{t=1}^T$ 由算法 2 在 S 和 $S^{(i)}$下生成。

不失一般性，设 $\tau \ge 1$ 在算法 2 中。当 $\min_{1\le j\le d}(v_t)_j \ge \tau \ge 1$ 时，有 $\theta_t = \theta_{t-1} - \eta(\frac{m_t}{\sqrt{v_t}+\varepsilon} + \lambda\theta_{t-1})$ 和 $\theta_t^{(i)} = \theta_{t-1}^{(i)} - \eta(\frac{m_t^{(i)}}{\sqrt{\hat{v}_t^{(i)}}+\varepsilon} + \lambda\theta_{t-1}^{(i)})$。于是有

$$\theta_t - \theta_t^{(i)} = (1-\eta\lambda)(\theta_{t-1} - \theta_{t-1}^{(i)}) - \eta\Big(\frac{\hat{m}_t}{\hat{v}_t+\varepsilon} - \frac{\hat{m}_t^{(i)}}{\hat{v}_t^{(i)}+\varepsilon}\Big). \tag{55}$$

设 $\rho_t = \min_{1\le j\le d}(v_t)_j$，则有

$$\begin{aligned}
\|\theta_t - \theta_t^{(i)}\| &= \|(1-\eta\lambda)(\theta_{t-1}-\theta_{t-1}^{(i)}) - \eta\big(\frac{\hat m_t}{\hat v_t+\varepsilon} - \frac{\hat m_t^{(i)}}{\hat v_t^{(i)}+\varepsilon}\big)\| \\
&\le (1-\eta\lambda)\|\theta_{t-1}-\theta_{t-1}^{(i)}\| + \eta\|\frac{\hat m_t}{\hat v_t+\varepsilon} - \frac{\hat m_t^{(i)}}{\hat v_t^{(i)}+\varepsilon}\| \\
&\overset{(i)}{\le}(1-\eta\lambda)\|\theta_{t-1}-\theta_{t-1}^{(i)}\| + \frac{\eta\sqrt{d}}{(1-\beta_1^t)(\rho_t+\varepsilon)}\big\|m_t - m_t^{(i)}\big\| \\
&\quad + \frac{\eta G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)(\rho_t+\varepsilon)^2}\big\|v_t - v_t^{(i)}\big\| \\
&\le (1-\eta\lambda)\|\theta_{t-1}-\theta_{t-1}^{(i)}\| + \frac{\eta\sqrt{d}}{(1-\beta_1^t)(\tau+\varepsilon)}\big\|m_t - m_t^{(i)}\big\| \\
&\quad + \frac{\eta G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)(\tau+\varepsilon)^2}\big\|v_t - v_t^{(i)}\big\| \\
&\le (1-\eta\lambda)\|\theta_{t-1}-\theta_{t-1}^{(i)}\| + \frac{\eta\sqrt{d}}{(1-\beta_1^t)}\big\|m_t - m_t^{(i)}\big\| + \frac{\eta G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)}\big\|v_t - v_t^{(i)}\big\|, \qquad (56)
\end{aligned}$$

其中，上述不等式(i) 由引理4.4得出，最后一个不等式由 $\tau+\varepsilon \ge 1$成立。

当 $\min_{1\le j\le d}(v_t)_j < \tau$ 时，我们有 $\theta_t = \theta_{t-1} - \eta(\hat m_t + \lambda\theta_{t-1})$ 和 $\theta_t^{(i)} = \theta_{t-1}^{(i)} - \eta(\hat m_t^{(i)} + \lambda\theta_{t-1}^{(i)})$。于是有

$$\begin{aligned}
\|\theta_t - \theta_t^{(i)}\| &= \|(1-\eta\lambda)(\theta_{t-1}-\theta_{t-1}^{(i)}) - \eta\big(\hat m_t - \hat m_t^{(i)}\big)\| \\
&\le (1-\eta\lambda)\|\theta_{t-1}-\theta_{t-1}^{(i)}\| + \eta\|\hat m_t - \hat m_t^{(i)}\| \\
&\le (1-\eta\lambda)\|\theta_{t-1}-\theta_{t-1}^{(i)}\| + \frac{\eta\sqrt{d}}{(1-\beta_1^t)}\big\|m_t - m_t^{(i)}\big\| + \frac{\eta G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)}\big\|v_t - v_t^{(i)}\big\|. \qquad (57)
\end{aligned}$$

因此，我们有

$$\|\theta_t - \theta_t^{(i)}\| \le (1-\eta\lambda)\|\theta_{t-1}-\theta_{t-1}^{(i)}\| + \frac{\eta\sqrt{d}}{(1-\beta_1^t)}\big\|m_t - m_t^{(i)}\big\| + \frac{\eta G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)}\big\|v_t - v_t^{(i)}\big\|. \qquad (58)$$

根据上述不等式（58），按照定理4.5的上述证明，令 $\eta = O(\frac{1}{\sqrt{d}})$，$\beta_1 = O(1)$，$\beta_2 = O(1)$，$\sigma = O(1)$，$G = O(1)$和 $L = O(1)$，我们有

$$\begin{aligned}
&\phi_1 = 2(1-\beta_1)\sigma = O(1), \quad \psi_1 = 2(1-\beta_2)G^2 = O(1), \quad \varphi_1 = 2\eta\sqrt{d}\sigma + \frac{2\eta\sqrt{d}G^3}{(1-\beta_1)} = O(1), \\
&\mathbb{E}\|m_1 - m_1^{(i)}\| \le \frac{\phi_1}{N}, \quad \mathbb{E}\|v_1 - v_1^{(i)}\| \le \frac{\psi_1}{N}, \quad \mathbb{E}\|\theta_1 - \theta_1^{(i)}\| \le \frac{\varphi_1}{N} = O(\frac{1}{N}). \qquad (59)
\end{aligned}$$

当迭代次数较少时（即 $T = O(1)$），令 $\eta = \frac{1}{\sqrt{d}}$, $\lambda \in [0, \frac{1}{\eta})$, $\beta_1 = O(1)$，其中 $\beta_1 \in (0,1)$，$\beta_2 = O(1)$，且 $\beta_2 \in (0,1)$，$\sigma = O(1)$，$G = O(1)$ 和 $L = O(1)$。按照上述定理 4.5 的证明，假设 $\mathbb{E}\|m_t - m_t^{(i)}\| \le \frac{\phi_t}{N}$with $\phi_t = O(1)$，且 $\mathbb{E}\|v_t - v_t^{(i)}\| \le \frac{\psi_t}{N}$ with $\psi_t = O(1)$，我们可以得到 $\phi_{t+1} = O(1)$ 和 $\psi_{t+1} = O(1)$。利用上述不等式 (58)，可得

$$\mathbb{E}\|\theta_{t+1} - \theta_{t+1}^{(i)}\| \le \frac{\varphi_{t+1}}{N} = O(\frac{1}{N}), \qquad (60)$$

其中 $\varphi_{t+1} = (1-\eta\lambda)\varphi_t + \frac{\eta\sqrt{d}\phi_{t+1}}{(1-\beta_1^{t+1})} + \frac{\eta G\sqrt{d}\psi_{t+1}}{(1-\beta_1^{t+1})(1-\beta_2^{t+1})} = O(1)$。通过数学归纳法，我们得到

$$\mathbb{E}\|\theta_T - \theta_T^{(i)}\| \le \frac{\varphi_T}{N} = O(\frac{1}{N}). \qquad (61)$$

若迭代次数 T 较大，我们在泛化分析中考虑迭代次数 $t \geq 1$ 。根据上述不等式（58）的递归关系，采用定理4.5的上述证明方法，我们假设$\mathbb{E}\|m_t - m_t^{(i)}\| \leq \frac{\phi_t}{N}$满足 $\phi_t = O(t)$，且 $\mathbb{E}\|v_t - v_t^{(i)}\| \leq \frac{\psi_t}{N}$满足 $\psi_t = O(t)$。

令 $\eta = O(\frac{1}{\sqrt{d}})$，$\lambda \in [0, \frac{1}{\eta})$，$\beta_1 = O(1)$，其中 $\beta_1 \in (0,\ 1)$，$\beta_2 = O(1)$，其中 $\beta_2 \in (0,\ 1)$，$\sigma = O(1)$、$G = O(1)$ 且$L = O(1)$，则有

$$\phi_{t+1} = \beta_1\phi_t + 2(1-\beta_1)\sigma + (1-\beta_1)L\varphi_t = O(t+1),$$
$$\psi_{t+1} = \beta_2\psi_t + 4(1-\beta_2)G\sigma + 2(1-\beta_2)GL\varphi_t = O(t+1).$$

利用上述不等式（58），可得

$$\mathbb{E}\|\theta_{t+1} - \theta_{t+1}^{(i)}\| \leq \frac{\varphi_{t+1}}{N} = O(\frac{t+1}{N}). \tag{62}$$

利用数学归纳法，可得

$$\mathbb{E}\|\theta_T - \theta_T^{(i)}\| \leq \frac{\varphi_T}{N} = O(\frac{T}{N}). \tag{63}$$

进一步令 $\eta = O(\frac{1}{T\sqrt{d}})$，$\lambda \in [0, \frac{1}{\eta})$，$\beta_1 = O(1)$，其中 $\beta_1 \in (0, 1)$，$\beta_2 = O(1)$，其中 $\beta_2 \in (0, 1)$，$\sigma = O(1)$、$G = O(1)$ 且$L = O(1)$，则有 $\phi_T = O(1)$、$\psi_T = O(1)$、$\phi_T = O(1)$，

$$\mathbb{E}\|\theta_T - \theta_T^{(i)}\| \leq \frac{\varphi_T}{N} = O(\frac{1}{N}). \tag{64}$$

B根据假设 4.2，即满足 G-Lipschitz $f(\theta;\ z)$ 对任意 $z \in \mathcal{D}$，则有

$$\mathbb{E}|f(\theta_T; z) - f(\theta_T^{(i)}; z)| \leq G\mathbb{E}\|\theta_T - \theta_T^{(i)}\| = O(\frac{1}{N}). \tag{65}$$

对 S，$S^{(i)}$ 以及算法的随机性，对上述不等式（65）两边取期望，并根据上述引理2.2，可得

$$|\mathbb{E}[F(\theta_T) - F_S(\theta_T)]| \leq O(\frac{1}{N}). \tag{66}$$

□

C. Adam(W)-srf 算法的收敛分析

本节将对我们的 Adam(W)-srf 算法进行详细的收敛分析。

引理 C.1. （引理 5.3的重述）假设序列 $\{m_t\}_{t=0}^T$ 由算法 1生成，令 $\beta_1 = 1 - c\eta \in (0,\ 1)$，$\beta_2 \in (0,\ 1)$，则有

$$\mathbb{E}\|\nabla F(\theta_t) - m_{t+1}\|^2 \leq (1-c\eta)\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{2}{c\eta}L^2\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + c^2\eta^2\sigma^2, \tag{67}$$

其中 $c > 0$。

证明。 在算法 1 的第 6 行中，有 $m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t$ 对于 $t \geq 1$。由于 $m_{t+1} = \beta_1 m_t + (1-\beta_1)g_{t+1} =$

$\beta_1 m_t + (1-\beta_1)\nabla f(\theta_t;\, z_{t+1})$，则有

$$\begin{aligned}
\mathbb{E}\|\nabla F(\theta_t) - m_{t+1}\|^2 &= \mathbb{E}\|\nabla F(\theta_{t-1}) - m_t + \nabla F(\theta_t) - \nabla F(\theta_{t-1}) - (m_{t+1} - m_t)\|^2 \\
&= \mathbb{E}\|\nabla F(\theta_{t-1}) - m_t + \nabla F(\theta_t) - \nabla F(\theta_{t-1}) + (1-\beta_1)m_t - (1-\beta_1)\nabla f(\theta_t; z_{t+1})\|^2 \\
&= \mathbb{E}\|\beta_1(\nabla F(\theta_{t-1}) - m_t) + \beta_1\big(\nabla F(\theta_t) - \nabla F(\theta_{t-1})\big) + (1-\beta_1)(\nabla F(\theta_t) - \nabla f(\theta_t; z_{t+1}))\|^2 \\
&\overset{(i)}{=} \beta_1^2\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t + \nabla F(\theta_t) - \nabla F(\theta_{t-1})\|^2 + (1-\beta_1)^2\mathbb{E}\|\nabla F(\theta_t) - \nabla f(\theta_t; z_{t+1})\|^2 \\
&\overset{(ii)}{\le} \beta_1^2(2-\beta_1)\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \beta_1^2(1 + \frac{1}{1-\beta_1})\mathbb{E}\|\nabla F(\theta_t) - \nabla F(\theta_{t-1})\|^2 \\
&\quad + (1-\beta_1)^2\mathbb{E}\|\nabla F(\theta_t) - \nabla f(\theta_t; z_{t+1})\|^2 \\
&\overset{(iii)}{\le} \beta_1\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{2}{1-\beta_1}\mathbb{E}\|\nabla F(\theta_t) - \nabla F(\theta_{t-1})\|^2 \\
&\quad + (1-\beta_1)^2\mathbb{E}\|\nabla F(\theta_t) - \nabla f(\theta_t; z_{t+1})\|^2 \\
&\le \beta_1\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{2}{1-\beta_1}L^2\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + (1-\beta_1)^2\sigma^2,
\end{aligned} \tag{68}$$

其中，等式 (i) 根据 $\mathbb{E}[\nabla f(\theta_t;\, z_{t+1})] = \nabla F(\theta_t)$ 成立，不等式 (ii) 根据杨氏不等式成立，而不等式 (iii) 则归因于 $0 < \beta_1 < 1$，使得 $\beta_1^2(2-\beta_1) = (1-(1-\beta_1))^2(1+1-\beta_1) = 1-(1-\beta_1)-(1-\beta_1)^2+(1-\beta_1)^3 \le 1-(1-\beta_1)$，且 $\beta_1^2(1 + \frac{1}{1-\beta_1}) \le \frac{2}{1-\beta_1}$；最后一个不等式根据假设 5.1、4.3 成立。于是我们有

$$\begin{aligned}
&\mathbb{E}\|\nabla F(\theta_t) - m_{t+1}\|^2 - \mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 \\
&\le -(1-\beta_1)\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{2}{1-\beta_1}L^2\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + (1-\beta_1)^2\sigma^2.
\end{aligned} \tag{69}$$

