arXiv:2303.00565v1 [cs.LG] 1 Mar 2023

# AdaSAM：通过自适应学习率和动量提升尖锐度感知最小化，以训练深度神经网络

Hao Sun、Li Shen、Qihuang Zhong、Liang Ding、Shixiang Chen、
Jingwei Sun、Jing Li、Guangzhong Sun 和 Dacheng Tao*IEEE* 会士

**摘要—尖锐度感知最小化（SAM）优化器已得到广泛研究，因为它通过引入额外的扰动步来平坦化深度学习模型的损失景观，从而能在训练深度神经网络时获得更好的泛化能力。将 SAM 与自适应学习率和动量加速相结合的方法称为 AdaSAM，尽管已有实证研究将其用于训练大规模深度神经网络，但由于分析耦合的扰动步骤、自适应学习率和动量步骤面临三重困难，尚无理论保证。本文尝试分析随机非凸设定下 AdaSAM 的收敛速率。我们从理论上证明，AdaSAM 达到 $\mathcal{O}(1/\sqrt{bT})$ 收敛速率，并且相对于小批量大小 $b$ 具有线性加速性质。具体而言，为了将随机梯度步与自适应学习率及扰动梯度解耦，我们引入延迟二阶动量项，在分析中取期望时对其进行分解，使它们彼此独立。随后，我们通过证明自适应学习率的取值范围有限来对其进行界定，从而使分析得以进行。据我们所知，我们首次给出了结合自适应学习率和动量加速的 SAM 的非平凡收敛速率。最后，我们在若干自然语言处理任务上进行了实验，结果表明，与随机梯度下降、AMSGrad 和 SAM 优化器相比，AdaSAM 能取得更优的性能。**

**索引词—尖锐度感知最小化、自适应学习率、非凸优化、线性加速。**

## I 引言

SHARPNESS-AWARE 最小化（SAM） [1] 是一种强大的优化器，通过显式最小化训练性能与泛化性能之间的差距来训练大规模深度学习模型。它在训练各种深度神经网络方面取得了显著成果，包括残差网络 [1]–[3], 视觉 Transformer [4], [5],语言模型 [6]–[8], 在广泛的基准测试中表现出色。

然而，SAM 类方法在训练深度神经网络期间面临一些问题，尤其是巨大的计算成本和繁重的超参数调优过程。每次迭代中，与随机梯度下降、Adam 等经典优化器相比，SAM 需要计算两倍的梯度 [9], 由于额外的扰动步骤，AMSGrad [10],因此，SAM 每次参数更新都需要进行两次前向传播和反向传播，因此比经典优化器多出一次计算开销。此外，由于训练过程中包含两个步骤，需要设置双倍数量的超参数，这使得学习率调优变得难以承受且成本高昂。

自适应学习率优化方法 [11] 根据历史梯度信息对梯度进行缩放，通过自动调节学习率来加速收敛。这些方法（如 Adagrad [12]、Adam [9],和 AMS-Grad [10], ）已被提出，用于解决计算机视觉、自然语言处理和生成式神经网络任务[11], [13]–[15]。最近，一些研究尝试在 SAM 中结合 SAM、自适应学习率和动量加速这三者的优势，以简化学习率调优。例如， [16] 和 [17]分别使用自适应学习率和动量加速来训练视觉Transformer模型和自然语言处理模型。尽管已经取得了显著的性能，但由于在 SAM 中使用了自适应学习率和动量加速，它们的收敛性仍然未知。由于优化的三个耦合步骤，即自适应学习率估计与 SAM 的动量步骤和扰动步骤相耦合，直接分析其收敛性十分复杂且困难。

本文在非凸随机设定下，分析了采用自适应学习率和动量加速的 SAM（称为 AdaSAM）的收敛速率。为克服分析中的困难，我们提出了一种新方法，将 SAM 的三步训练过程与自适应学习率及动量步骤解耦。分析过程主要分为三部分。首先，分析 SAM 的过程。然后分析采用自适应学习率方法的第二步。我们引入一个来自前一次迭代的二阶动量项，该项与自适应学习率相关，并且在取期望时独立于 SAM。随后，借助自适应学习率的上界，我们可以对由 SAM 和前一次迭代的二阶动量构成的项进行界定。最后，我们分析与 SAM 和自适应学习率相结合的动量加速。动量加速会在收敛性分析中引入一个额外项。在这里，我们引入一个辅助序列来吸收该项，并证明其在所有迭代上的总和是可控的。我们证明，AdaSAM 具有相对于批量大小的线性加速特性，即 $\mathcal{O}(1/\sqrt{bT})$，其中

Hao Sun、Jingwei Sun、Jing Li 和 Guangzhong Sun 就职于中国科学技术大学计算机科学与技术学院，中国合肥 230000。（电子邮箱：ustcsh@mail.ustc.edu.cn、sunjw@ustc.edu.cn、lj@ustc.edu.cn、gzsun@ustc.edu.cn。）

Qihuang Zhong 就职于湖北武汉大学计算机科学学院，邮编 430000。（电子邮箱：zhongqihuang@whu.edu.cn）

Li Shen、Liang Ding、Shixiang Chen 和 Dacheng Tao 任职于京东探索研究院，北京，100000。（电子邮件：mathshenli@gmail.com，liangding.liam@gmail.com，chenshxi- ang@gmail.com，dacheng.tao@gmail.com。）

$b$ 是小批量大小。在实证实验中，我们将 AdaSAM 应用于训练 GLUE 基准上的 RoBERTa 模型，以评估理论发现。结果表明，AdaSAM 在实验中取得了最佳性能，在 8 项任务中赢得了 6 项，并且可以清楚地观察到线性加速。

最后，我们将贡献总结如下：

- 我们首次为随机非凸设定下结合动量加速的自适应 SAM 方法提供了收敛保证。我们的结果表明，由于已确立相对于批量大小的线性加速，较大的小批量有助于收敛。
- 我们在各种任务上进行了一系列实验。结果表明，AdaSAM 优于大多数最先进的优化器，并验证了线性加速。

## II. 预备知识与相关工作

本节首先介绍基本的问题设定，然后概述若干关于 SAM、自适应学习率和动量步的相关工作。

### A. 问题设定

本文聚焦于随机非凸优化

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

其中 $d$ 是变量 $x$的维度， $D$ 是数据样本的未知分布，$f_\xi(x)$ 是一个光滑且可能非凸的函数，而$f_{\xi_i}(x)$ 表示根据数据分布 $\xi_i$ ，在采样数据点 $D$处的目标函数。在机器学习中，经验风险最小化是其特例，且 $f$ 是数据集 $D$包含 $N$ 个数据点时的损失函数，即 $D=\{\xi_i, i=1,2,\dots,N\}$。问题（1）可化为以下有限和问题：

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{N}\sum_i f_{\xi_i}(x). \tag{2}$$

*a)* 符号说明。：若无特别说明，为简化起见，我们将 $f_i(x)$表示为 $f_{\xi_i}(x)$，它是第 $i$个损失函数； $x\in\mathbb{R}^d$ 是模型参数， $d$是参数维度。我们将 $l_2$ 范数记为 $\|\cdot\|_2$。Hadamard 积记为 $a\odot b$ ，其中 $a$、$b$ 是两个向量。对于向量 $a\in\mathbb{R}^d$， $\sqrt{a}$ 表示一个向量，其第 $j$个值$(\sqrt{a})^{(j)}$等于 $a_j$的平方根。

### B. 相关工作

*a*）尖锐度感知最小化：许多研究试图提升深度学习模型训练过程中的泛化能力。一些方法，如 Dropout [18],权重衰减 [19], 和正则化方法 [20], [21]能够以显式方式提升泛化能力。以往研究表明，尖锐极小值可能导致较差的泛化效果，而平坦极小值的表现更好 [22]–[24]。因此，人们普遍认为尖锐度与泛化能力密切相关。尖锐度感知最小化（SAM） [1] 旨在通过在整个邻域内最小化训练损失，显式地寻找平坦极小化解在整个邻域内均匀地。具体而言，SAM 旨在求解以下极小极大鞍点问题：

$$\min_x \max_{\|\delta\|\le\rho} f(x+\delta) + \lambda\|x\|_2^2, \tag{3}$$

其中， $\rho\ge 0$ 和 $\lambda\ge 0$ 是两个超参数。也就是说，最小化的是邻域内 $f(x)$的扰动损失函数，而不是原始损失函数 $f(x)$。通过对 $f(x+\delta)$关于 $\delta$进行泰勒展开，内层最大化问题可近似求解为

$$\begin{aligned}\delta^*(x) &= \arg\max_{\|\delta\|\le\rho} f(x+\delta)\\ &\approx \arg\max_{\|\delta\|\le\rho} f(x) + \delta^\top\nabla f(x)\\ &= \arg\max_{\|\delta\|\le\rho} \delta^\top\nabla f(x) = \rho\frac{\nabla f(x)}{\|\nabla f(x)\|}.\end{aligned}$$

通过舍弃二次项， （3）可简化为以下最小化问题

$$\min_x f\left(x+\rho\frac{\nabla f(x)}{\|\nabla f(x)\|}\right). \tag{4}$$

批次数据$b$ 上的 $f\left(x+\rho\frac{\nabla f}{\|\nabla f}{}^{(x)}{}^{(x)}\|\right)$ 的随机梯度包含海森矩阵-向量积，SAM进一步将梯度近似为

$$\nabla_x f_b\left(x+\rho\frac{\nabla f_b(x)}{\|\nabla f_b(x)\|}\right) \approx \nabla_x f_b(x)|_{x+\rho\frac{\nabla f_b(x)}{\|\nabla f_b(x)\|}}.$$

随后，沿着负方向 $-\nabla_x f_b(x)|_{x+\rho\ \nabla f_b{}^{(x)}\|\nabla f_b{}^{(x)}\|}$，使用随机梯度下降求解替代最小化问题（4）。显然，SAM需要进行两次梯度反向传播，即 $\nabla f_b(x)$和 $\nabla_x f_b(x)|_{x+\rho\ \nabla f_b{}^{(x)}\|\nabla f_b{}^{(x)}\|}$。由于存在超参数 $\rho$，因此需要仔细调节SAM中的 $\rho$ 和学习率。在实践中，预先设定 $\rho$ 以控制邻域的半径。

最近，已有多种 SAM 变体被提出，以提升其性能。例如， [8], [16], [17] 在经验上将自适应学习率与 SAM 结合，并展现出出色的泛化准确率，但其收敛性尚未得到研究。ESAM [25] 提出了一种通过稀疏化梯度来降低反向传播双重计算开销的高效方法。ASAM [17]通过自适应缩放邻域来改进 SAM，使尖锐度不受参数重新缩放的影响。GSAM[16] 同时最小化扰动函数和新定义的替代间隙函数，以进一步提升极小化点的平坦性。Liu 等人 [26] 还研究了大批量训练场景下的 SAM，并定期更新扰动梯度。最近，[3], [8] 采用稀疏梯度扰动技术提升了 SAM 的效率。 [27], [28]将 SAM 扩展至联邦学习场景，并显著提升了性能。另一方面，也有一些工作分析了 SAM 的收敛性，例如 [29] ，但未考虑归一化步骤，即 $\nabla f_{b(x)}\|\nabla f_b(x)\|$中的归一化。

*b)* 自适应优化器：自适应优化器可以根据历史梯度自动调整学习率。第一种自适应方法 Adagrad [12],在凸情形下能够取得比其他一阶方法更好的结果。然而，在训练深度神经网络时，Adagrad 会迅速降低学习率，导致性能下降。Adadelta [30] 旨在改变这一状况，并引入了基于历史梯度指数平均的学习率。Adam [9] 进一步加入动量步骤以稳定训练过程，并在许多任务中展现出优异性能。然而，Reddi 等人[10] 给出了一个反例，表明即使目标函数是凸函数，该方法也可能无法收敛；他们还提出了一种名为 AMSGrad 的替代方法，并为其提供了收敛保证。随后，许多工作 [31]–[44] 被提出，用于研究自适应方法及其变体在非凸情形下的收敛性。然而，由于扰动步骤与自适应学习率相互耦合，这些分析技术无法直接用于证明采用自适应学习率的 SAM 的收敛性。

*c)* 动量加速：诸如 Polyak 重球法 [45],Nestrov 加速梯度下降法 [46] 以及加速投影法[47]等动量方法被用于优化深度神经网络的参数。在实践中，它们已被用于加速联邦学习任务 [48], 非负潜因子模型[49] 和推荐系统 [50]。有许多理论工作 [51]–[53]专注于分析针对非凸问题优化的动量加速。 [54] 表明，在训练深度神经网络时调节动量十分重要。 [55]首次指出了随机动量方法的线性收敛结果。 [56] 提出了一类加速零阶和一阶动量方法，用于解决小规模优化和极小极大优化问题。[57]通过引入 RNA 方案和具有非线性更新的约束形式 RNA，扩展了动量方法。[58]提出了一种启发式自适应重启方法，而 [59] 提出了一种名为 SRSGD 的定期重启动量加速 SGD 方法，有助于缩短训练时间。 [60] 在分布式梯度算法中加入了一个动量项。

## III. 方法

本节介绍一种采用自适应学习率和动量加速的 SAM 方法，称为 AdaSAM，旨在稳定 SAM 的训练过程并降低学习率调优的难度。随后，我们给出 AdaSAM 的收敛结果。最后，我们概述主定理的证明。

### A. *AdaSAM* 算法

用于求解问题（1）的 AdaSAM 算法如算法1所示。在每次迭代中，计算批量大小为 $g_t$ 、取点$x+\epsilon$（$x$）处的小批量梯度估计 ，即:

$$g_t = \nabla_x f_b(x)|_{x_t+\epsilon(x_t)} = \frac{1}{b}\sum_{i\in B}\nabla f_{\xi_i}(x_t+\delta(x_t)).$$

**算法 1：** AdaSAM：采用自适应学习率的 SAM 率和动量加速

**输入：** 初始参数 $x_0$, $m_{-1}=0$, $\hat{v}_{-1}=\epsilon^2$(a 用于避免分母为零的小正标量 递减）、基础学习率 $\gamma$, 邻域大小 $\rho$ 和动量 参数 $\beta_1$、 $\beta_2$。

**输出：** 优化后的参数 $x_{T+1}$

1 **对于** 迭代 $t \in \{0,1,2,...,T-1\}$ **执行**

2 采样小批量 $B=\{\xi_{t_1},\xi_{t_2},...,\xi_{t\ \ \ \ |B|}\}$;