设 $\beta_1 = 1 - c\eta$，我们可以得到

$$\begin{aligned}
&\mathbb{E}\|\nabla F(\theta_t) - m_{t+1}\|^2 - \mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 \\
&\le -c\eta\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{2}{c\eta}L^2\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + c^2\eta^2\sigma^2.
\end{aligned} \tag{70}$$

□

定理 C.2.（定理 *5.4* 的重述）假设序列 $\{\theta_t\}_{t=0}^T$ 由算法 *1* 生成。在假设 *5.1*、*4.2*、*4.3*、*5.2* 下，并令 $0 \le \lambda < \min(\frac{1}{\eta}, \frac{1}{\eta T^\gamma \bar{G}\hat{}G})$，$\|\theta_0\| \le \eta\bar{G}$, $c \ge \frac{16L}{\breve{\rho}}$, $\beta_1 = 1 - c\eta \in (0,\ 1)$，$\beta_2 \in (0,\ 1)$ 且 $0 < \eta \le \frac{\rho}{4L}$，则有

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| \le \hat{G}\Big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\breve{\rho}}} + \frac{4\sqrt{2}}{T^{\gamma-1}\breve{\rho}} + \frac{4c\sigma\sqrt{\eta}}{\sqrt{L\breve{\rho}}}\Big) + \frac{1}{T^{\gamma-1}}, \tag{71}$$

其中 $\bar{G} = \frac{G}{(1-\beta_1)(\rho+\varepsilon)}$，$\hat{G} = \frac{G^2+\varepsilon}{1-\beta_2}$，$\breve{\rho} = (1-\beta_1)(\rho + (1-\beta_2)\varepsilon)$ 以及 $\rho = \min_{t\ge 1}\min_{j\in[d]}(\hat{v}_t)_j$。

证明。 由算法 1可知， m_t 是 g_t的指数移动平均，根据假设 4.2，我们有 $\|m_t\| \le G$。

由于 $\theta_t = \theta_{t-1} - \eta(\frac{m_t}{\hat{v}_t+\varepsilon} + \lambda\theta_{t-1})$ 根据算法 1第 10 行，令 $\rho =$ 最小$_{t\geq 1}$最小$_{j\in[d]}(\hat{v}_t)_j$，我们有 e

$$
\begin{aligned}
\|\theta_t\| &= \|\theta_{t-1} - \eta(\frac{\hat{m}_t}{\hat{v}_t+\varepsilon} + \lambda\theta_{t-1})\| \\
&= \|(1-\eta\lambda)\theta_{t-1} - \eta\frac{\hat{m}_t}{\hat{v}_t+\varepsilon}\| \\
&\leq (1-\eta\lambda)\|\theta_{t-1}\| + \eta\|\frac{\hat{m}_t}{\hat{v}_t+\varepsilon}\| \\
&\leq (1-\eta\lambda)\|\theta_{t-1}\| + \frac{\eta G}{(1-\beta_1^t)(\rho+\varepsilon)} \\
&\leq (1-\eta\lambda)^t\|\theta_0\| + \frac{t\eta G}{(1-\beta_1)(\rho+\varepsilon)} \\
&\leq (t+1)\frac{\eta G}{(1-\beta_1)(\rho+\varepsilon)},
\end{aligned} \tag{72}
$$

其中，第一个不等式由 $0 \leq \lambda < \frac{1}{\eta}$得出，最后一个不等式根据 $\|\theta_0\| \leq \frac{\eta G}{(1-\beta_1)(\rho+\varepsilon)}$成立。令 $\bar{G} = \frac{G}{(1-\beta_1)(\rho+\varepsilon)}$，我们有 $\|\theta_t\| \leq (t+1)\eta G$ 对所有 $t \geq 1$均成立。

我们可以改写算法 1第 10 行，对所有 $j = 1, 2, \cdots, d$

$$
\begin{aligned}
(\theta_t)_j &= (\theta_{t-1})_j - \eta\big(\frac{(\hat{m}_t)_j}{(\hat{v}_t)_j+\varepsilon} - \lambda(\theta_{t-1})_j\big) \\
&= (1-\lambda\eta)(\theta_{t-1})_j - \eta\frac{\frac{1}{1-\beta_1^t}(m_t)_j}{\frac{(v_t)_j}{1-\beta_2^t}+\varepsilon},
\end{aligned} \tag{73}
$$

其中，$(\cdot)_j$ 表示向量的第 j个元素。根据假设 4.2，由于 v_t 是 g_t^2的指数移动平均，我们有 1 1 $(v_t)_j \leq G^2$对所有 $j = 1$、2、$\cdots$、d均成立。于是我们有 $\frac{1-\beta_2}{G^2+\varepsilon} \leq \frac{1-\beta_1^t}{v_t}(\cdot)_{\frac{j}{1-\beta_2^t}+\varepsilon} \leq \frac{1}{1-\beta_1}\frac{1}{\rho+}(\cdot)(\cdot)(\frac{1}{1-\beta_2})\varepsilon)$。令 $H_t = \mathrm{diag}(\frac{1-\beta_1^t}{v_t}(\cdot)_{\frac{j}{1-\beta_2^t}+\varepsilon})$为一个对角矩阵，我们有$\frac{1-\beta_2}{G^2+\varepsilon}I_d \preceq H_t \preceq \frac{1}{1-\beta_1}\frac{1}{\rho+}(\cdot)(\cdot)(\frac{1}{1-\beta_2})\varepsilon)\, I_d$。因此，我们也可以将算法 1 的第 10 行改写如下：

$$
\begin{aligned}
\theta_t &= \theta_{t-1} - \eta(H_t m_t + \lambda\theta_{t-1}) \\
&= (1-\eta\lambda)\theta_{t-1} - \eta H_t m_t \\
&= \arg\min_{\theta\in\mathbb{R}^d}\Big\{\langle m_t, \theta\rangle + \frac{1}{2\eta}\big(\theta - (1-\lambda\eta)\theta_{t-1}\big)^T H_t^{-1}\big(\theta - (1-\lambda\eta)\theta_{t-1}\big)\Big\}.
\end{aligned} \tag{74}
$$

利用子问题的最优性条件（74），可得

$$
\langle m_t + \frac{1}{\eta}H_t^{-1}\big(\theta_t - (1-\lambda\eta)\theta_{t-1}\big), \theta - \theta_t\rangle \geq 0, \quad \forall\theta\in\mathbb{R}^d. \tag{75}
$$

将 $\theta = \theta_{t-1}$ 代入上述不等式（75），可得

$$
\langle m_t + \frac{1}{\eta}H_t^{-1}\big(\theta_t - (1-\lambda\eta)\theta_{t-1}\big), \theta_{t-1} - \theta_t\rangle \geq 0. \tag{76}
$$

因此，我们可以得到

$$
\begin{aligned}
\langle m_t, \theta_{t-1} - \theta_t\rangle &\geq \frac{1}{\eta}\langle H_t^{-1}(\theta_t - \theta_{t-1}), \theta_t - \theta_{t-1}\rangle + \lambda\langle H_t^{-1}\theta_{t-1}, \theta_t - \theta_{t-1}\rangle \\
&\geq \frac{\hat{\rho}}{\eta}\|\theta_t - \theta_{t-1}\|^2 + \lambda\langle H_t^{-1}\theta_{t-1}, \theta_t - \theta_{t-1}\rangle,
\end{aligned} \tag{77}
$$

其中最后一个不等式由 $\breve{\rho}I_d \preceq H_t^{-1} \preceq \hat{G}I_d$ 结合 $\breve{\rho} = (1-\beta_1)(\rho + (1-\beta_2)\varepsilon)$ 以及 $\hat{G} = \frac{G^2+\varepsilon}{1-\beta_2}$成立。

由于 $\check{\rho}I_d \preceq H_t^{-1} \preceq \hat{G}I_d$ 与 $\|\theta_{t-1}\| \le t\eta\bar{G}$，我们有 $\|H_t^{-1}\theta_{t-1}\|^2 \le \hat{G}^2\|\theta_{t-1}\|^2 \le t^2\eta^2\bar{G}^2\hat{G}^2$ 对所有 $t \ge 1$。根据假设 5.1，即 $F(\theta)$ 是 L-光滑，我们有

$$
\begin{aligned}
\mathbb{E}[F(\theta_t)] &\le \mathbb{E}[F(\theta_{t-1}) + \nabla F(\theta_{t-1})^T(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2] \\
&= \mathbb{E}[F(\theta_{t-1}) + (\nabla F(\theta_{t-1}) - m_t)^T(\theta_t - \theta_{t-1}) + m_t^T(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2] \\
&\overset{(i)}{\le} \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{2\check{\rho}}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{\check{\rho}}{2\eta}\|\theta_t - \theta_{t-1}\|^2 + m_t^T(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2] \\
&\overset{(ii)}{\le} \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{2\check{\rho}}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{\check{\rho}}{2\eta}\|\theta_t - \theta_{t-1}\|^2 - \frac{\check{\rho}}{\eta_t}\|\theta_t - \theta_{t-1}\|^2 \\
&\quad - \lambda\langle H_t^{-1}\theta_{t-1}, \theta_t - \theta_{t-1}\rangle + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2] \\
&\le \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{2\check{\rho}}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{\check{\rho}}{2\eta}\|\theta_t - \theta_{t-1}\|^2 - \frac{\check{\rho}}{\eta}\|\theta_t - \theta_{t-1}\|^2 \\
&\quad + \frac{\lambda^2\eta}{\check{\rho}}\|H_t^{-1}\theta_{t-1}\|^2 + \frac{\check{\rho}}{4\eta}\|\theta_t - \theta_{t-1}\|^2 + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2] \\
&\le \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{\check{\rho}T^{2\gamma-2}} + \frac{\eta}{2\check{\rho}}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\check{\rho}}{8\eta}\|\theta_t - \theta_{t-1}\|^2], \qquad (78)
\end{aligned}
$$

其中，上述不等式（i）由杨氏不等式成立，上述不等式（77）由上述不等式得出，而最后一个不等式由 $0 < \eta \le \frac{\check{\rho}}{4L}$ 和 $0 \le \lambda \le \frac{1}{\eta T^\gamma \bar{G}\hat{G}}$ 成立。

我们定义一个有用的李雅普诺夫函数 $\Phi_t = F(\theta_t) + \frac{1}{2L}\|\nabla F(\theta_t) - m_{t+1}\|^2$。于是有

$$
\begin{aligned}
\mathbb{E}[\Phi_t - \Phi_{t-1}] &= \mathbb{E}[F(\theta_t)] - \mathbb{E}[F(\theta_{t-1})] + \frac{1}{2L}(\mathbb{E}\|\nabla F(\theta_t) - m_{t+1}\|^2 - \mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2) \\
&\le \frac{\eta}{\check{\rho}T^{2\gamma-2}} + \frac{\eta}{2\check{\rho}}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\check{\rho}}{8\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 \\
&\quad + \frac{1}{2L}\big(-c\eta\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{2}{c\eta}L^2\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + c^2\eta^2\sigma^2\big) \\
&\overset{(i)}{\le} \frac{\eta}{\check{\rho}T^{2\gamma-2}} - \frac{c\eta}{4L}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\check{\rho}}{16\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + \frac{c^2\eta^2\sigma^2}{2L} \\
&\le \frac{\eta}{\check{\rho}T^{2\gamma-2}} - \frac{4\eta}{\check{\rho}}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\check{\rho}}{16\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + \frac{c^2\eta^2\sigma^2}{2L}, \qquad (79)
\end{aligned}
$$

其中，第一个不等式由引理 C.1得出，而上述不等式 (i) 由 $c \ge \frac{16L}{\check{\rho}}$ 例如 $\frac{\check{\rho}}{16\eta} \ge \frac{L}{c\eta}$ 和 $\frac{c\eta}{4L} \ge \frac{c\eta}{32L} \ge \frac{\eta}{2\check{\rho}}$ 得出，最后一个不等式也由 $c \ge \frac{16L}{\check{\rho}}$ 得出。于是有

$$
\frac{4\eta}{\check{\rho}}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{\check{\rho}}{16\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 \le \mathbb{E}[\Phi_{t-1} - \Phi_t] + \frac{\eta}{\check{\rho}T^{2\gamma-2}} + \frac{c^2\eta^2\sigma^2}{2L}. \qquad (80)
$$

将上述不等式两边同时乘以（80），得到 $\frac{16}{\eta\check{\rho}}$

$$
\begin{aligned}
\frac{1}{\check{\rho}^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 &\le \frac{4}{\check{\rho}^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 \\
&\le \mathbb{E}[\frac{16(\Phi_{t-1} - \Phi_t)}{\eta\check{\rho}}] + \frac{16}{T^{2\gamma-2}\check{\rho}^2} + \frac{8c^2\eta\sigma^2}{L\check{\rho}}. \qquad (81)
\end{aligned}
$$

由于 $m_1 = \beta_1 m_0 + (1-\beta_1)g_1 = (1-\beta_1)\nabla f(\theta_0; z_1)$，我们有

$$
\begin{aligned}
\|\nabla F(\theta_0) - m_1\|^2 &= \|\nabla F(\theta_0) - (1-\beta_1)\nabla f(\theta_0; z_1)\|^2 \\
&= \|\nabla F(\theta_0) - \nabla f(\theta_0; z_1) + \beta_1\nabla f(\theta_0; z_1)\|^2 \le 2\sigma^2 + 2\beta_1^2 G^2. \qquad (82)
\end{aligned}
$$