3 计算梯度 $s_t = \nabla_x f_B(x)|_{x\ t} = \frac{1}{b}\sum_{i\in B}\nabla f_{t_i}(x_t)$;

4 计算 $\delta(x_t) = \rho_t \ \ \frac{s_t}{\|s_t\|}$;

5 计算 SAM 梯度 $g_t = \nabla_x f_B(x)|_{x_t+\delta(x_t)}$;

6 $m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t$;

7 $v_t = \beta_2 v_{t-1} + (1-\beta_2)[g_t]^2$;

8 $\hat{v}_t = \max(\hat{v}_{t-1},\ v_t)$;

9 $\eta_t = 1/\sqrt{\hat{v}_t}$;

10 $x_{t+1} = x_t - \gamma m_t \odot \eta_t$;

11 **结束**

这里， $\delta(x_t)$ 是 SAM 中额外的扰动梯度步，定义如下

$$\delta(x_t) = \rho\frac{s_t}{\|s_t\|},\ \text{where}\ s_t = \nabla_x f_b(x)|_{x_t} = \frac{1}{b}\sum_{i\in B}\nabla f_{\xi_i}(x_t).$$

随后， $g_t$ 的动量项和二阶动量项 $[g_t]^2$ 分别累积计算为 $m_t$ 和 $v_t$。AdaSAM 随后沿着 $-m_t$ ，以自适应学习率 $\gamma\eta_t$更新迭代点。

**注 1.** 下面，我们对 *AdaSAM* 作几点说明：

- 当 $\beta_2 = 1$ 时，自适应学习率会退化为与随机梯度下降相同的递减学习率。此时，*AdaSAM* 退化为经典的 *SAM* 优化器。
- 如果去掉第 *8* 行 $\hat{v}_t = \max(\hat{v}_{t-1},\ v_t)$，那么我们的算法就会变成 *Adam* 的一个变体。*Adam* 在 *[10]*中不收敛的反例同样适用于 *SAM* 变体，而 *AdaSAM* 则可以收敛。

### B. 收敛性分析

在给出 AdaSAM 算法的收敛结果之前，我们先介绍一些必要的假设。

假设 *1*（*L*-光滑）。 $f_i$ 和 $f$ 可微，并具有梯度利普希茨性质： $\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x-y\|$, $\|\nabla f(x)-\nabla f(y)\| \le L\|x-y\|$, $\forall x,y\in\mathbb{R}^d, i=1,2,...,N$，这也意味着下降不等式，即 $f_i(y) \le f_i(x)+\langle\nabla f_i(x)y-x\rangle + \frac{L}{2}\|y-x\|^2$。

**假设 2 (有界方差)**。 梯度的估计量是无偏的，且随机梯度的方差有界。即，

$$\mathbb{E}\nabla f_i(x) = \nabla f(x),\quad \mathbb{E}\|\nabla f_i(x)-\nabla f(x)\|^2 \le \sigma^2.$$

当采用小批量大小 $b$ 时，有 $\mathbb{E}\|\nabla f_b(x)$ $-$ $\nabla f\ x\ \|^2 \le \frac{\sigma^2}{b}$. ()

**假设 3 (有界随机梯度)。** 随机梯度一致有界，即，

$$\|\nabla f_i(x)\|_\infty \le G, for\ any\ i = 1, \dots, N.$$

**注2.** 上述假设通常用于证明诸如 *[31], [32], [61], [62]* 的自适应随机梯度方法的收敛性。

下面我们简要说明分析 AdaSAM 算法收敛性的主要思路。首先，我们讨论将自适应学习率应用于 SAM 时所面临的困难。我们注意到，收敛性分析中包含自适应学习率的主要步骤，是估计期望 $\mathbb{E}[x_{t+1} - x_t] = -\mathbb{E}m_t \odot \eta_t = -\mathbb{E}(1-\beta_1)g_t \odot \eta_t - \mathbb{E}\beta_1 m_{t-1} \odot \eta_t$，该期望以滤过 $\sigma(x_t)$ 为条件。在这一部分中，我们考虑 $\beta_1 = 0$ 不包含动量的情况。然后，我们应用延迟技术来解耦 $g_t$ 与 $\eta_t$ 之间的依赖关系，即

$$\begin{aligned}\mathbb{E}g_t \odot \eta_t &= \mathbb{E}[g_t \odot \eta_{t-1}] + \mathbb{E}[g_t \odot (\eta_t - \eta_{t-1})] \\ &= \nabla f(x_t) \odot \eta_{t-1} + \mathbb{E}[g_t \odot (\eta_t - \eta_{t-1})].\end{aligned}$$

第二项 $\mathbb{E}[g_t \odot (\eta_t - \eta_{t-1})]$ 被第一项 $\nabla f(x_t) \odot \eta_{t-1}$所支配。因此，不难得到采用 AMSGrad 等自适应学习率方法的随机梯度下降的收敛结果。然而，当我们将相同策略应用于 AdaSAM 时，发现$\mathbb{E}g_t \odot \eta_{t-1} = \mathbb{E}\nabla_x f_b\left(x + \rho \frac{\nabla f_b}{\|\nabla f_b\|}\right) \ne \nabla f(x_t)$ [29, 无法以类似方式处理，因为 $\eta_{t-1}$ 和 $\nabla f_b(x)$。受引理 16], 的启发，我们的关键观察是

$$\begin{aligned}\mathbb{E}\nabla_x f_b\left(x + \rho\frac{\nabla f_b(x)}{\|\nabla f_b(x)\|}\right) &\approx \mathbb{E}\nabla_x f_b\left(x + \rho\frac{\nabla f(x)}{\|\nabla f(x)\|}\right) \\ &= \nabla_x f\left(x + \rho\frac{\nabla f(x)}{\|\nabla f(x)\|}\right)\end{aligned}$$

并且我们证明了其他项，例如 $\mathbb{E}\left(\nabla_x f_b\left(x + \rho\frac{\nabla f_b(x)}{\|\nabla f_b(x)\|}\right) - \nabla_x f_b\left(x + \rho\frac{\nabla f}{\|\nabla f(x)\|}\right)\right) \odot \eta_{t-1}$ 的值较小，不会主导收敛速率。

另一方面，当我们采用动量步时，会发现项 $\mathbb{E}m_{t-1} \odot \eta_t$ 不能被忽略。通过引入辅助序列 $z_t = x_t + \frac{\beta_1}{1-\beta_1}(x_t - x_{t-1})$，我们有 $\mathbb{E}[z_{t+1} - z_t] = -\mathbb{E}[\frac{\beta_1}{1-\beta_1}\gamma m_{t-1} \odot (\eta_{t-1} - \eta_t) - \gamma g_t \odot \eta_t]$。第一项包含动量项，由于自适应学习率 $\eta_t$的差异，其值较小。因此，它会逐渐减小，而不会损害收敛速率。

**定理 1。** 在假设 *1*、*2*、*3*以及 $\gamma$ 是满足 $\gamma \le \frac{\epsilon}{16L}$的固定数时，对于由算法 *1*生成的序列$\{x_t\}$，我们有如下收敛速率

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|_2^2 \le \frac{2G(f(x_0)-f^*)}{\gamma T} + \frac{8G\gamma L}{\epsilon}\frac{\sigma^2}{b\epsilon} + \Phi \quad (5)$$

其中

$$\begin{aligned}\Phi = &\frac{45GL^2\rho_t^2}{\epsilon} + \frac{2G^3}{(1-\beta_1)T}d(\frac{1}{\epsilon} - \frac{1}{G}) + \frac{6\gamma^2L^2\beta_1^2}{(1-\beta_1)^2}\frac{dG^3}{\epsilon^3} \\ &+ \frac{2(4 + (\frac{\beta_1}{1-\beta_1})^2)\gamma LG^3}{T}d(\epsilon^{-2} - G^{-2}) + \frac{8G\gamma L}{\epsilon}\frac{L\rho_t^2}{\epsilon}, \quad (6)\end{aligned}$$

其中，$T$ 是迭代次数，$f^*$ 是函数 $f$ 的最小值，$\gamma$ 是基础学习率，$b$ 是迷你批量大小，$d$ 是参数 $x$ 的维度。$\beta_1$、$G$、$L$、$\epsilon$、$\sigma^2$、$d$是固定常数。

定理 1 刻画了由 AdaSAM 生成的序列 $\{x_t\}$ 关于随机梯度残差的收敛速率。不等式（5）右侧的前两项是主导收敛速率的项。与前两项相比，$\Phi$ 是一个较小的值，同时我们将邻域大小 $\rho$ 和学习率$\gamma$ 设为与较大迭代次数相关的较小值$T$。于是，我们可以直接得到以下推论。

**推论 1（小批量线性加速）。** 在定理*1*的相同条件下。此外，当我们选择基础学习率 $\gamma = O(\sqrt{\frac{b}{T}})$ 和邻域大小 $\rho = O(\sqrt{\frac{1}{bT}})$ 时，有如下结果：

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|_2^2 = &O\left(\frac{1}{\sqrt{bT}}\right) + O\left(\frac{1}{bT}\right) + O\left(\frac{1}{T}\right) \\ &+ O\left(\frac{1}{b^{\frac{1}{2}}T^{\frac{3}{2}}}\right) + O\left(\frac{b^{\frac{1}{2}}}{T^{\frac{3}{2}}}\right) + O\left(\frac{b}{T}\right).\end{aligned}$$

当 $T$ 足够大时，我们可以获得关于小批量大小 $b$的线性加速收敛速率，即

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|_2^2 = O\left(\frac{1}{\sqrt{bT}}\right). \quad (7)$$

**备注 3。** 关于上述结果，有以下两点说明：

- 要达到一个 $O(\delta)$ 驻点，当批量大小为 *1* 时，需要进行 $T = O(\frac{1}{\delta^2})$ 次迭代。当批量大小为 $b$时，需要运行 $T = O(\frac{1}{b\delta^2})$ 步数。批量大小为 $b$ 的方法比批量大小为 *1* 时快 $b$ 倍，这意味着它具有小批量线性加速特性。
- 根据 *[37], [63], [64], AdaSAM* 可以扩展为分布式版本，并且在参数服务器设置下，关于工作节点数量实现线性加速特性。

### C. 证明概要

本部分给出定理 1 的证明概要。完整证明请参见附录。下面，我们先介绍一个辅助序列 $z_t = x_t + \frac{\beta_1}{1-\beta_1}(x_t - x_{t-1})$。应用 $L$-光滑条件，可得

$$f(z_{t+1}) \le f(z_t) + \langle \nabla f(z_t), z_{t+1} - z_t\rangle + \frac{L}{2}\|z_{t+1} - z_t\|^2. \quad (8)$$

将其应用于序列 $\{z_t\}$ 并采用延迟策略，可得

$$\begin{aligned}&f(z_{t+1}) - f(z_t) \\ &\le \langle \nabla f(z_t), \frac{\gamma\beta_1}{1-\beta_1}m_{t-1} \odot (\eta_{t-1} - \eta_t)\rangle + \frac{L}{2}\|z_{t+1} - z_t\|^2 \\ &\quad + \langle \nabla f(z_t), \frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t + \rho_t\frac{s_t}{\|s_t\|}) \odot (\eta_{t-1} - \eta_t)\rangle\end{aligned}$$

表 I：使用 $\beta_1 = 0.9$ 在 GLUE 基准上评估 SGD、SAM、AMSGrad 和 AdaSAM

| 模型 | CoLA mcc. | SST-2 Acc. | MRPC 准确率/F1值 | STS-B 皮尔逊相关系数/斯皮尔曼相关系数 | RTE Acc. | MNLI 匹配/不匹配 | QNLI Acc. | QQP F1值/准确率 | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| SGD | 9.25 | 50.92 | 68.38/ 81.22 | 3.22/ 1.9 | 55.6 | 84.94/ 84.87 | 63.61 | 85.6/ 80.14 | 55.8 |
| SAM($\rho$ =0.01) | 4.64 | 95.87 | 70.58/ 81.98 | 84.74/ 85.57 | 52.71 | 90.5/ 90.19 | 94.44 | 84.7/ 87.88 | 76.98 |
| SAM($\rho$ =0.005) | 66.76 | 95.76 | 68.38/ 81.22 | 2/ 2 | 52.71 | 90.42/ 89.74 | 94.6 | 86.72/ 89.94 | 68.35 |
| SAM(best) | 66.76 | 95.87 | 70.58/ 81.98 | 84.74/ 85.57 | 52.71 | 90.5/ 90.19 | 94.6 | 86.72/ 89.94 | 82.51 |
| AMSGrad | 68.0 | 96.33 | 90.2/ 92.72 | 91.72/ 91.48 | 87.73 | 90.67/ 90.41 | 94.82 | 88.7/ 91.41 | 89.52 |
| AdaSAM($\rho$ =0.01) | 65.29 | 96.33 | 91.18/ 93.64 | 90.13/ 90.36 | 84.84 | 90.97/ 90.42 | 94.65 | 88.55/ 91.23 | 88.97 |
| AdaSAM($\rho$ =0.005) | 68.74 | 96.67 | 90.93/ 93.36 | 91.64/ 91.38 | 87.73 | 90.88/ 90.4 | 94.56 | 88.69/ 91.33 | 89.69 |
| AdaSAM($\rho$ =0.001) | 67.3 | 96.1 | 90.2/ 92.96 | 91.9/ 91.62 | 85.92 | 90.45/ 90.4 | 94.56 | 88.64/ 91.27 | 89.28 |
| AdaSAM（最佳） | 68.74 | 96.67 | 91.18/ 93.64 | 91.9/ 91.62 | 87.73 | 90.97/ 90.42 | 94.65 | 88.69/ 91.33 | 89.8 |

$$
\begin{aligned}
&+\langle\nabla f(z_t)-\nabla f(x_t),-\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|})\odot\eta_{t-1}\rangle\\
&+\langle\nabla f(x_t),-\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})\odot\eta_{t-1}\rangle\\
&+\langle\nabla f(x_t),\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})\odot\eta_{t-1}\\
&\quad-\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|})\odot\eta_{t-1}\rangle. \qquad (9)
\end{aligned}
$$