给定 $\Phi_0 = F(\theta_0) + \frac{1}{2L}\|\nabla F(\theta_0) - m_1\|^2$，我们有

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}[\frac{1}{\breve{\rho}^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{16(\Phi_{t-1} - \Phi_t)}{\eta\breve{\rho}}] + \frac{16}{T^{2\gamma-2}\breve{\rho}^2} + \frac{8c^2\eta\sigma^2}{L\breve{\rho}} \\
&\leq \frac{16(F(\theta_0) + \frac{1}{2L}(2\sigma^2 + 2\beta_1^2 G^2) - F^*)}{T\eta\breve{\rho}} + \frac{16}{T^{2\gamma-2}\breve{\rho}^2} + \frac{8c^2\eta\sigma^2}{L\breve{\rho}}.
\end{aligned} \tag{83}$$

令 $\Delta = F(\theta_0) + \frac{1}{L}(\sigma^2 + \beta_1^2 G^2) - F^*$，我们可以将上述不等式（80）重写如下：

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}[\frac{1}{\breve{\rho}^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2] \\
&\leq \frac{16\Delta}{T\eta\breve{\rho}} + \frac{16}{T^{2\gamma-2}\breve{\rho}^2} + \frac{8c^2\eta\sigma^2}{L\breve{\rho}}.
\end{aligned} \tag{84}$$

根据詹森不等式，我们可以得到

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{1}{\breve{\rho}}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\theta_t - \theta_{t-1}\|] \\
&\leq \Big(\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{1}{\breve{\rho}^2}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\|\theta_t - \theta_{t-1}\|^2]\Big)^{1/2} \\
&\leq \sqrt{\frac{32\Delta}{T\eta\breve{\rho}} + \frac{32}{T^{2\gamma-2}\breve{\rho}^2} + \frac{16c^2\eta\sigma^2}{L\breve{\rho}}} \leq \frac{4\sqrt{2\Delta}}{\sqrt{T\eta\breve{\rho}}} + \frac{4\sqrt{2}}{T^{\gamma-1}\breve{\rho}} + \frac{4c\sigma\sqrt{\eta}}{\sqrt{L\breve{\rho}}}.
\end{aligned} \tag{85}$$

通过使用 $\theta_t = \theta_{t-1} - \eta(H_t m_t + \lambda\theta_{t-1})$，我们有

$$\begin{aligned}
\frac{1}{\breve{\rho}}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\theta_t - \theta_{t-1}\| &= \frac{1}{\breve{\rho}}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\eta(H_t m_t + \lambda\theta_{t-1})\| \\
&\geq \frac{1}{\breve{\rho}}\|\nabla F(\theta_{t-1}) - m_t\| + \|H_t m_t\| - \|\lambda\theta_{t-1}\| \\
&= \frac{1}{\breve{\rho}}\|H_t^{-1}H_t(\nabla F(\theta_{t-1}) - m_t)\| + \|H_t m_t\| - \|\lambda\theta_{t-1}\| \\
&\geq \|H_t(\nabla F(\theta_{t-1}) - m_t)\| + \|H_t m_t\| - \|\lambda\theta_{t-1}\| \\
&\geq \|H_t\nabla F(\theta_{t-1})\| - \|\lambda\theta_{t-1}\| \\
&\geq \frac{\|\nabla F(\theta_{t-1})\|}{\hat{G}} - \|\lambda\theta_{t-1}\|,
\end{aligned} \tag{86}$$

其中，上述不等式由 $\breve{\rho}I_d \preceq H_t^{-1} \preceq \hat{G}I_d$ 和 $\frac{1}{G}I_d \preceq H_t \preceq \frac{1}{\breve{\rho}}I_d$ 得出。

将上述不等式（86）代入（85），可得

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\| &\le \frac{\hat{G}}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{1}{\breve{\rho}}\|\nabla F(\theta_{t-1})-m_t\|+\frac{1}{\eta}\|\theta_t-\theta_{t-1}\|+\|\lambda\theta_{t-1}\|]\\ &\le \hat{G}\big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\breve{\rho}}}+\frac{4\sqrt{2}}{T^{\gamma-1}\breve{\rho}}+\frac{4c\sigma\sqrt{\eta}}{\sqrt{L\breve{\rho}}}\big)+\frac{\hat{G}}{T}\sum_{t=1}^{T}\mathbb{E}\|\lambda\theta_{t-1}\|\\ &\overset{(i)}{\le} \hat{G}\big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\breve{\rho}}}+\frac{4\sqrt{2}}{T^{\gamma-1}\breve{\rho}}+\frac{4c\sigma\sqrt{\eta}}{\sqrt{L\breve{\rho}}}\big)+\frac{\hat{G}}{T}\sum_{t=1}^{T}\lambda t\eta\bar{G}\\ &\overset{(ii)}{\le} \hat{G}\big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\breve{\rho}}}+\frac{4\sqrt{2}}{T^{\gamma-1}\breve{\rho}}+\frac{4c\sigma\sqrt{\eta}}{\sqrt{L\breve{\rho}}}\big)+\frac{\hat{G}}{T}\sum_{t=1}^{T}\frac{1}{\eta T^{\gamma}\bar{G}\hat{G}}t\eta\bar{G}\\ &\le \hat{G}\big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\breve{\rho}}}+\frac{4\sqrt{2}}{T^{\gamma-1}\breve{\rho}}+\frac{4c\sigma\sqrt{\eta}}{\sqrt{L\breve{\rho}}}\big)+\frac{1}{T^{\gamma-1}},\end{aligned}\tag{87}$$

其中，上述不等式（i）得益于 $\|\theta_{t-1}\|\le t\eta\bar{G}$，且上述不等式（$ii$）由 $0\le\lambda\le\frac{1}{\eta T^{\gamma}\bar{G}\hat{G}}$成立。因此我们有

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| &\le \hat{G}\big(\frac{4\sqrt{2\Delta}}{\sqrt{(T+1)\eta\breve{\rho}}}+\frac{4\sqrt{2}}{(T+1)^{\gamma-1}\breve{\rho}}+\frac{4c\sigma\sqrt{\eta}}{\sqrt{L\breve{\rho}}}\big)+\frac{1}{(T+1)^{\gamma-1}}\\ &\le \hat{G}\big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\breve{\rho}}}+\frac{4\sqrt{2}}{T^{\gamma-1}\breve{\rho}}+\frac{4c\sigma\sqrt{\eta}}{\sqrt{L\breve{\rho}}}\big)+\frac{1}{T^{\gamma-1}}.\end{aligned}\tag{88}$$

令 $\eta=\frac{1}{\sqrt{T}}$和 $\gamma=\frac{3}{4}$，可得

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\|\le\hat{G}\big(\frac{4\sqrt{2\Delta}}{\sqrt{\breve{\rho}}T^{1/4}}+\frac{4\sqrt{2}}{\breve{\rho}T^{1/4}}+\frac{4c\sigma}{\sqrt{L\breve{\rho}}T^{1/4}}\big)+\frac{1}{T^{1/4}}.\tag{89}$$

进一步令 $G=O(1)$、$L=O(1)$、$\sigma=O(1)$、$\beta_1=O(1)$，其中 $\beta_1\in(0,1)$，$\beta_2=O(1)$，其中 $\beta_2\in(0,1)$，以及 $c=O(1)$，则有 $\hat{G}=\frac{G^2+\varepsilon}{1-\beta_2}=O(1)$ 和 $\Delta=F(\theta_0)+\frac{1}{L}(\sigma^2+\beta_1^2G^2)-F^*=O(1)$。由于 $\breve{\rho}=(1-\beta_1)(\rho+(1-\beta_2)\varepsilon)\le\rho+\varepsilon$且 ε非常小，因此 $\breve{\rho}$非常小。于是我们有

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\|\le O(\frac{1}{\breve{\rho}T^{1/4}}).\tag{90}$$

□

D. HomeAdam(W) 算法的收敛分析

本节将对我们的 HomeAdam(W) 算法进行详细的收敛分析 .

引理 D.1. 假设 序列 $\{m_t\}_{t=0}^{T}$由算法 2生成，令 $\beta_1=1-c\eta\in(0,1)$，$\beta_2\in(0,1)$，我们有 e

$$\mathbb{E}\|\nabla F(\theta_t)-m_{t+1}\|^2\le(1-c\eta)\mathbb{E}\|\nabla F(\theta_{t-1})-m_t\|^2+\frac{2}{c\eta}L^2\mathbb{E}\|\theta_t-\theta_{t-1}\|^2+c^2\eta^2\sigma^2,\tag{91}$$

其中 $c>0$。

证明。 此证明可沿用引理 C.1的上述证明。

□

定理 D.2。（定理 5.6的重述）假设序列 $\{\theta_t\}_{t=0}^T$ 由算法 2生成。在假设 5.1、4.2、4.3、5.2下，并令 $0 \le \lambda < \min(\frac{1}{\eta}, \frac{1}{\eta T^{\gamma} \hat{G}^{\gamma} G})$、$\|\theta_0\| \le \eta \tilde{G}$、$c \ge \frac{32L}{\breve{\tau}}$、$\beta_1 = 1 - c\eta \in (0, 1)$、$\beta_2 \in (0, 1)$、$0 < \eta \le \frac{\tau}{4L}$ 和 $\tau > 0$，则有

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| \le \frac{8\sqrt{\Delta}\breve{G}}{\sqrt{T\eta\breve{\tau}}} + \frac{8\breve{G}}{\breve{\tau}T^{\gamma-1}} + \frac{4\sqrt{2\eta}c\sigma\breve{G}}{\sqrt{L\breve{\tau}}} + \frac{\breve{G}}{\hat{G}T^{\gamma-1}}, \tag{92}$$

其中 $\tilde{G} = \max(\frac{G}{(1-\beta_1)(\tau+(1-\beta_2)\varepsilon)}, \frac{G}{1-\beta_1})$, $\hat{G} = \frac{G^2+\varepsilon}{1-\beta_2}$, $\breve{G} = \max(1, \frac{G^2+\varepsilon}{1-\beta_2})$, $\hat{\tau} = (1-\beta_1)(\tau + (1-\beta_2)\varepsilon)$ 以及 $\breve{\tau} = \min(1-\beta_1, \hat{\tau})$。

证明。根据我们的算法 2，由于 m_t 是 g_t 的指数移动平均，利用假设 4.2，我们有 $\|m_t\| \le G$。当 $\min_{1\le j\le d}(v_t)_j \ge \tau > 0$ 时，我们有 $\theta_t = \theta_{t-1} - \eta(\frac{m_t}{\hat{v}_t+\varepsilon} + \lambda\theta_{t-1})$。令 $0 \le \lambda < \frac{1}{\eta}$，我们有

$$\begin{aligned}
\|\theta_t\| &= \|\theta_{t-1} - \eta(\frac{\hat{m}_t}{\hat{v}_t+\varepsilon} + \lambda\theta_{t-1})\| \\
&= \|(1-\eta\lambda)\theta_{t-1} - \eta\frac{\hat{m}_t}{\hat{v}_t+\varepsilon}\| \\
&\le (1-\eta\lambda)\|\theta_{t-1}\| + \eta\|\frac{\frac{m_t}{1-\beta_1^t}}{\frac{v_t}{1-\beta_2^t}+\varepsilon}\| \\
&\le (1-\eta\lambda)\|\theta_{t-1}\| + \frac{\eta G}{(1-\beta_1)(\tau+(1-\beta_2)\varepsilon)} \\
&\le (1-\eta\lambda)^t\|\theta_0\| + \frac{t\eta G}{(1-\beta_1)(\tau+(1-\beta_2)\varepsilon)} \\
&\le \frac{(t+1)\eta G}{(1-\beta_1)(\tau+(1-\beta_2)\varepsilon)},
\end{aligned} \tag{93}$$

其中，最后一个不等式由 $\|\theta_0\| \le \frac{\eta G}{(1-\beta_1)(\tau+(1-\beta_2)\varepsilon)}$ 得出。

当 $\min_{1\le j\le d}(v_t)_j < \tau$ 时，我们有 $\theta_t = \theta_{t-1} - \eta(\hat{m}_t + \lambda\theta_{t-1})$。令 $0 \le \lambda < \frac{1}{\eta}$，我们有

$$\begin{aligned}
\|\theta_t\| &= \|\theta_{t-1} - \eta(\hat{m}_t + \lambda\theta_{t-1})\| \\
&= \|(1-\eta\lambda)\theta_{t-1} - \eta\hat{m}_t\| \\
&\le (1-\eta\lambda)\|\theta_{t-1}\| + \eta\|\frac{m_t}{1-\beta_1^t}\| \\
&\le (1-\eta\lambda)\|\theta_{t-1}\| + \frac{\eta G}{1-\beta_1} \\
&\le (1-\eta\lambda)^t\|\theta_0\| + \frac{t\eta G}{1-\beta_1} \\
&\le \frac{(t+1)\eta G}{1-\beta_1},
\end{aligned} \tag{94}$$

其中，最后一个不等式由 $\|\theta_0\| \le \frac{\eta G}{1-\beta_1}$ 得出。

令 $\tilde{G} = \max(\frac{G}{(1-\beta_1)(\tau+(1-\beta_2)\varepsilon)}, \frac{G}{1-\beta_1})$，我们有 $\|\theta_t\| \le (t+1)\eta\tilde{G}$。

When 最小值$_{1\le j\le d}(v_t)_j \ge \tau > 0$，我们可以将算法 2 的第 11 行改写为，对于所有 $j = 1, 2, \cdots, d$

$$\begin{aligned}
(\theta_t)_j &= (\theta_{t-1})_j - \eta\big(\frac{(\hat{m}_t)_j}{(\hat{v}_t)_j+\varepsilon} - \lambda(\theta_{t-1})_j\big) \\
&= (1-\lambda\eta)(\theta_{t-1})_j - \eta\frac{\frac{1}{1-\beta_1^t}(m_t)_j}{\frac{1}{1-\beta_2^t}(v_t)_j+\varepsilon},
\end{aligned} \tag{95}$$