根据附录中的引理5、引理6和引理7，我们可以对上面的（9）中的各项作如下界定

$$
\begin{aligned}
&\langle\nabla f(z_t),\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|})\odot(\eta_{t-1}-\eta_t)\rangle\\
&\le\gamma G^2\|\eta_{t-1}-\eta_t\|_1, \qquad (10)\\
&\langle\nabla f(z_t),\frac{\gamma\beta_1}{1-\beta_1}m_{t-1}\odot(\eta_{t-1}-\eta_t)\rangle\\
&\le\frac{\gamma\beta_1}{1-\beta_1}G^2\|\eta_{t-1}-\eta_t\|_1, \qquad (11)\\
&\langle\nabla f(x_t),\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})\odot\eta_{t-1}\\
&\quad-\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|})\odot\eta_{t-1}\rangle\\
&\le\frac{\gamma}{2\mu^2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{2\mu^2\gamma L^2\rho_t^2}{\epsilon}. \qquad (12)
\end{aligned}
$$

然后将它们代入（9），并取条件期望，得到

$$
\begin{aligned}
&\mathbb{E}f(z_{t+1})-f(z_t)\\
&\le\mathbb{E}\langle\nabla f(x_t),-\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})\odot\eta_{t-1}\rangle\\
&+\frac{\gamma}{2\mu^2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{\gamma}{1-\beta_1}G^2\|\eta_{t-1}-\eta_t\|_1\\
&+\mathbb{E}\langle\nabla f(z_t)-\nabla f(x_t),-\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|})\odot\eta_{t-1}\rangle\\
&+\frac{2\mu^2\gamma L^2\rho_t^2}{\epsilon}+\frac{L}{2}\mathbb{E}\|z_{t+1}-z_t\|^2, \qquad (13)
\end{aligned}
$$

其中 $\mu>0$ 是待确定的常数。接下来，根据附录中的引理8、引理 10 和引理 9，我们有

$$
\begin{aligned}
&\mathbb{E}\langle\nabla f(x_t),-\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})\odot\eta_{t-1}\rangle\\
&\le-\gamma\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\mathbb{E}\frac{\gamma}{2\alpha^2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2\\
&\quad+\frac{\gamma\alpha^2L^2\rho_t^2}{2\epsilon}, \qquad (14)\\
&\frac{L}{2}\mathbb{E}\|z_{t+1}-z_t\|^2\le\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2\\
&\quad+\gamma^2L(3\frac{1+\beta}{\beta\epsilon}(\frac{L\rho_t^2}{\epsilon}+\frac{\sigma^2}{b\epsilon}+\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2)\\
&\quad+(1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2), \qquad (15)\\
&\mathbb{E}\langle\nabla f(z_t)-\nabla f(x_t),-\frac{\gamma}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|})\odot\eta_{t-1}\rangle\\
&\le\frac{\gamma^3L^2\beta_1^2}{2\epsilon(1-\beta_1)^2}(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2})\frac{dG_\infty^2}{\epsilon^2}+\frac{\gamma L^2\rho_t^2}{2\epsilon}(\lambda_2^2+4\lambda_3^2)\\
&\quad+\frac{\gamma\lambda_1^2}{2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2. \qquad (16)
\end{aligned}
$$

接下来，将其代入（13）。对所有历史信息取期望可得

$$
\begin{aligned}
&\mathbb{E}f(x_{t+1})-\mathbb{E}f(x_t)\\
&\le-\gamma(1-\frac{1}{2\mu^2}-\frac{1}{2\alpha^2}-\frac{3\gamma L(1+\beta)}{\beta\epsilon}-\frac{\lambda_1^2}{2})\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2\\
&+\frac{2\mu^2\gamma L^2\rho_t^2}{\epsilon}+\frac{\gamma}{1-\beta_1}G^2\mathbb{E}\|\eta_{t-1}-\eta_t\|_1+\frac{\gamma\alpha^2L^2\rho^2}{2\epsilon}\\
&+\frac{\gamma^3L^2\beta_1^2}{2\epsilon(1-\beta_1)^2}(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2})\frac{dG_\infty^2}{\epsilon^2}+\frac{\gamma L^2\rho_t^2}{2\epsilon}(\lambda_2^2+4\lambda_3^2)\\
&+\gamma^2LG^2((\frac{\beta_1}{1-\beta_1})^2+1+\beta)\mathbb{E}\|\eta_t-\eta_{t-1}\|^2\\
&+\frac{3\gamma^2L(1+\beta)}{\beta\epsilon}(\frac{L\rho_t^2}{\epsilon}+\frac{\sigma^2}{b\epsilon}). \qquad (17)
\end{aligned}
$$

我们设定 $\mu^2=\alpha^2=8$、$\beta=3$、$\lambda_1^2=\frac{1}{4}$、$\lambda_2^2=\lambda_3^2=1$，并选择 $\frac{2\gamma L}{\epsilon}\le\frac{1}{8}$。注意，$\eta_t$ 有界。我们有

$$
\begin{aligned}
&\frac{\gamma}{2G}\mathbb{E}\|\nabla f(x_t)\|^2\le\frac{\gamma}{2}\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2 \qquad (18)\\
&\le-\mathbb{E}f(x_{t+1})+\mathbb{E}f(x_t)+\frac{45\gamma L^2\rho_t^2}{2\epsilon}+\frac{4\gamma^2L}{\epsilon}(\frac{L\rho_t^2}{\epsilon}+\frac{\sigma^2}{b\epsilon})\\
&+\frac{\gamma}{1-\beta_1}G^2\mathbb{E}\|\eta_{t-1}-\eta_t\|_1+\frac{3\gamma^3L^2\beta_1^2}{(1-\beta_1)^2}\frac{dG_\infty^2}{\epsilon^3}\\
&+(4+(\frac{\beta_1}{1-\beta_1})^2)\gamma^2LG^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2. \qquad (19)
\end{aligned}
$$

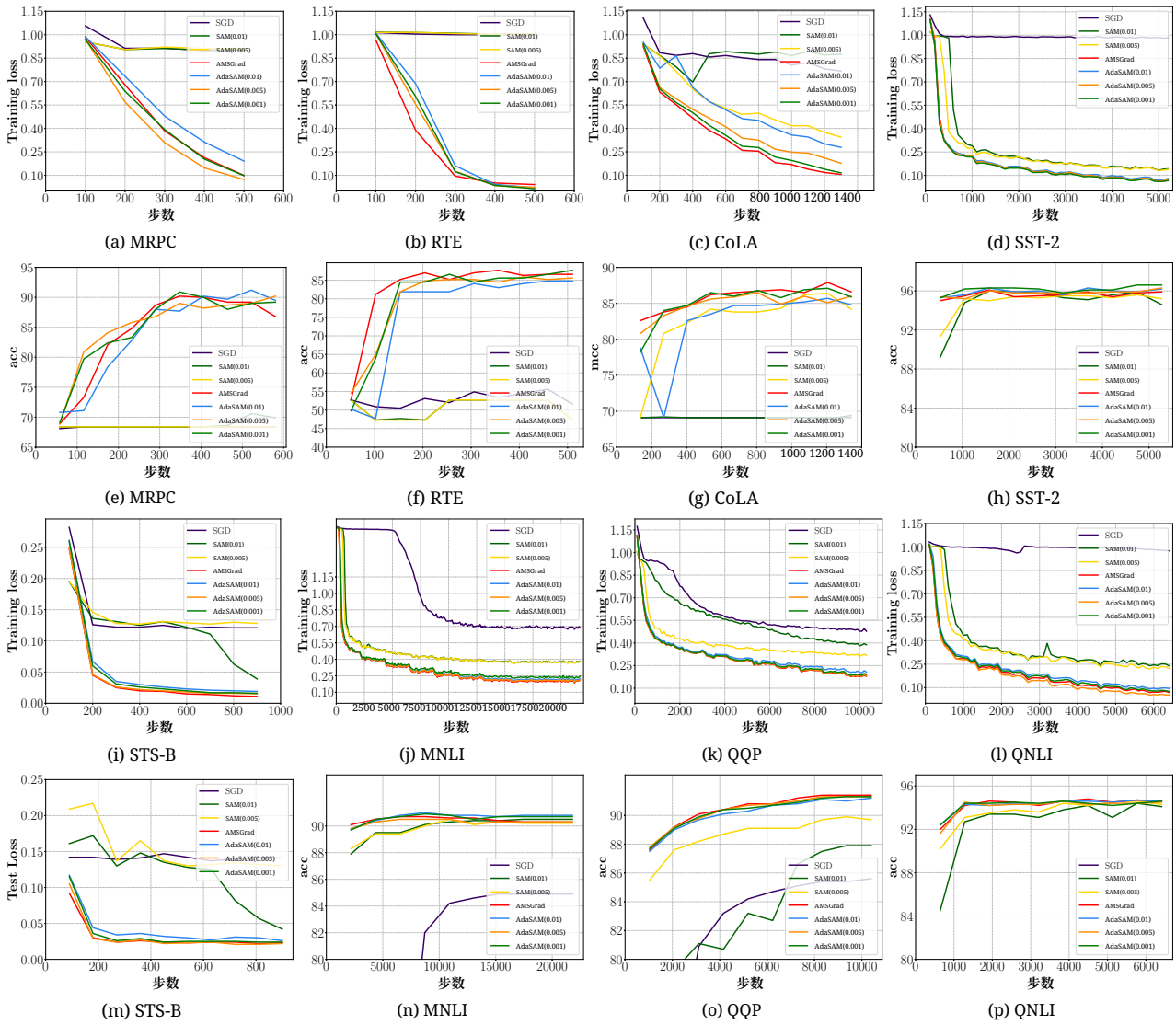



图 1：MRPC、RTE、CoLA、SST-2、STS-B、MNLI、QQP 和 QNLI 上的损失与评估指标随步数的变化。($\beta_1 = 0.9$)

然后，将其从 $t = 0$ 到 $t = T - 1$ 进行求和，并假设 $\gamma$ 为常数，可得

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{2G(f(x_0)-f^*)}{\gamma T} + \frac{8G\gamma L}{\epsilon}\frac{\sigma^2}{b\epsilon} + \frac{45GL^2\rho_t^2}{\epsilon} + \frac{2G^3}{(1-\beta_1)T}d(\frac{1}{\epsilon}-\frac{1}{G}) + \frac{6\gamma^2L^2\beta_1^2}{(1-\beta_1)^2}\frac{dG^3}{\epsilon^3} + \frac{8G\gamma L}{\epsilon}\frac{L\rho_t^2}{\epsilon} + \frac{2(4+(\frac{\beta_1}{1-\beta_1})^2)\gamma LG^3}{T}d(\epsilon^{-2}-G^{-2}), \quad (20)$$

从而完成证明。

## IV. 实验

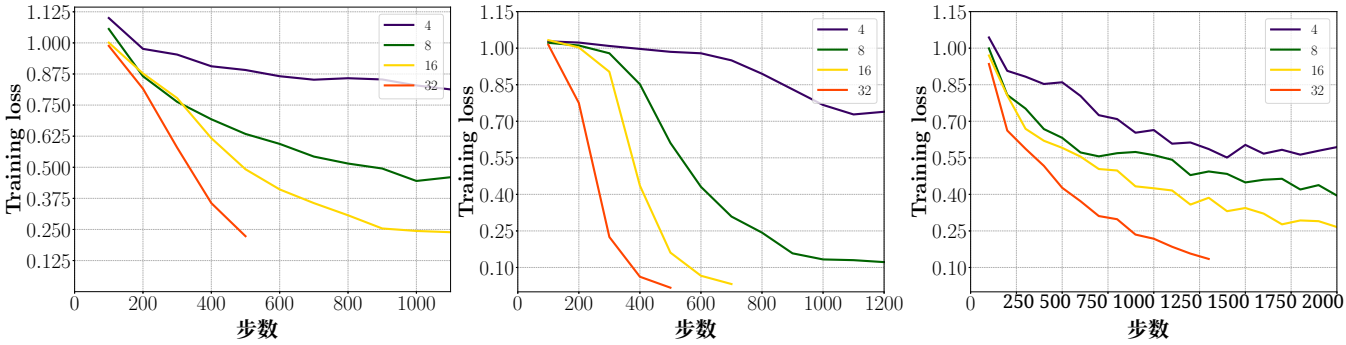
在本节中，我们使用 AdaSAM 训练语言模型，并将其与随机梯度下降、AMSGrad 和 SAM 进行比较，以展示其有效性。受篇幅限制，更多实验（包括可视化、任务描述、实现细节和结果描述）置于附录中。

### A. 实验设置

**任务与数据集**。我们在一个热门基准 即 通用语言理解评估（GLUE） [65],上评估 AdaSAM，该基准包含若干语言理解任务，包括情感分析、问答和文本蕴含。为确保比较公平，我们报告单任务训练的结果，不采用多任务训练或集成训练。对于大多数任务，我们使用准确率（“准确率”）指标评估性能；例外情况是 QQP 和 MRPC 使用 F1分数，STS-B 使用皮尔逊-斯皮尔曼相关系数（“*Pcor/Scor*”），CoLA 使用马修斯相关系数（“*Mcc*”）。指标值越高，性能越好。

**实现**。我们使用广泛采用的预训练语言模型 RoBERTa-large[1] ，并借助开源工具包 fairseq[2],开展实验。该模型包含 24 个 Transformer层，隐藏层大小为 1024。针对每项任务进行微调时，我们采用不同的超参数组合，包括

[1]https://dl.fbaipublicfiles.com/fairseq/models/roberta.large.tar.gz [2] https://github.com/facebookresearch/fairseq