1其中 $(\cdot)_j$ 表示向量的第 j 个元素。于是我们有 $\frac{1-\beta_2}{G^2+\varepsilon} \le \frac{1-\beta_1^t}{\frac{1}{1-\beta_2^t}v_t+\varepsilon} \le \frac{1}{1-\beta_1}\frac{1}{\tau+((1-\beta_2)\varepsilon)_\circ}$ 令 $H_t = 1\mathrm{diag}(\frac{1-\beta_1^t}{\frac{1}{1-\beta_2^t}v_t+\varepsilon})$ 为一个对角矩阵，则有 $\frac{1-\beta_2}{G^2+\varepsilon}I_d \preceq H_t \preceq \frac{1}{1-\beta_1}\frac{1}{\tau+((1-\beta_2)\varepsilon)}I_{d\circ}$ 于是我们也可以将算法 2 的第 11 行改写如下：

$$\begin{aligned}\theta_t &= \theta_{t-1} - \eta(H_t m_t + \lambda\theta_{t-1})\\ &= (1-\eta\lambda)\theta_{t-1} - \eta H_t m_t\\ &= \arg\min_{\theta\in\mathbb{R}^d}\Big\{\langle m_t,\theta\rangle + \frac{1}{2\eta}\big(\theta-(1-\lambda\eta)\theta_{t-1}\big)^T H_t^{-1}\big(\theta-(1-\lambda\eta)\theta_{t-1}\big)\Big\}.\end{aligned} \tag{96}$$

利用子问题的最优性条件（96），我们得到

$$\langle m_t + \frac{1}{\eta}H_t^{-1}\big(\theta_t-(1-\lambda\eta)\theta_{t-1}\big), \theta-\theta_t\rangle \ge 0, \quad \forall\theta\in\mathbb{R}^d. \tag{97}$$

将 $\theta=\theta_{t-1}$ 代入上述不等式（97），我们得到

$$\langle m_t + \frac{1}{\eta}H_t^{-1}\big(\theta_t-(1-\lambda\eta)\theta_{t-1}\big), \theta_{t-1}-\theta_t\rangle \ge 0. \tag{98}$$

令 $\hat{}\tau = (1-\beta_1)(\tau+(1-\beta_2)\varepsilon)$，我们可以得到

$$\begin{aligned}\langle m_t, \theta_{t-1}-\theta_t\rangle &\ge \frac{1}{\eta}\langle H_t^{-1}(\theta_t-\theta_{t-1}), \theta_t-\theta_{t-1}\rangle + \lambda\langle H_t^{-1}\theta_{t-1}, \theta_t-\theta_{t-1}\rangle\\ &\ge \frac{\hat{\tau}}{\eta}\|\theta_t-\theta_{t-1}\|^2 + \lambda\langle H_t^{-1}\theta_{t-1}, \theta_t-\theta_{t-1}\rangle.\end{aligned} \tag{99}$$

由于 $\frac{1-\beta_2}{G^2+\varepsilon}I_d \preceq H_t \preceq \frac{1}{1-\beta_1}\frac{1}{\tau+((1-\beta_2)\varepsilon)}I_d$，令 $\hat{}\tau=(1-\beta_1)(\tau+(1-\beta_2)\varepsilon)$ 和 $\hat{}G=\frac{G^2+\varepsilon}{1-\beta_2}$，我们有 $\hat{}\tau I_d \preceq H_t^{-1} \preceq \hat{}GI_d$和 $\|H_t^{-1}\theta_{t-1}\|^2 \le \hat{}G^2\|\theta_{t-1}\|^2 \le t^2\eta^2\tilde{G}^2\,\hat{}G^2$。

根据假设 5.1，即 $F(\theta)$ 是 L-光滑的，我们有

$$\begin{aligned}\mathbb{E}[F(\theta_t)] &\le \mathbb{E}[F(\theta_{t-1}) + \nabla F(\theta_{t-1})^T(\theta_t-\theta_{t-1}) + \frac{L}{2}\|\theta_t-\theta_{t-1}\|^2]\\ &= \mathbb{E}[F(\theta_{t-1}) + (\nabla F(\theta_{t-1})-m_t)^T(\theta_t-\theta_{t-1}) + m_t^T(\theta_t-\theta_{t-1}) + \frac{L}{2}\|\theta_t-\theta_{t-1}\|^2]\\ &\le \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{2\hat{\tau}}\|\nabla F(\theta_{t-1})-m_t\|^2 + \frac{\hat{\tau}}{2\eta}\|\theta_t-\theta_{t-1}\|^2 + m_t^T(\theta_t-\theta_{t-1}) + \frac{L}{2}\|\theta_t-\theta_{t-1}\|^2]\\ &\overset{(i)}{\le} \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{2\hat{\tau}}\|\nabla F(\theta_{t-1})-m_t\|^2 + \frac{\hat{\tau}}{2\eta}\|\theta_t-\theta_{t-1}\|^2 - \frac{\hat{\tau}}{\eta_t}\|\theta_t-\theta_{t-1}\|^2\\ &\quad - \lambda\langle H_t^{-1}\theta_{t-1}, \theta_t-\theta_{t-1}\rangle + \frac{L}{2}\|\theta_t-\theta_{t-1}\|^2]\\ &\overset{(ii)}{\le} \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{2\hat{\tau}}\|\nabla F(\theta_{t-1})-m_t\|^2 + \frac{\hat{\tau}}{2\eta}\|\theta_t-\theta_{t-1}\|^2 - \frac{\hat{\tau}}{\eta}\|\theta_t-\theta_{t-1}\|^2\\ &\quad + \frac{\lambda^2\eta}{\hat{\tau}}\|H_t^{-1}\theta_{t-1}\|^2 + \frac{\hat{\tau}}{4\eta}\|\theta_t-\theta_{t-1}\|^2 + \frac{L}{2}\|\theta_t-\theta_{t-1}\|^2]\\ &\le \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{\hat{\tau}T^{2\gamma-2}} + \frac{\eta}{2\hat{\tau}}\|\nabla F(\theta_{t-1})-m_t\|^2 - \frac{\hat{\tau}}{8\eta}\|\theta_t-\theta_{t-1}\|^2],\end{aligned} \tag{100}$$

其中，不等式（i）由上述不等式（99）成立，最后一个不等式则由 $\eta \le \hat{}\frac{\tau}{4L}$ 和 $\lambda \le \frac{1}{\eta T^\gamma \tilde{G}\,\hat{}G}$ 得出。于是我们有

$$\mathbb{E}\|\theta_t-\theta_{t-1}\|^2 \le \frac{8\eta(F(\theta_{t-1})-F(\theta_t))}{\hat{\tau}} + \frac{4\eta^2}{\hat{\tau}^2}\|\nabla F(\theta_{t-1})-m_t\|^2 + \frac{8\eta^2}{\hat{\tau}^2T^{2\gamma-2}}. \tag{101}$$

当 $\min_{1\le j\le d}(v_t)_j < \tau$ 时，我们有

$$\theta_t = \theta_{t-1} - \eta(\hat{m}_t + \lambda\theta_{t-1}) = \theta_{t-1} - \eta(\frac{m_t}{1-\beta_1^t} + \lambda\theta_{t-1}). \tag{102}$$

设 $\tilde{H}_t = \frac{1}{1-\beta_1^t} I_d$ 为对角矩阵，且有 $I_d \preceq \tilde{H}_t \preceq \frac{1}{1-\beta_1} I_d$。根据上述证明，可得

$$\begin{aligned}\langle m_t, \theta_{t-1} - \theta_t\rangle &\ge \frac{1}{\eta}\langle \tilde{H}_t^{-1}(\theta_t - \theta_{t-1}), \theta_t - \theta_{t-1}\rangle + \lambda\langle \tilde{H}_t^{-1}\theta_{t-1}, \theta_t - \theta_{t-1}\rangle \\ &\ge \frac{1-\beta_1}{\eta}\|\theta_t - \theta_{t-1}\|^2 + \lambda\langle \tilde{H}_t^{-1}\theta_{t-1}, \theta_t - \theta_{t-1}\rangle. \end{aligned}\tag{103}$$

于是可得

$$\mathbb{E}[F(\theta_t)] \le \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{T^{2\gamma-2}} + \frac{\eta}{2(1-\beta_1)}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{1-\beta_1}{8\eta}\|\theta_t - \theta_{t-1}\|^2]. \tag{104}$$

因此，我们可以得到

$$\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 \le \frac{8\eta(F(\theta_{t-1}) - F(\theta_t))}{1-\beta_1} + \frac{8\eta^2}{(1-\beta_1)T^{2\gamma-2}} + \frac{4\eta^2}{(1-\beta_1)^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2. \tag{105}$$

设 $\breve{\tau} = \min(1-\beta_1, \hat{\tau})$，其中 $\hat{\tau} = (1-\beta_1)(\tau + (1-\beta_2)\varepsilon)$，且 $\breve{G} = \max(1, \frac{G^2+\varepsilon}{1-\beta_2})$；进一步设 $\frac{1}{\breve{G}} I_d = \min(1, \frac{1-\beta_2}{G^2+\varepsilon}) I_d \preceq \hat{H}_t \preceq \max(\frac{1}{1-\beta_1}, \frac{1}{(1-\beta_1)(\tau+(1-\beta_2)\varepsilon)}) I_d = \frac{1}{\breve{\tau}} I_d$。

根据上述不等式（101）和（105），我们有

$$\begin{aligned}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 &\le \frac{8\eta(F(\theta_{t-1}) - F(\theta_t))}{\hat{\tau}} + \frac{8\eta(F(\theta_{t-1}) - F(\theta_t))}{1-\beta_1} + \frac{4\eta^2}{\hat{\tau}^2}\|\nabla F(\theta_{t-1}) - m_t\|^2 \\ &\quad + \frac{4\eta^2}{(1-\beta_1)^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{8\eta^2}{\hat{\tau}^2 T^{2\gamma-2}} + \frac{8\eta^2}{(1-\beta_1)T^{2\gamma-2}} \\ &\le \frac{16\eta(F(\theta_{t-1}) - F(\theta_t))}{\breve{\tau}} + \frac{8\eta^2}{\breve{\tau}^2}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{16\eta^2}{\breve{\tau}^2 T^{2\gamma-2}}, \end{aligned}\tag{106}$$

其中，最后一个不等式是由 $\breve{\tau} = \min(1-\beta_1, \hat{\tau})$ 得出的，其中 $\hat{\tau} = (1-\beta_1)(\tau + (1-\beta_2)\varepsilon)$。

于是我们可以得到

$$F(\theta_{t-1}) \le F(\theta_t)) - \frac{\breve{\tau}}{16\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + \frac{\eta}{2\breve{\tau}}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{\eta}{\breve{\tau} T^{2\gamma-2}}. \tag{107}$$

我们定义一个有用的李雅普诺夫函数 $\Omega_t = F(\theta_t) + \frac{1}{2L}\|\nabla F(\theta_t) - m_{t+1}\|^2$。根据上面的不等式（107），我们有

$$\begin{aligned}\mathbb{E}[\Omega_t - \Omega_{t-1}] &= \mathbb{E}[F(\theta_t) - F(\theta_{t-1})] + \frac{1}{2L}(\mathbb{E}\|\nabla F(\theta_t) - m_{t+1}\|^2 - \mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2) \\ &\le \frac{\eta}{\breve{\tau} T^{2\gamma-2}} + \frac{\eta}{2\breve{\tau}}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\breve{\tau}}{16\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 \\ &\quad + \frac{1}{2L}\big(-c\eta\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{2}{c\eta}L^2\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + c^2\eta^2\sigma^2\big) \\ &\overset{(i)}{\le} \frac{\eta}{\breve{\tau} T^{2\gamma-2}} - \frac{c\eta}{4L}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\breve{\tau}}{32\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + \frac{c^2\eta^2\sigma^2}{2L} \\ &\le \frac{\eta}{\breve{\tau} T^{2\gamma-2}} - \frac{8\eta}{\breve{\tau}}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\breve{\tau}}{32\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + \frac{c^2\eta^2\sigma^2}{2L}, \end{aligned}\tag{108}$$

其中，第一个不等式由引理 D.1 得出，上述不等式 (i) 由 $c \geq \frac{32L}{\check{\tau}}$（例如 $\frac{\check{\tau}}{32\eta} \geq \frac{L}{c\eta}$ 和 $\frac{c\eta}{4L} \geq \frac{c\eta}{64L} \geq \frac{\eta}{2\check{\tau}}$）得出，最后一个不等式也归因于 $c \geq \frac{32L}{\check{\tau}}$。于是我们有

$$\frac{8\eta}{\check{\tau}}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{\check{\tau}}{32\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 \leq \mathbb{E}[\Omega_{t-1} - \Omega_t] + \frac{\eta}{\check{\tau}T^{2\gamma-2}} + \frac{c^2\eta^2\sigma^2}{2L}. \tag{109}$$

将不等式（109）两边乘以 $\frac{32}{\eta\check{\tau}}$，可得

$$\begin{aligned}\frac{1}{\check{\tau}^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 &\leq \frac{4}{\check{\tau}^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 \\ &\leq \mathbb{E}[\frac{32(\Omega_{t-1} - \Omega_t)}{\eta\check{\tau}}] + \frac{32}{\check{\tau}^2T^{2\gamma-2}} + \frac{16c^2\eta\sigma^2}{L\check{\tau}}.\end{aligned} \tag{110}$$

由于 $m_1 = \beta_1 m_0 + (1-\beta_1)g_1 = (1-\beta_1)\nabla f(\theta_0; z_1)$，因此有

$$\begin{aligned}\|\nabla F(\theta_0) - m_1\|^2 &= \|\nabla F(\theta_0) - (1-\beta_1)\nabla f(\theta_0; z_1)\|^2 \\ &= \|\nabla F(\theta_0) - \nabla f(\theta_0; z_1) + \beta_1\nabla f(\theta_0; z_1)\|^2 \leq 2\sigma^2 + 2\beta_1^2G^2.\end{aligned} \tag{111}$$