(a) MRPC (b) RTE (c) CoLA

图 2：对批量大小为 4、8、16、32 时 AdaSAM 的线性加速进行验证。

表 II：在不使用动量的情况下，随机梯度下降、SAM、AMSGrad 和 AdaSAM 在 GLUE 基准上的结果，即 $\beta_1 = 0$

| 模型 | CoLA mcc. | SST-2 Acc. | MRPC 准确率/F1值 | STS-B 皮尔逊相关系数/斯皮尔曼相关系数 | RTE Acc. | MNLI 匹配/不匹配 | QNLI Acc. | QQP F1值/准确率 | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| SGD | 0 | 51.722 | 68.38/ 81.22 | 5.55/ 7.2 | 51.27 | 32.51/ 32.42 | 53.32 | 0/ 63.18 | 37.23 |
| SAM($\rho$ =0.01) | 41.91 | 95.3 | 68.38/ 81.22 | 9.21/ 10.38 | 53.07 | 87.99/ 87.8 | 51.24 | 83.44/ 87.27 | 63.1 |
| SAM($\rho$ =0.005) | 58.79 | 81.54 | 68.38/ 81.22 | 13.52/ 16.6 | 53.79 | 88.42/ 88.15 | 92.95 | 83.84/ 87.7 | 67.91 |
| SAM（最佳） | 58.79 | 95.3 | 68.38/ 81.22 | 13.52/ 16.6 | 53.79 | 88.42/ 88.15 | 92.95 | 83.84/ 87.7 | 69.06 |
| AMSGrad | 63.78 | 96.44 | 89.71/ 92.44 | 89.98/ 90.35 | 87.36 | 90.65/ 90.35 | 94.53 | 88.59/ 91.27 | 88.79 |
| AdaSAM($\rho$ =0.01) | 69.23 | 96.22 | 89.96/ 92.84 | 88.83/ 89.07 | 87 | 90.83/ 90.41 | 94.8 | 88.67/ 91.38 | 89.1 |
| AdaSAM($\rho$ =0.005) | 68.47 | 96.22 | 89.96/ 92.82 | 91.59/ 91.22 | 73.65 | 90.75/ 90.42 | 94.73 | 88.72/ 91.46 | 88.33 |
| AdaSAM（最佳） | 69.23 | 96.22 | 89.96/ 92.84 | 91.59/ 91.22 | 87 | 90.83/ 90.42 | 94.8 | 88.72/ 91.46 | 89.52 |

学习率、训练轮数、批量大小，等等 [3]。具体而言，对于 GLUE 基准中的 RTE、STS-B 和 MRPC，我们先在 MNLI数据集上微调预训练的 RoBERTa-large模型，再在相应的单任务语料库上继续微调 RoBERTa-large-MNLI模型，以获得更好的性能；许多先前的工作也采用了这种做法 [7], [66]。所有模型均在 NVIDIA DGX SuperPOD集群上训练，其中每台机器配备 8×40 GB A100 GPU。

### *B. GLUE* 基准上的结果

表I 展示了随机梯度下降、SAM、AMSGrad 和 AdaSAM 的性能。对于 AdaSAM，我们将扰动参数的邻域大小分别设为 0.01、0.005 和 0.001 进行调节。结果表明，在 8 项任务中的 6 项上，AdaSAM 的表现优于 AMSGrad，只有 QNLI 和 QQP 除外。总体而言，其平均分比 AMSGrad 高 0.28。另一方面，表格 I 表明，在 8 项任务中的 7 项上，SAM 优于随机梯度下降，只有 RTE 除外。而且，SAM 能显著提升性能。比较表格 I 的结果可以发现，自适应学习率方法优于采用手工调整的学习率进行调参的随机梯度下降。AdaSAM 在 6 项任务上取得了最佳指标，分别是 CoLA、SST-2、MRPC、STS-B、RTE、QNLI 和 MNLI。总体而言，AdaSAM 优于其他方法。

此外，图3分别展示了训练过程中详细损失和评估指标随步数变化的收敛速度。AdaSAM 的损失曲线在所有任务中的下降速度都快于 SAM 和随机梯度下降，并且其下降速度与 AMSGrad 相近。AdaSAM 和 AMSGrad 的评估指标曲线表明，AdaSAM 优于随机梯度下降和 SAM，并且在所有任务中降低损失值的速度与 AMSGrad 一样快。

[3]由于篇幅限制，数据集和训练设置的详细信息见附录A。

### *C.* 小批量加速

在这一部分中，我们使用不同的批量大小测试性能，以验证线性加速特性。实验在 MRPC、RTE 和 CoLA 任务上进行。批量大小分别设为 4、8、16 和 32。我们按比例缩放学习率$\sqrt{N}$，这与 [67], 类似，其中 $N$ 表示批量大小。结果表明，随着批量大小增加，训练损失下降得更快；批量大小为 32 时的损失曲线所需迭代次数几乎只有批量大小为 16 时的一半。

### *D.* 消融实验

在本小节中，我们进行实验，将动量超参数 $\beta_1$ 设为 0，以评估动量加速和自适应学习率的影响。表II 显示，除 SST-2 和 RTE 外，AdaSAM 在 8 个任务中的 6 个任务上优于 AMSGrad。在表II中，我们还比较了 SGD 和 SAM；在不使用动量的情况下，SAM 在所有任务上均优于 SGD。在这种情况下，不使用动量加速方法的 AdaSAM 优于其他方法。

比较表 I 和表 II的结果后，我们发现，自适应学习率方法和动量加速都有助于提升模型的泛化能力。在没有动量项时，采用自适应

学习率的 SAM 将平均分数提高了 0.74，达到 AMSGrad 的水平。加入动量项后，AdaSAM 将平均分数提高了 0.28，达到 AMSGrad 的水平。这表明，无论是否采用动量加速，自适应方法都能提升性能；而采用动量加速时，性能最佳。我们还发现，动量加速能够提升 SAM、AMSGrad 和 AdaSAM 的性能。

## V. 结论

在本文中，我们研究了在随机非凸设定下，采用自适应学习率和动量加速的尖锐度感知最小化优化器（称为 AdaSAM）的收敛速率。据我们所知，我们首次给出了 AdaSAM 非平凡的 $\mathcal{O}(1/\sqrt{bT})$ 收敛速率，其相对于小批量大小 $b$具有线性加速性质。我们在若干自然语言处理任务上进行了广泛实验，验证了 AdaSAM 相较于 AMSGrad 和 SAM 优化器能够取得更优性能。未来工作包括将 AdaSAM扩展到分布式设置，并降低两次梯度反向传播的成本。

## 参考文献


[1] P. Foret、A. Kleiner、H. Mobahi 和 B. Neyshabur，“用于高效提升泛化能力的尖锐度感知最小化”，载于国际学习表征会议，2021 1, 2 。K. He、X. Zhang、S. Ren

[2] 和 J. Sun，“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，2016， 770 778 1 第 – 页。P. Mi、L. Shen、

[3] T. Ren、Y. Zhou、X. Sun、R. Ji 和 D. Tao，“增强尖锐度感知最小化：一种稀疏化扰动方法”，载于神经信息处理系统进展。1，2[4] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly、J. Uszkoreit 和 N. Houlsby，“一张图像胜过16×16个词：用于大规模图像识别的 Transformer”，载于国际学习表征会议，2021 1 。X. Chen、C.-J. Hsieh 和 B. Gong，“在无需预

[5] 训练或强数据增强的情况下，视觉Transformer何时优于残差网络”，载于国际学习表征会议，2022 1 。J.

[6] Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向Transformer预训练”，arXiv 预印本 arXiv: *1810 04805*, 2018 1 。

[7] P. He、X. Liu、J. Gao 和 W. Chen，“DeBERTa：采用解耦注意力的解码增强型 BERT”，载于 ICLR，2020 1, 7 。Q.