由于 $\Omega_0 = F(\theta_0) + \frac{1}{2L}\|\nabla F(\theta_0) - m_1\|^2$，因此有

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}[\frac{1}{\check{\tau}^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2] \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{32(\Omega_{t-1} - \Omega_t)}{\eta\check{\tau}}] + \frac{32}{\check{\tau}^2T^{2\gamma-2}} + \frac{16c^2\eta\sigma^2}{L\check{\tau}} \\ &\leq \frac{32(F(\theta_0) + \frac{1}{2L}(2\sigma^2 + 2\beta_1^2G^2) - F^*)}{T\eta\check{\tau}} + \frac{32}{\check{\tau}^2T^{2\gamma-2}} + \frac{16c^2\eta\sigma^2}{L\check{\tau}}.\end{aligned} \tag{112}$$

设 $\Delta = F(\theta_0) + \frac{1}{L}(\sigma^2 + \beta_1^2G^2) - F^*$，则有

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}[\frac{1}{\check{\tau}^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2] \\ &\leq \frac{32\Delta}{T\eta\check{\tau}} + \frac{32}{\check{\tau}^2T^{2\gamma-2}} + \frac{16c^2\sigma^2\eta}{L\check{\tau}}.\end{aligned} \tag{113}$$

根据詹森不等式，我们可以得到

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{1}{\check{\tau}}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\theta_t - \theta_{t-1}\|] \\ &\leq \Big(\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{1}{\check{\tau}^2}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\|\theta_t - \theta_{t-1}\|^2]\Big)^{1/2} \\ &\leq \sqrt{\frac{64\Delta}{T\eta\check{\tau}} + \frac{64}{\check{\tau}^2T^{2\gamma-2}} + \frac{32c^2\eta\sigma^2}{L\check{\tau}}} \leq \frac{8\sqrt{\Delta}}{\sqrt{T\eta\check{\tau}}} + \frac{8}{\check{\tau}T^{\gamma-1}} + \frac{4\sqrt{2\eta}c\sigma}{\sqrt{L\check{\tau}}}.\end{aligned} \tag{114}$$

使用 $\theta_t = \theta_{t-1} - \eta(\hat{H}_t m_t + \lambda\theta_{t-1})$，我们有

$$\begin{aligned}\frac{1}{\breve{\tau}}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\theta_t - \theta_{t-1}\| &= \frac{1}{\breve{\tau}}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\eta(\hat{H}_t m_t + \lambda\theta_{t-1})\| \\ &\geq \frac{1}{\breve{\tau}}\|\nabla F(\theta_{t-1}) - m_t\| + \|\hat{H}_t m_t\| - \|\lambda\theta_{t-1}\| \\ &= \frac{1}{\breve{\tau}}\|\hat{H}_t^{-1}\hat{H}_t(\nabla F(\theta_{t-1}) - m_t)\| + \|\hat{H}_t m_t\| - \|\lambda\theta_{t-1}\| \\ &\geq \|\hat{H}_t(\nabla F(\theta_{t-1}) - m_t)\| + \|\hat{H}_t m_t\| - \|\lambda\theta_{t-1}\| \\ &\geq \|\hat{H}_t\nabla F(\theta_{t-1})\| - \|\lambda\theta_{t-1}\| \\ &\geq \frac{\|\nabla F(\theta_{t-1})\|}{\breve{G}} - \|\lambda\theta_{t-1}\|, \end{aligned} \tag{115}$$

其中，上述不等式由 $\breve{\tau} I_d \preceq \hat{H}_t^{-1} \preceq \breve{G} I_d$和 $\frac{1}{\breve{G}} I_d \preceq \hat{H}_t \preceq \frac{1}{\breve{\tau}} I_d$ 成立。

将上述不等式（115）代入（114），可得

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\| &\leq \frac{\breve{G}}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{1}{\breve{\tau}}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\theta_t - \theta_{t-1}\| + \|\lambda\theta_{t-1}\|] \\ &\leq \breve{G}\big(\frac{8\sqrt{\Delta}}{\sqrt{T\eta\breve{\tau}}} + \frac{8}{\breve{\tau}T^{\gamma-1}} + \frac{4\sqrt{2\eta}c\sigma}{\sqrt{L\breve{\tau}}}\big) + \frac{\breve{G}}{T}\sum_{t=1}^{T}\mathbb{E}\|\lambda\theta_{t-1}\| \\ &\overset{(i)}{\leq} \breve{G}\big(\frac{8\sqrt{\Delta}}{\sqrt{T\eta\breve{\tau}}} + \frac{8}{\breve{\tau}T^{\gamma-1}} + \frac{4\sqrt{2\eta}c\sigma}{\sqrt{L\breve{\tau}}}\big) + \frac{\breve{G}}{T}\sum_{t=1}^{T}\lambda t\eta\tilde{G} \\ &\overset{(ii)}{\leq} \breve{G}\big(\frac{8\sqrt{\Delta}}{\sqrt{T\eta\breve{\tau}}} + \frac{8}{\breve{\tau}T^{\gamma-1}} + \frac{4\sqrt{2\eta}c\sigma}{\sqrt{L\breve{\tau}}}\big) + \frac{\breve{G}}{T}\sum_{t=1}^{T}\frac{1}{\eta T^{\gamma}\tilde{G}\hat{G}}t\eta\tilde{G} \\ &\leq \breve{G}\big(\frac{8\sqrt{\Delta}}{\sqrt{T\eta\breve{\tau}}} + \frac{8}{\breve{\tau}T^{\gamma-1}} + \frac{4\sqrt{2\eta}c\sigma}{\sqrt{L\breve{\tau}}}\big) + \frac{\breve{G}}{\hat{G}T^{\gamma-1}}, \end{aligned} \tag{116}$$

其中，上述不等式（i）由 $\|\theta_t\| \leq (t+1)\eta\tilde{G}$ 成立，而上述不等式（ii）则是因为 $\lambda \leq \frac{1}{\eta T^{\gamma}\tilde{G}\hat{G}}$。因此，我们有

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| &\leq \breve{G}\big(\frac{8\sqrt{\Delta}}{\sqrt{T+1\eta\breve{\tau}}} + \frac{8}{\breve{\tau}(T+1)^{\gamma-1}} + \frac{4\sqrt{2\eta}c\sigma}{\sqrt{L\breve{\tau}}}\big) + \frac{\breve{G}}{\hat{G}(T+1)^{\gamma-1}} \\ &\leq \frac{8\sqrt{\Delta}\breve{G}}{\sqrt{T\eta\breve{\tau}}} + \frac{8\breve{G}}{\breve{\tau}T^{\gamma-1}} + \frac{4\sqrt{2\eta}c\sigma\breve{G}}{\sqrt{L\breve{\tau}}} + \frac{\breve{G}}{\hat{G}T^{\gamma-1}}. \end{aligned} \tag{117}$$

令 $\eta = \frac{1}{\sqrt{T}}$和 $\gamma = \frac{3}{4}$，可得

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| \leq \breve{G}\big(\frac{8\sqrt{\Delta}}{\sqrt{\breve{\tau}}T^{1/4}} + \frac{8}{\breve{\tau}T^{1/4}} + \frac{4\sqrt{2}c\sigma}{\sqrt{L\breve{\tau}}T^{1/4}} + \frac{1}{\hat{G}T^{1/4}}\big). \tag{118}$$

由于 $\eta = \frac{1}{\sqrt{T}}$并设 $c = O(1)$，我们有 $\beta_1 = 1 - c\eta = O(1)$。设 $\tau = O(1)$，我们有 $\breve{\tau} = \min(1-\beta_1, \hat{\tau}) = \min(1-\beta_1, (1-\beta_1)(\tau + (1-\beta_2)\varepsilon)) = O(1)$。进一步令 $G = O(1)$ 和 $L = O(1)$，我们可以得到

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| \leq O(\frac{1}{T^{1/4}}). \tag{119}$$

□

E. 详细实验设置

E.1. 计算机视觉任务

在实验中，我们将所有算法的小批量大小设为 64。

在 CIFAR-10 数据集上训练 VGG16 时，我们为随机梯度下降和带动量的随机梯度下降设置学习率 10^{-4}，并为带动量的随机梯度下降设置动量参数 $\beta = 0.9$。Adam、AdamW、AdaBelief 和 MiAdam 使用基础学习率 10^{-6}、调节参数 $\varepsilon = 10^{-8}$、一阶动量参数 $\beta_1 = 0.9$ 和二阶动量参数 $\beta_2 = 0.99$。同时，AdamW 使用权重衰减参数 $\lambda = 10^{-5}$，MiAdam 使用多重积分率 $\kappa = 0.9$。SWATS 使用基础学习率 10^{-5}、调节参数 $\varepsilon = 10^{-8}$、$\beta_1 = 0.9$ 和 $\beta_2 = 0.99$。我们的 Adam(W)-srf 和 HomeAdam(W) 使用基础学习率 10^{-6}、$\varepsilon = 10^{-7}$、$\beta_1 = 0.9$ 和 $\beta_2 = 0.99$。同时，我们的 AdamW-srf 和 HomeAdamW 使用权重衰减参数 $\lambda = 10^{-5}$。

在 Tiny-ImageNet 数据集上训练 ResNet34 时，我们为随机梯度下降和带动量的随机梯度下降设置学习率 4×10^{-4}，并为带动量的随机梯度下降设置 $\beta = 0.9$。Adam、AdamW、AdaBelief 和 MiAdam 使用的基础学习率分别为 10^{-6}、$\varepsilon = 10^{-8}$、$\beta_1 = 0.9$ 和 $\beta_2 = 0.99$。同时，AdamW 使用权重衰减参数 $\lambda = 10^{-4}$，MiAdam 使用多重积分率 $\kappa = 0.999$。SWATS 使用的基础学习率为 10^{-5}、$\varepsilon = 10^{-8}$、$\beta_1 = 0.9$ 和 $\beta_2 = 0.99$。我们的 Adam(W)-srf 和 HomeAdam(W) 使用的基础学习率分别为 10^{-6}、$\varepsilon = 10^{-7}$、$\beta_1 = 0.9$ 和 $\beta_2 = 0.99$。同时，我们的 AdamW-srf 和 HomeAdamW 使用权重衰减参数 $\lambda = 10^{-4}$。

E.2. 自然语言处理任务

第一个语言模型采用 8 层 Transformer（Vaswani 等人，2017年）编码器，嵌入维度为 768，每层有 8 个注意力头，前馈网络维度为 1024，并使用正弦位置编码。同时，整个网络的丢弃率为 0.1。最终输出层将表示映射回词汇表大小，用于词元预测。

第二个语言模型采用 24 层 Transformer 编码器，嵌入维度为 768，每层有 8 个注意力头，前馈网络维度为 2048，并使用正弦位置编码。同时，整个网络的丢弃率为 0.15。最终输出层将表示映射回词汇表大小，用于词元预测。

在 WikiText2 数据集上训练 8 层 Transformer 模型时，我们将所有算法的小批量大小设为 32。随机梯度下降和带动量的随机梯度下降的学习率设为 2×10^{-5}，带动量的随机梯度下降的动量参数设为 $\beta = 0.9$。Adam、AdamW 和 AdaBelief 使用基本学习率 10^{-6}、调节参数 $\varepsilon = 10^{-8}$、一阶动量参数 $\beta_1 = 0.9$ 和二阶动量参数 $\beta_2 = 0.999$。同时，AdamW 使用权重衰减参数 $\lambda = 10^{-4}$，MiAdam 使用基本学习率 10^{-6}、$\varepsilon = 10^{-8}$、$\beta_1 = 0.9$、$\beta_2 = 0.99$ 和多重积分率 $\kappa = 0.999$。SWATS 使用基本学习率 10^{-5}、调节参数 $\varepsilon = 10^{-8}$、$\beta_1 = 0.9$ 和 $\beta_2 = 0.99$。我们的 Adam(W)-srf 和 HomeAdam(W) 使用基本学习率 10^{-6}、$\varepsilon = 10^{-5}$、$\beta_1 = 0.9$ 和 $\beta_2 = 0.999$。同时，我们的 AdamW-srf 和 HomeAdamW 使用权重衰减参数 $\lambda = 10^{-4}$。

在 WikiText103 数据集上训练 24 层 Transformer 模型时，我们将所有算法的小批量大小设为 10。我们为随机梯度下降和带动量的随机梯度下降设置学习率 2×10^{-5}，并为带动量的随机梯度下降设置动量参数 $\beta = 0.9$。Adam、AdamW 和 AdaBelief 使用基本学习率 10^{-6}、调节参数 $\varepsilon = 10^{-8}$、一阶动量参数$\beta_1 = 0.9$和二阶动量参数 $\beta_2 = 0.999$。同时，AdamW 使用权重衰减参数$\lambda = 10^{-4}$，MiAdam 使用基本学习率 $10^{-6}, \varepsilon = 10^{-8}, \beta_1 = 0.9,$ $\beta_2 = 0.999$ 以及多重积分率 $\kappa = 0.999$。SWATS 使用基本学习率 10^{-5}、调节参数 $\varepsilon = 10^{-8}, \beta_1 = 0.9$和 $\beta_2 = 0.99$。我们的 Adam(W)-srf 和 HomeAdam(W) 使用基本学习率 $10^{-6}, \varepsilon = 10^{-5}, \beta_1 = 0.9$和 $\beta_2 = 0.999$。同时，我们的 AdamW-srf 和 HomeAdamW 使用权重衰减参数 $\lambda = 10^{-4}$。

F. HomeAdam 和 HomeAdamW 算法的逐元素变体

在本节中，我们给出 HomeAdam 和 HomeAdamW（HomeAdam-ew 和 HomeAdamW-ew）算法的逐元素变体。由于它与反向传播框架相匹配，因此更适合用于训练深度学习模型。算法3 展示了 HomeAdam-ew 和 HomeAdamW-ew 算法的算法框架。

F.1. 我们的 HomeAdam(W)-ew 算法的泛化分析

在本小节中，我们证明 HomeAdam-ew 和 HomeAdamW-ew 优化器的泛化误差也更小，为 $O(\frac{1}{N})$。

在理论分析中，我们首先定义一个有用的梯度映射 $\mathcal{M}(\cdot,\ \cdot)$，如下所示：

$$\mathcal{M}((\hat{m}_t)_j,(\hat{v}_t)_j)=\begin{cases}\frac{(\hat{m}_t)_j}{(\hat{v}_t)_j+\varepsilon}, & \text{if } (\hat{v}_t)_j\geq\tau\\ (\hat{m}_t)_j, & \text{otherwise}\end{cases} \tag{120}$$

然后，我们可以将算法 3 的第 12 行和第 14 行改写如下：

$$(\theta_t)_j=(\theta_{t-1})_j-\eta(\mathcal{M}((\hat{m}_t)_j,(\hat{v}_t)_j)+\lambda(\theta_{t-1})_j). \tag{121}$$

算法 3 逐元素 HomeAdam(W) 算法

Input: $\eta>0$, $\beta_1,\beta_2\in(0,1)$, $\varepsilon\geq 0$, $\lambda\geq 0$ and $\tau>0$;
Initialize: $\theta_0\in\mathbb{R}^d$, $m_0=0$ and $v_0=0$;
for $t=1,2,\ldots,T$ **do**
  Draw a sample $z_t\sim\mathcal{D}$;
  for $j=1,2,\ldots,d$ (Parallel Execution in Each Element or Layer) **do**
    $(g_t)_j=(\nabla f(\theta_{t-1};z_t))_j$;
    $(m_t)_j=\beta_1(m_{t-1})_j+(1-\beta_1)(g_t)_j$;
    $(v_t)_j=\beta_2(v_{t-1})_j+(1-\beta_2)(g_t)_j^2$;
    $(\hat{m}_t)_j=\frac{(m_t)_j}{1-\beta_1^t}$;
    $(\hat{v}_t)_j=\frac{(v_t)_j}{1-\beta_2^t}$;
    if $(\hat{v}_t)_j\geq\tau$ **then**
      $(\theta_t)_j=(\theta_{t-1})_j-\eta\big(\frac{(\hat{m}_t)_j}{(\hat{v}_t)_j+\varepsilon}+\lambda(\theta_{t-1})_j\big)$;
    else
      $(\theta_t)_j=(\theta_{t-1})_j-\eta((\hat{m}_t)_j+\lambda(\theta_{t-1})_j)$.
    end if
  end for
end for
Output: θ_T.

引理 F.1. 假设序列 $\{\hat{m}_t、\hat{v}_t、m_t、v_t\}_{t=1}^T$ 和 $\{\hat{m}_t^{(i)}、\hat{v}_t^{(i)}、m_t^{(i)}、v_t^{(i)}\}_{t=1}^T$ 分别基于数据集 S 和 $S^{(i)}$通过算法 3生成。不失一般性，设 $\tau\geq 1$，则有

$$\big\|\mathcal{M}(\hat{m}_t,\hat{v}_t)-\mathcal{M}(\hat{m}_t^{(i)},\hat{v}_t^{(i)})\big\|\leq\frac{\sqrt{d}}{(1-\beta_1^t)}\big\|m_t-m_t^{(i)}\big\|+\frac{G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)}\big\|v_t-v_t^{(i)}\big\|. \tag{122}$$

证明。根据算法 3，由于 $S,S^{(i)}\sim\mathcal{D}$，我们有 $(m_t)_j=\beta_1(m_{t-1})_j+(1-\beta_1)(g_t)_j$、$(v_t)_j=\beta_2(v_{t-1})_j+(1-\beta_2)(g_t)_j^2$、$(m_t^{(i)})_j=\beta_1(m_{t-1}^{(i)})_j+(1-\beta_1)(g_t^{(i)})_j$ 以及 $(v_t^{(i)})_j=\beta_2(v_{t-1}^{(i)})_j+(1-\beta_2)(g_t^{(i)})_j^2$。同时，我们有 $(\hat{m}_t)_j=\frac{(m_t)_j}{1-\beta_1^t}$、$(\hat{v}_t)_j=\frac{(v_t)_j}{1-\beta_2^t}$、$(\hat{m}_t^{(i)})_j=\frac{(m_t^{(i)})_j}{1-\beta_1^t}$ 以及 $(\hat{v}_t^{(i)})_j=\frac{(v_t^{(i)})_j}{1-\beta_2^t}$。

由于 $m_t^{(i)}$ 是 $g_t^{(i)}$ 和 $\|g_t^{(i)}\|\leq G$的指数移动平均，因此有 $(\hat{m}_t^{(i)})_j\leq\|\hat{m}_t^{(i)}\|\leq\frac{G}{1-\beta_1^t}$。根据算法 3，

当 $(\hat{v}_t)_j \geq \tau$ 和 $(\hat{v}_t^{(i)})_j \geq \tau$ 时，我们有

$$\begin{aligned}
&\Big|\frac{(\hat{m}_t)_j}{(\hat{v}_t)_j+\varepsilon}-\frac{(\hat{m}_t^{(i)})_j}{(\hat{v}_t^{(i)})_j+\varepsilon}\Big| \\
&=\frac{1}{((\hat{v}_t)_j+\varepsilon)((\hat{v}_t^{(i)})_j+\varepsilon)}\big|((\hat{v}_t^{(i)})_j+\varepsilon)(\hat{m}_t)_j-(\hat{m}_t^{(i)})_j((\hat{v}_t)_j+\varepsilon)\big| \\
&\leq\frac{1}{((\hat{v}_t)_j+\varepsilon)((\hat{v}_t^{(i)})_j+\varepsilon)}\Big(\big|((\hat{v}_t^{(i)})_j+\varepsilon)(\hat{m}_t)_j-(\hat{m}_t^{(i)})_j((\hat{v}_t^{(i)})_j+\varepsilon)\big| \\
&\quad+\big|(\hat{m}_t^{(i)})_j((\hat{v}_t^{(i)})_j+\varepsilon)-(\hat{m}_t^{(i)})_j((\hat{v}_t)_j+\varepsilon)\big|\Big) \\
&=\frac{1}{(\hat{v}_t)_j+\varepsilon}\big|(\hat{m}_t)_j-(\hat{m}_t^{(i)})_j\big|+\frac{(\hat{m}_t^{(i)})_j}{((\hat{v}_t)_j+\varepsilon)((\hat{v}_t^{(i)})_j+\varepsilon)}\big|(\hat{v}_t^{(i)})_j-(\hat{v}_t)_j\big| \\
&\leq\frac{1}{\tau+\varepsilon}\big|(\hat{m}_t)_j-(\hat{m}_t^{(i)})_j\big|+\frac{G}{(1-\beta_1^t)(\tau+\varepsilon)^2}\big|(\hat{v}_t^{(i)})_j-(\hat{v}_t)_j\big| \\
&\leq\frac{1}{(1-\beta_1^t)}\big|(m_t)_j-(m_t^{(i)})_j\big|+\frac{G}{(1-\beta_1^t)(1-\beta_2^t)}\big|(v_t^{(i)})_j-(v_t)_j\big|, \qquad (123)
\end{aligned}$$

其中，最后一个不等式由 $\tau+\varepsilon\geq 1$ 得出。根据算法 3，当 $(\hat{v}_t)_j<\tau$ 和 $(\hat{v}_t^{(i)})_j<\tau$ 时，我们有

$$\big|(\hat{m}_t)_j-(\hat{m}_t^{(i)})_j\big|\leq\frac{1}{(1-\beta_1^t)}\big|(m_t)_j-(m_t^{(i)})_j\big|+\frac{G}{(1-\beta_1^t)(1-\beta_2^t)}\big|(v_t^{(i)})_j-(v_t)_j\big|. \qquad (124)$$

根据上述不等式（123）和（124），我们有

$$\big|\mathcal{M}((\hat{m}_t)_j,(\hat{v}_t)_j)-\mathcal{M}((\hat{m}_t^{(i)})_j,(\hat{v}_t^{(i)})_j)\big|\leq\frac{1}{(1-\beta_1^t)}\big|(m_t)_j-(m_t^{(i)})_j\big|+\frac{G}{(1-\beta_1^t)(1-\beta_2^t)}\big|(v_t^{(i)})_j-(v_t)_j\big|. \qquad (125)$$

因此我们可以得到

$$\big\|\mathcal{M}(\hat{m}_t,\hat{v}_t)-\mathcal{M}(\hat{m}_t^{(i)},\hat{v}_t^{(i)})\big\|_1\leq\frac{1}{(1-\beta_1^t)}\big\|m_t-m_t^{(i)}\big\|_1+\frac{G}{(1-\beta_1^t)(1-\beta_2^t)}\big\|v_t-v_t^{(i)}\big\|_1. \qquad (126)$$

由于 $\|\cdot\|\leq\|\cdot\|_1\leq\sqrt{d}\|\cdot\|$，我们有

$$\begin{aligned}
\big\|\mathcal{M}(\hat{m}_t,\hat{v}_t)-\mathcal{M}(\hat{m}_t^{(i)},\hat{v}_t^{(i)})\big\| &\leq\big\|\mathcal{M}(\hat{m}_t,\hat{v}_t)-\mathcal{M}(\hat{m}_t^{(i)},\hat{v}_t^{(i)})\big\|_1 \\
&\leq\frac{1}{(1-\beta_1^t)}\big\|m_t-m_t^{(i)}\big\|_1+\frac{G}{(1-\beta_1^t)(1-\beta_2^t)}\big\|v_t-v_t^{(i)}\big\|_1 \\
&\leq\frac{\sqrt{d}}{(1-\beta_1^t)}\big\|m_t-m_t^{(i)}\big\|+\frac{G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)}\big\|v_t-v_t^{(i)}\big\|. \qquad (127)
\end{aligned}$$

□

定理 *F.2.* 假设序列 $\{\theta_t\}_{t=1}^T$ 是在数据集 $S=\{z_1,z_2,\cdots,z_N\}$ 上由算法 *3* 生成的。在假设 *4.1*、*4.2*、*4.3* 下，不失一般性，令 $\tau\geq 1$, $\lambda\in[0,\frac{1}{\eta})$, $\beta_1=O(1)$，其中 $\beta_1\in(0,1)$，$\beta_2=O(1)$，其中 $\beta_2\in(0,1)$，$\sigma=O(1)$，$G=O(1)$ 且 $L=O(1)$。若迭代次数较少（即 $T=O(1)$），则令 $\eta=\frac{1}{\sqrt{d}}$；否则令 $\eta=\frac{1}{\sqrt{dT}}$，我们有

$$|\mathbb{E}[F(\theta_T)-F_S(\theta_T)]|\leq O(\frac{1}{N}). \qquad (128)$$

证明。 在数据集上实施算法 3， S 和 $S^{(i)}$，并使用相同的随机索引序列 $\{j_t\}_{t=1}^T$，并令 $\{\theta_t\}_{t=1}^T$ 和 $\{\theta_t^{(i)}\}_{t=1}^T$ 由算法3 在 S 和 $S^{(i)}$ 下生成。不失一般性，令 $\tau\geq 1$ 在算法 3中。

根据上述梯度映射 $\mathcal{M}(\cdot,\ \cdot)$，我们有 $(\theta_t)_j = (\theta_{t-1})_j - \eta\big(\mathcal{M}((\hat{m}_t)_j,\ (\hat{v}_t)_j) + \lambda(\theta_{t-1})_j\big)$ 以及 $(\theta_t^{(i)})_j = (\theta_{t-1}^{(i)})_j - \eta\big(\mathcal{M}((\hat{m}_t^{(i)})_j,\ (\hat{v}_t^{(i)})_j) + \lambda(\theta_{t-1}^{(i)})_j\big)$，对所有 $j = 1,\ 2,\ \cdots,\ d$均成立。因此我们有

$$\theta_t - \theta_t^{(i)} = (1-\eta\lambda)(\theta_{t-1} - \theta_{t-1}^{(i)}) - \eta\big(\mathcal{M}(\hat{m}_t, \hat{v}_t) - \mathcal{M}(\hat{m}_t^{(i)}, \hat{v}_t^{(i)})\big). \tag{129}$$

我们可以得到

$$\begin{aligned}\|\theta_t - \theta_t^{(i)}\| &= \|(1-\eta\lambda)(\theta_{t-1} - \theta_{t-1}^{(i)}) - \eta\big(\mathcal{M}(\hat{m}_t, \hat{v}_t) - \mathcal{M}(\hat{m}_t^{(i)}, \hat{v}_t^{(i)})\big)\| \\ &\le (1-\eta\lambda)\|\theta_{t-1} - \theta_{t-1}^{(i)}\| + \eta\|\mathcal{M}(\hat{m}_t, \hat{v}_t) - \mathcal{M}(\hat{m}_t^{(i)}, \hat{v}_t^{(i)})\| \\ &\le (1-\eta\lambda)\|\theta_{t-1} - \theta_{t-1}^{(i)}\| + \frac{\eta\sqrt{d}}{(1-\beta_1^t)}\big\|m_t - m_t^{(i)}\big\| + \frac{\eta G\sqrt{d}}{(1-\beta_1^t)(1-\beta_2^t)}\big\|v_t - v_t^{(i)}\big\|,\end{aligned} \tag{130}$$

其中，最后一个不等式由引理F.1得出。

根据上述不等式（130），按照定理 4.5 的上述证明，令 $\eta = O(\frac{1}{\sqrt{d}})$，$\beta_1 = O(1)$，$\beta_2 = O(1)$，$\sigma = O(1)$，$G = O(1)$和 $L = O(1)$，我们有

$$\begin{aligned}&\phi_1 = 2(1-\beta_1)\sigma = O(1), \quad \psi_1 = 2(1-\beta_2)G^2 = O(1), \quad \varphi_1 = 2\eta\sqrt{d}\sigma + \frac{2\eta\sqrt{d}G^3}{(1-\beta_1)} = O(1), \\ &\mathbb{E}\|m_1 - m_1^{(i)}\| \le \frac{\phi_1}{N}, \quad \mathbb{E}\|v_1 - v_1^{(i)}\| \le \frac{\psi_1}{N}, \quad \mathbb{E}\|\theta_1 - \theta_1^{(i)}\| \le \frac{\varphi_1}{N} = O(\frac{1}{N}).\end{aligned} \tag{131}$$

若迭代次数较少（即 $T = O(1)$），令 $\eta = \frac{1}{\sqrt{d}}$, $\lambda \in [0, \frac{1}{\eta}]$, $\beta_1 = O(1)$，其中 $\beta_1 \in (0, 1)$，$\beta_2 = O(1)$，其中 $\beta_2 \in (0, 1)$，$\sigma = O(1)$，$G = O(1)$ 且 $L = O(1)$。按照定理 4.5 的上述证明，假设 $\mathbb{E}\|m_t - m_t^{(i)}\| \le \frac{\phi_t}{N}$with $\phi_t = O(1)$，且 $\mathbb{E}\|v_t - v_t^{(i)}\| \le \frac{\psi_t}{N}$with $\psi_t = O(1)$，我们可以得到 $\phi_{t+1} = O(1)$ 和 $\psi_{t+1} = O(1)$。利用上述不等式（130），则有

$$\mathbb{E}\|\theta_{t+1} - \theta_{t+1}^{(i)}\| \le \frac{\varphi_{t+1}}{N} = O(\frac{1}{N}), \tag{132}$$

其中 $\varphi_{t+1} = (1-\eta\lambda)\varphi_t + \frac{\eta\sqrt{d}\phi_{t+1}}{(1-\beta_1^{t+1})} + \frac{\eta G\sqrt{d}\psi_{t+1}}{(1-\beta_1^{t+1})(1-\beta_2^{t+1})} = O(1)$。通过数学归纳法，我们有

$$\mathbb{E}\|\theta_T - \theta_T^{(i)}\| \le \frac{\varphi_T}{N} = O(\frac{1}{N}). \tag{133}$$

如果迭代次数 T 较大，则我们在泛化分析中考虑迭代次数 $t \ge 1$ 。利用上述不等式（130）的递归关系，并通过数学归纳法，按照上述定理4.5的证明，我们假设 $\mathbb{E}\|m_t - m_t^{(i)}\| \le \frac{\phi_t}{N}$满足 $\phi_t = O(t)$，并且 $\mathbb{E}\|v_t - v_t^{(i)}\| \le \frac{\psi_t}{N}$满足 $\psi_t = O(t)$。

令 $\eta = O(\frac{1}{\sqrt{d}})$, $\lambda \in [0, \frac{1}{\eta}]$, $\beta_1 = O(1)$，其中 $\beta_1 \in (0, 1)$，$\beta_2 = O(1)$，其中 $\beta_2 \in (0, 1)$，$\sigma = O(1)$、$G = O(1)$ 和$L = O(1)$，我们有

$$\begin{aligned}&\phi_{t+1} = \beta_1\phi_t + 2(1-\beta_1)\sigma + (1-\beta_1)L\varphi_t = O(t+1), \\ &\psi_{t+1} = \beta_2\psi_t + 4(1-\beta_2)G\sigma + 2(1-\beta_2)GL\varphi_t = O(t+1).\end{aligned}$$

利用上述不等式（130），即可得到

$$\mathbb{E}\|\theta_{t+1} - \theta_{t+1}^{(i)}\| \le \frac{\varphi_{t+1}}{N} = O(\frac{t+1}{N}). \tag{134}$$

通过数学归纳法，我们得到

$$\mathbb{E}\|\theta_T - \theta_T^{(i)}\| \le \frac{\varphi_T}{N} = O(\frac{T}{N}). \tag{135}$$

进一步，设 $\eta = O(\frac{1}{T\sqrt{d}})$，$\lambda \in [0, \frac{1}{\eta})$，$\beta_1 = O(1)$，其中 $\beta_1 \in (0, 1)$，$\beta_2 = O(1)$，其中 $\beta_2 \in (0, 1)$，$\sigma = O(1)$、$G = O(1)$ 和 $L = O(1)$，我们有 $\phi_T = O(1)$、$\psi_T = O(1)$、$\phi_T = O(1)$，

$$\mathbb{E}\|\theta_T - \theta_T^{(i)}\| \leq \frac{\varphi_T}{N} = O(\frac{1}{N}). \tag{136}$$

通过采用假设4.2，即条件 G-Lipschitz $f(\theta;\ z)$ 对任意 $z \in \mathcal{D}$成立，我们有

$$\mathbb{E}|f(\theta_T; z) - f(\theta_T^{(i)}; z)| \leq G\mathbb{E}\|\theta_T - \theta_T^{(i)}\| = O(\frac{1}{N}). \tag{137}$$

对 S、$S^{(i)}$ 以及算法的随机性对上述不等式（137）取期望，并根据上述引理2.2，可得

$$|\mathbb{E}[F(\theta_T) - F_S(\theta_T)]| \leq O(\frac{1}{N}). \tag{138}$$

□

F.2. 我们的 HomeAdam(W)-ew 算法的收敛分析

在本小节中，我们证明我们的 HomeAdam-ew 和 HomeAdamW-ew 优化器也具有快速收敛速率$O(\frac{1}{T^{1/4}})$。

引理 F.3. 假设序列 $\{m_t\}_{t=0}^T$ 由算法 3生成，令 $\beta_1 = 1 - c\eta \in (0, 1)$，$\beta_2 \in (0,1)$，则有

$$\mathbb{E}\|\nabla F(\theta_t) - m_{t+1}\|^2 \leq (1 - c\eta)\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{2}{c\eta}L^2\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + c^2\eta^2\sigma^2, \tag{139}$$

其中 $c > 0$。

证明. 本证明可沿用上述引理 C.1的证明。 □

定理F.4. 假设序列 $\{\theta_t\}_{t=0}^T$ 由算法 3 生成。在假设 5.1、4.2、4.3、5.2 下，并令 $0 \leq \lambda < \min(\frac{1}{\eta}, \frac{1}{\eta T^{\gamma}\sqrt{dr}\hat{G}})$、$\|\theta_0\|_\infty \leq \eta\breve{G}$、$c \geq \frac{16L}{\nu}$、$\beta_1 = 1 - c\eta \in (0,\ 1)$、$\beta_2 \in (0,\ 1)$ 以及 $0 < \eta \leq \frac{\nu}{4L}$，则有

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| \leq r\left(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\nu}} + \frac{4\sqrt{2}}{T^{\gamma-1}\nu} + \frac{4c\sigma\sqrt{\eta}}{\sqrt{L\nu}}\right) + \frac{1}{T^{\gamma-1}}, \tag{140}$$

其中 $\hat{G} = \max(\frac{G}{1-\beta_1}, \frac{G}{(1-\beta_1)(\tau+\varepsilon)})$，$\breve{G} = \min(\frac{G}{1-\beta_1}, \frac{G}{(1-\beta_1)(\tau+\varepsilon)})$，$r = \max(1, G^2+\varepsilon)$，$\nu = \min(1-\beta_1, (\tau+\varepsilon)(1-\beta_1))$和 $\Delta = F(\theta_0) + \frac{1}{L}(\sigma^2 + \beta_1^2 G^2) - F^*$。

证明。 根据我们的算法 3，由于 $(m_t)_j$ 是 $(g_t)_j$ 对所有 $j = 1$、2、$\cdots$、d的指数移动平均，通过使用假设 4.2，即 $|(g_t)_j| \leq \|g_t\| \leq G$，我们有 $|(m_t)_j| \leq \|m_t\| \leq G$。

Let $\hat{G} = \max(\frac{G}{1-\beta_1}, \frac{G}{(1-\beta_1)(\tau+\varepsilon)})$ and $\breve{G} = \min(\frac{G}{1-\beta_1}, \frac{G}{(1-\beta_1)(\tau+\varepsilon)})$.

算法 3 的第 12 行，我们有 $(\theta_t)_j = (\theta_{t-1})_j - \eta\big(\frac{(\hat{m}_t)_j}{(\hat{v}_t)_j+\varepsilon} + \lambda(\theta_{t-1})_j\big)$，于是可以得到当 $(\hat{v}_t)_j \geq \tau$ 时，由

$$\begin{aligned}
|(\theta_t)_j| &= \big|(\theta_{t-1})_j - \eta\big(\frac{(\hat{m}_t)_j}{(\hat{v}_t)_j+\varepsilon} + \lambda(\theta_{t-1})_j\big)\big| \\
&= \big|(1-\eta\lambda)(\theta_{t-1})_j - \eta\frac{(\hat{m}_t)_j}{(\hat{v}_t)_j+\varepsilon}\big| \\
&\leq (1-\eta\lambda)\big|(\theta_{t-1})_j\big| + \eta\big|\frac{(\hat{m}_t)_j}{(\hat{v}_t)_j+\varepsilon}\big| \\
&\leq (1-\eta\lambda)\big|(\theta_{t-1})_j\big| + \frac{\eta G}{(1-\beta_1^t)(\tau+\varepsilon)} \\
&\leq (1-\eta\lambda)^t\big|(\theta_0)_j\big| + \frac{t\eta G}{(1-\beta_1)(\tau+\varepsilon)} \\
&\leq (t+1)\frac{\eta G}{(1-\beta_1)(\tau+\varepsilon)} \leq (t+1)\eta\hat{G}, \qquad (141)
\end{aligned}$$

其中，第一个不等式由 $0 \leq \lambda < \frac{1}{\eta}$ 得出，倒数第二个不等式则由 $\|\theta_0\|_\infty \leq \eta\breve{G} \leq \frac{\eta G}{(1-\beta_1)(\tau+\varepsilon)}$ 得出。

当 $(\hat{v}_t)_j < \tau$ 时，由算法 3 第 14 行可得 $(\theta_t)_j = (\theta_{t-1})_j - \eta\big((\hat{m}_t)_j + \lambda(\theta_{t-1})_j\big)$，进而可得

$$\begin{aligned}
|(\theta_t)_j| &= \big|(\theta_{t-1})_j - \eta\big((\hat{m}_t)_j + \lambda(\theta_{t-1})_j\big)\big| \\
&= \big|(1-\eta\lambda)(\theta_{t-1})_j - \eta(\hat{m}_t)_j\big| \\
&\leq (1-\eta\lambda)\big|(\theta_{t-1})_j\big| + \eta\big|(\hat{m}_t)_j\big| \\
&\leq (1-\eta\lambda)\big|(\theta_{t-1})_j\big| + \frac{\eta G}{(1-\beta_1^t)} \\
&\leq (1-\eta\lambda)^t\big|(\theta_0)_j\big| + \frac{t\eta G}{1-\beta_1} \\
&\leq (t+1)\frac{\eta G}{1-\beta_1} \leq (t+1)\eta\hat{G}, \qquad (142)
\end{aligned}$$

其中，第一个不等式由 $0 \leq \lambda < \frac{1}{\eta}$ 得出，倒数第二个不等式则由 $\|\theta_0\|_\infty \leq \eta\breve{G} \leq \frac{\eta G}{1-\beta_1}$ 得出。利用上述不等式（141）和（142），因此对所有 $j = 1, 2, \cdots, d$

$$\big|(\theta_t)_j\big| \leq (t+1)\eta\hat{G}. \qquad (143)$$

然后我们可以得到

$$\|\theta_t\|_\infty \leq (t+1)\eta\hat{G},\ \|\theta_t\| \leq \sqrt{d}\|\theta_t\|_\infty \leq (t+1)\sqrt{d}\eta\hat{G}. \qquad (144)$$

接下来，我们定义一个对角矩阵 H_t，以统一算法 3 中的第 12 行和第 14 行，如下所示

$$(\theta_t)_j = (\theta_{t-1})_j - \eta\big((H_t)_{jj}(m_t)_j + \lambda(\theta_{t-1})_j\big), \qquad (145)$$

其中 $(H_t)_{jj} = \frac{1}{((\hat{v}_t)_j+\varepsilon)(1-\beta_1^t)}$ 当 $(\hat{v}_t)_j \geq \tau$ 时，否则 $(H_t)_{jj} = \frac{1}{1-\beta_1^t}$。然后我们可以将上述等式（145）改写为向量形式

$$\theta_t = \theta_{t-1} - \eta\big(H_t m_t + \lambda\theta_{t-1}\big). \qquad (146)$$

根据假设 4.2，我们有 $|(g_t)_j^2| \leq G^2$。由于 $(v_t)_j$ 是 $(g_t)_j^2$ 的指数移动平均，因此对于所有 $j = 1, 2, \cdots, d$，我们有 $(v_t)_j \leq G^2$。进一步令 $\frac{1}{r} = \min(1, \frac{1}{G^2+\varepsilon})$，$\frac{1}{\nu} = \max(\frac{1}{1-\beta_1}, \frac{1}{(\tau+\varepsilon)(1-\beta_1)})$，则有 $\frac{1}{r}I_d \preceq H_t \preceq \frac{1}{\nu}I_d$。

接下来，我们可以将上述等式（146）进一步改写如下：

$$\begin{aligned}
\theta_t &= \theta_{t-1} - \eta(H_t m_t + \lambda\theta_{t-1}) \\
&= (1-\eta\lambda)\theta_{t-1} - \eta H_t m_t \\
&= \arg\min_{\theta\in\mathbb{R}^d}\Big\{\langle m_t, \theta\rangle + \frac{1}{2\eta}\big(\theta - (1-\lambda\eta)\theta_{t-1}\big)^T H_t^{-1}\big(\theta - (1-\lambda\eta)\theta_{t-1}\big)\Big\}. \qquad (147)
\end{aligned}$$

利用子问题的最优性条件（147），可得

$$\langle m_t + \frac{1}{\eta} H_t^{-1}\big(\theta_t - (1-\lambda\eta)\theta_{t-1}\big), \theta - \theta_t\rangle \geq 0, \quad \forall \theta \in \mathbb{R}^d. \tag{148}$$

将 $\theta = \theta_{t-1}$ 代入上述不等式（148），可得

$$\langle m_t + \frac{1}{\eta} H_t^{-1}\big(\theta_t - (1-\lambda\eta)\theta_{t-1}\big), \theta_{t-1} - \theta_t\rangle \geq 0. \tag{149}$$

因此，我们可以得到

$$\begin{aligned}\langle m_t, \theta_{t-1} - \theta_t\rangle &\geq \frac{1}{\eta}\langle H_t^{-1}(\theta_t - \theta_{t-1}), \theta_t - \theta_{t-1}\rangle + \lambda\langle H_t^{-1}\theta_{t-1}, \theta_t - \theta_{t-1}\rangle \\ &\geq \frac{\nu}{\eta}\|\theta_t - \theta_{t-1}\|^2 + \lambda\langle H_t^{-1}\theta_{t-1}, \theta_t - \theta_{t-1}\rangle,\end{aligned} \tag{150}$$

其中，最后一个不等式由 $\nu I_d \preceq H_t^{-1} \preceq r I_d$可得。

由于 $\nu I_d \preceq H_t^{-1} \preceq r I_d$ 且 $\|\theta_{t-1}\| \leq \sqrt{d}\|\theta_{t-1}^-\|_\infty \leq t\eta\sqrt{d}\hat{G},^-$因此对所有$t \geq 1$，都有 $\|H_t^{-1}\theta_{t-1}\|^2 \leq r^2\|\theta_{t-1}\|^2 \leq t^2 r^2\eta^2 d\hat{G}^2$。

根据假设 5.1，即 $F(\theta)$ 是 L-光滑的，我们有

$$\begin{aligned}\mathbb{E}[F(\theta_t)] &\leq \mathbb{E}[F(\theta_{t-1}) + \nabla F(\theta_{t-1})^T(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2] \\ &= \mathbb{E}[F(\theta_{t-1}) + (\nabla F(\theta_{t-1}) - m_t)^T(\theta_t - \theta_{t-1}) + m_t^T(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2] \\ &\overset{(i)}{\leq} \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{2\nu}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{\nu}{2\eta}\|\theta_t - \theta_{t-1}\|^2 + m_t^T(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2] \\ &\overset{(ii)}{\leq} \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{2\nu}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{\nu}{2\eta}\|\theta_t - \theta_{t-1}\|^2 - \frac{\nu}{\eta_t}\|\theta_t - \theta_{t-1}\|^2 \\ &\quad - \lambda\langle H_t^{-1}\theta_{t-1}, \theta_t - \theta_{t-1}\rangle + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2] \\ &\leq \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{2\nu}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{\nu}{2\eta}\|\theta_t - \theta_{t-1}\|^2 - \frac{\nu}{\eta}\|\theta_t - \theta_{t-1}\|^2 \\ &\quad + \frac{\lambda^2\eta}{\nu}\|H_t^{-1}\theta_{t-1}\|^2 + \frac{\nu}{4\eta}\|\theta_t - \theta_{t-1}\|^2 + \frac{L}{2}\|\theta_t - \theta_{t-1}\|^2] \\ &\leq \mathbb{E}[F(\theta_{t-1}) + \frac{\eta}{\nu T^{2\gamma-2}} + \frac{\eta}{2\nu}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\nu}{8\eta}\|\theta_t - \theta_{t-1}\|^2],\end{aligned} \tag{151}$$

其中，上述不等式（i）由杨氏不等式得出，上述不等式（150）由上文的不等式得出，而最后一个不等式由 $0 < \eta \leq \frac{\nu}{4L}$和 $0 \leq \lambda \leq \frac{1}{\eta T^\gamma\sqrt{dr}\hat{G}}$得出。

H这里我们定义一个有用的李雅普诺夫函数 $\Psi_t = F(\theta_t) + \frac{1}{2L}\|\nabla F(\theta_t) - m_{t+1}\|^2$。然后我们有 e

$$\begin{aligned}\mathbb{E}[\Psi_t - \Psi_{t-1}] &= \mathbb{E}[F(\theta_t)] - \mathbb{E}[F(\theta_{t-1})] + \frac{1}{2L}(\mathbb{E}\|\nabla F(\theta_t) - m_{t+1}\|^2 - \mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2) \\ &\leq \frac{\eta}{\nu T^{2\gamma-2}} + \frac{\eta}{2\nu}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\nu}{8\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 \\ &\quad + \frac{1}{2L}\big(-c\eta\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{2}{c\eta}L^2\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + c^2\eta^2\sigma^2\big) \\ &\overset{(i)}{\leq} \frac{\eta}{\nu T^{2\gamma-2}} - \frac{c\eta}{4L}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\nu}{16\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + \frac{c^2\eta^2\sigma^2}{2L} \\ &\leq \frac{\eta}{\nu T^{2\gamma-2}} - \frac{4\eta}{\nu}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 - \frac{\nu}{16\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 + \frac{c^2\eta^2\sigma^2}{2L},\end{aligned} \tag{152}$$

其中，第一个不等式由引理 F.3得出，而上述不等式 (i) 由 $c \geq \frac{16L}{\nu}$ 得出，例如 $\frac{\nu}{16\eta} \geq \frac{L}{c\eta}$ 和$\frac{c\eta}{4L} \geq \frac{c\eta}{32L} \geq \frac{\eta}{2\nu}$，最后一个不等式也由 $c \geq \frac{16L}{\nu}$ 得出。然后我们有

$$\frac{4\eta}{\nu}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{\nu}{16\eta}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 \leq \mathbb{E}[\Psi_{t-1} - \Psi_t] + \frac{\eta}{\nu T^{2\gamma-2}} + \frac{c^2\eta^2\sigma^2}{2L}. \tag{153}$$

将上述不等式（153）的两边乘以 $\frac{16}{\eta\nu}$，即可得到

$$\begin{aligned}\frac{1}{\nu^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 &\leq \frac{4}{\nu^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2 \\ &\leq \mathbb{E}[\frac{16(\Psi_{t-1} - \Psi_t)}{\eta\nu}] + \frac{16}{\nu^2 T^{2\gamma-2}} + \frac{8c^2\eta\sigma^2}{L\nu}.\end{aligned} \tag{154}$$

由于 $m_1 = \beta_1 m_0 + (1-\beta_1)g_1 = (1-\beta_1)\nabla f(\theta_0; z_1)$，我们有

$$\begin{aligned}\|\nabla F(\theta_0) - m_1\|^2 &= \|\nabla F(\theta_0) - (1-\beta_1)\nabla f(\theta_0; z_1)\|^2 \\ &= \|\nabla F(\theta_0) - \nabla f(\theta_0; z_1) + \beta_1\nabla f(\theta_0; z_1)\|^2 \leq 2\sigma^2 + 2\beta_1^2 G^2.\end{aligned} \tag{155}$$

给定 $\Psi_0 = F(\theta_0) + \frac{1}{2L}\|\nabla F(\theta_0) - m_1\|^2$，我们有

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}[\frac{1}{\nu^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2] \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{16(\Psi_{t-1} - \Psi_t)}{\eta\nu}] + \frac{16}{T^{2\gamma-2}\nu^2} + \frac{8c^2\eta\sigma^2}{L\nu} \\ &\leq \frac{16(F(\theta_0) + \frac{1}{2L}(2\sigma^2 + 2\beta_1^2 G^2) - F^*)}{T\eta\nu} + \frac{16}{T^{2\gamma-2}\nu^2} + \frac{8c^2\eta\sigma^2}{L\nu}.\end{aligned} \tag{156}$$

令 $\Delta = F(\theta_0) + \frac{1}{L}(\sigma^2 + \beta_1^2 G^2) - F^*$，可将上述不等式（156）改写如下：

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}[\frac{1}{\nu^2}\mathbb{E}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\mathbb{E}\|\theta_t - \theta_{t-1}\|^2] \\ &\leq \frac{16\Delta}{T\eta\nu} + \frac{16}{T^{2\gamma-2}\nu^2} + \frac{8c^2\eta\sigma^2}{L\nu}.\end{aligned} \tag{157}$$

根据詹森不等式，我们可以得到

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{1}{\nu}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\theta_t - \theta_{t-1}\|] \\ &\leq \Big(\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{1}{\nu^2}\|\nabla F(\theta_{t-1}) - m_t\|^2 + \frac{1}{\eta^2}\|\theta_t - \theta_{t-1}\|^2]\Big)^{1/2} \\ &\leq \sqrt{\frac{32\Delta}{T\eta\nu} + \frac{32}{T^{2\gamma-2}\nu^2} + \frac{16c^2\eta\sigma^2}{L\nu}} \leq \frac{4\sqrt{2\Delta}}{\sqrt{T\eta\nu}} + \frac{4\sqrt{2}}{T^{\gamma-1}\nu} + \frac{4c\sigma\sqrt{\eta}}{\sqrt{L\nu}}.\end{aligned} \tag{158}$$

利用 $\theta_t = \theta_{t-1} - \eta\left(H_t m_t + \lambda\theta_{t-1}\right)$，我们有

$$\begin{aligned}
\frac{1}{\nu}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\theta_t - \theta_{t-1}\| &= \frac{1}{\nu}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\eta(H_t m_t + \lambda\theta_{t-1})\| \\
&\geq \frac{1}{\nu}\|\nabla F(\theta_{t-1}) - m_t\| + \|H_t m_t\| - \|\lambda\theta_{t-1}\| \\
&= \frac{1}{\nu}\|H_t^{-1}H_t(\nabla F(\theta_{t-1}) - m_t)\| + \|H_t m_t\| - \|\lambda\theta_{t-1}\| \\
&\geq \|H_t(\nabla F(\theta_{t-1}) - m_t)\| + \|H_t m_t\| - \|\lambda\theta_{t-1}\| \\
&\geq \|H_t\nabla F(\theta_{t-1})\| - \|\lambda\theta_{t-1}\| \\
&\geq \frac{\|\nabla F(\theta_{t-1})\|}{r} - \|\lambda\theta_{t-1}\|, \qquad (159)
\end{aligned}$$

其中，上述不等式由 $\nu I_d \preceq H_t^{-1} \preceq r I_d$ 和 $\frac{1}{r} I_d \preceq H_t \preceq \frac{1}{\nu} I_d$ 成立。

将上述不等式（159）代入（158，可得

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\| &\leq \frac{r}{T}\sum_{t=1}^{T}\mathbb{E}[\frac{1}{\nu}\|\nabla F(\theta_{t-1}) - m_t\| + \frac{1}{\eta}\|\theta_t - \theta_{t-1}\| + \|\lambda\theta_{t-1}\|] \\
&\leq r\big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\nu}} + \frac{4\sqrt{2}}{T^{\gamma-1}\nu} + \frac{4c\sigma\sqrt{\eta}}{\sqrt{L\nu}}\big) + \frac{r}{T}\sum_{t=1}^{T}\mathbb{E}\|\lambda\theta_{t-1}\| \\
&\overset{(i)}{\leq} r\big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\nu}} + \frac{4\sqrt{2}}{T^{\gamma-1}\nu} + \frac{4c\sigma\sqrt{\eta}}{\sqrt{L\nu}}\big) + \frac{r}{T}\sum_{t=1}^{T}\lambda t\sqrt{d}\eta\hat{G} \\
&\overset{(ii)}{\leq} r\big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\nu}} + \frac{4\sqrt{2}}{T^{\gamma-1}\nu} + \frac{4c\sigma\sqrt{\eta}}{\sqrt{L\nu}}\big) + \frac{r}{T}\sum_{t=1}^{T}\frac{1}{\eta T^{\gamma}\sqrt{d}r\hat{G}}t\sqrt{d}\eta\hat{G} \\
&\leq r\big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\nu}} + \frac{4\sqrt{2}}{T^{\gamma-1}\nu} + \frac{4c\sigma\sqrt{\eta}}{\sqrt{L\nu}}\big) + \frac{1}{T^{\gamma-1}}, \qquad (160)
\end{aligned}$$

其中，上述不等式（i）源于 $\|\theta_{t-1}\| \leq t\sqrt{d}\eta\hat{G}$，且上述不等式（$ii$）由 $0 \leq \lambda \leq \frac{1}{\eta T^{\gamma}\sqrt{d}r\hat{G}}$ 成立。因此我们有

$$\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| &\leq r\big(\frac{4\sqrt{2\Delta}}{\sqrt{(T+1)\eta\nu}} + \frac{4\sqrt{2}}{(T+1)^{\gamma-1}\nu} + \frac{4c\sigma\sqrt{\eta}}{\sqrt{L\nu}}\big) + \frac{1}{(T+1)^{\gamma-1}} \\
&\leq r\big(\frac{4\sqrt{2\Delta}}{\sqrt{T\eta\nu}} + \frac{4\sqrt{2}}{T^{\gamma-1}\nu} + \frac{4c\sigma\sqrt{\eta}}{\sqrt{L\nu}}\big) + \frac{1}{T^{\gamma-1}}. \qquad (161)
\end{aligned}$$

□

注 F.5。由上述定理，令 $\eta = \frac{1}{\sqrt{T}}$ 和 $\gamma = \frac{3}{4}$，可得

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| \leq r\big(\frac{4\sqrt{2\Delta}}{\sqrt{\nu}T^{1/4}} + \frac{4\sqrt{2}}{\nu T^{1/4}} + \frac{4c\sigma}{\sqrt{L\nu}T^{1/4}}\big) + \frac{1}{T^{1/4}}. \qquad (162)$$

进一步，令 $G = O(1)$、$L = O(1)$、$\sigma = O(1)$、$\tau = O(1)$、$\beta_1 = O(1)$，其中 $\beta_1 \in (0,\ 1)$，$\beta_2 = O(1)$，其中 $\beta_2 \in (0,\ 1)$，且 $c = O(1)$，我们有 $r = \max(1,\ G^2 + \varepsilon) = O(1)$，$\frac{1}{\nu} = \max(\frac{1}{1-\beta_1},\ \frac{1}{(\tau+\varepsilon)(1-\beta_1)}) = O(1)$，且 $\Delta = F(\theta_0) + \frac{1}{L}(\sigma^2 + \beta_1^2 G^2) - F^* = O(1)$。于是我们有

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla F(\theta_t)\| \leq O(\frac{1}{T^{1/4}}). \qquad (163)$$