[8] Zhong、L. Ding、L. Shen、P. Mi、J. Liu、B. Du 和 D. Tao，“利用 Fisher掩码改进尖锐度感知最小化，以提升语言模型的泛化能力”，arXiv 预印本 arXiv: *2210 05497*, 2022 1, 2 。D. P. King

[9] ma 和 J. Ba，“Adam：一种随机优化方法”，载于 ICLR（海报），2015年。 [在线]。可获取于：http: //arxiv.org/abs/ 1412 6980 1, 3 。S. J. Reddi、

[10] S. Kale 和 S. Kumar，“关于 Adam 及其扩展的收敛性”，载于国际学习表征会议，2018 1, 3 。H. Iiduka，“用于训练深度神经网络的自适应学习率优化算法的适

[11] 当学习率”，IEEE Trans. Cybern.，第 52， 12， 13 250 13 261, 2022 1 卷，第 – 页。J. Duchi、E. Hazan 和

[12] Y. Singer，“用于在线学习和随机优化的自适应次梯度方法”。机器学习研究期刊，第 12， 7, 2011 1, 3 卷，第 。S. Ruder，“ 梯度下降优化

[13] 算法概述”，arXiv 预印本 arXiv: *1609 04747*, 2016 1 。L.

[14] Liao、L. Shen、J. Duan、M. Kolar 和 D. Tao，“用于随机凸-凹优化的局部 Adagrad 型算法”，Machine Learning，第1–20页，2022年。1

[15] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar 和 S. Sra，“自适应方法为何适用于注意力模型？”神经信息处理系统进展，第33卷，第15 383–15 393页，2020年。1[16] J. Zhuang, B. Gong, L. Yuan, Y. Cui, H. Adam, N. C. Dvornek, sekhar tatikonda, J. s Duncan 和 T. Liu，“代理间隙最小化改进了尖锐性感知训练”，载于国际学习表征会议，2022 1, 2 。J. Kwon, J. Kim, H. Park 和 I. K.

[17] Choi，“ASAM：面向深度神经网络尺度不变学习的自适应尖锐度感知最小化”，载于国际机器学习会议。PMLR， 2021， 5905 5914 1, 2 页码 – 。N.

[18] Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever 和 R. Salakhutdinov，“Dropout：一种防止神经网络过拟合的简单方法”，机器学习研究期刊，第15卷，1958，2014 1，1929 2期，页码 – 。[19] I. Loshchilov 和 F. Hutter，“解耦权重衰减正则化”，arXiv 预印本 arXiv: *1711 05101*, 2017 2 。。Z. Li, H. Zhao, Y. Guo, Z.

[20] Yang 和 S. Xie，“对数正则化卷积变换学习的加速及其收敛保证”，IEEE 控制论汇刊，第 52， 10， 10 785 10 799, 2022 2 卷，期，

[21] 页码 – 。Y. Lu, Z. Zhang, G. Lu, Y. Zhou, J. Li 和 D. Zhang，“Addi-reg：一种针对卷积神经网络、具有更佳泛化-优化权衡的正则化方法”，IEEE 控制论汇刊，第52卷，第10期，页码 10 827 10 842, 2022 2 – 。Y. Jiang, B.

[22] Neyshabur, H. Mobahi, D. Krishnan 和 S. Bengio，“绝妙的泛化度量及其寻觅之处”，载于国际学习表征会议，2020 2 。N. S. Keskar, D.

[23] Mudigere, J. Nocedal, M. Smelyanskiy 和 P. T. P. Tang，“关于深度学习的大批量训练：泛化间隙与尖锐极小值”，载于第5届国际学习表征会议，ICLR *2017* 2017 2 。OpenReview.net， 。H.

[24] He, G. Huang 和 Y. Yuan，“非对称谷：超越尖锐和扁平的局部极小值”，神经信息处理系统进展，第 32, 2019 2 卷。J. Du, H. Yan, J. Feng, J. T.

[25] Zhou, L. Zhen, R. S. M. Goh 和 V. Tan，“用于改进神经网络训练的高效尖锐度感知最小化”，载于国际学习表征会议，2022年。2[26] Y. Liu, S. Mai, X. Chen, C.-J. Hsieh 和 Y. You，“迈向高效且可扩展的尖锐度感知最小化”，载于 IEEE/CVF 计算机视觉与模式识别会议论文集（CVPR），6月 2022， 12 360 12 370 2 页码 – 。Z. Qu, X. Li, R. Duan, Y.

[27] Liu, B. Tang 和 Z. Lu，“通过尖锐度感知最小化实现广义联邦学习”，载于国际机器学习会议。PMLR， 2022， 18 250 18 280 2 页码 – 。

[28] Y. Sun, L. Shen, T. Huang, L. Ding 和 D. Tao，“Fedspeed：更大的本地间隔、更少的通信轮次和更高的泛化准确率”，载于国际学习表征会议。

[29] M. Andriushchenko 和 N. Flammarion，“迈向对尖锐度感知最小化的理解”，载于国际机器学习会议。PMLR， 2022， 639 668 2, 4

[30] 页码 – 。M. D. Zeiler，“Adadelta：一种自适应学习率方法”，arXiv 预印本 arXiv: *1212 5701*, 2012 3 。。D. Zhou, J. Chen, Y. Cao, Y.

[31] Tang, Z. Yang 和 Q. Gu，“非凸优化中自适应梯度方法的收敛性”，arXiv 预印本 arXiv: *1808 05671*, 2018 3, 4 。。X. Chen, S. Liu, R.

[32] Sun 和 M. Hong，“非凸优化中一类 Adam 类算法的收敛性”，载于国际学习表征会议，2019 3, 4 。M. Zaheer,

[33] S. Reddi, D. Sachan, S. Kale 和 S. Kumar，“非凸优化的自适应方法”，神经信息处理系统进展，第 31, 2018 3 卷。R. Ward, X.

[34] Wu 和 L. Bottou，“Adagrad 步长：非凸景观上的尖锐收敛”，载于国际机器学习会议。PMLR， 2019， 6677 6686 3 页码 – 。A.

[35] Defossez, L. Bottou, F. Bach 和 N. Usunier，“Adam 和 Adagrad 的收敛性”，CoRR，第abs/2003.02395卷，2020年。 [线上]。网址：https://arxiv.org/abs/ 2003 02395 3 。F. Zou,

[36] L. Shen, Z. Jie, W. Zhang 和 W. Liu，“Adam 和 RMSProp 收敛的一个充分条件”，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2019年，页码11 127–11 135。3[37] C. Chen, L. Shen, F. Zou 和 W. Liu，“迈向实用 Adam：非凸性、收敛理论与小批量加速”，arXiv 预印本 arXiv:2101.05471，2021年。3，4

[38] C. Chen, L. Shen, H. Huang 和 W. Liu，“带误差反馈的量化 Adam”，ACM 智能系统与技术汇刊（TIST），第 12, 5, 1 26, 2021 3 卷，期，页码 – 。C. Chen, L. [39] Shen, F. Zou 和 W. Liu，“迈向实用 Adam：非凸性、收敛理论与小批量加速”，机器学习研究期刊，第 23, 1 47, 2022 3 卷，页码 – 。C. [40] Chen, L. Shen, W. Liu 和 Z.-Q. Luo，“Efficient-Adam：带复杂度分析的通信高效分布式 Adam”，arXiv 预印本 arXiv: *2205 14473*, 2022 3 。。F. Zou, L. Shen, Z. Jie, J. S [41] un 和 W. Liu，“带统一动量的加权 Adagrad”，arXiv 预印本 arXiv: *1808 03408*, 2018 3 。。H. Ii [42] duka，“训练深度神经网络的自适应学习率优化算法的适当学习率”，IEEE 控制论汇刊，第 52, 12, 13 250 13 261, 2021 3 卷，期，页码 – 。S. Sun, Z. Cao, H. Zhu 和 J. Zhao，“从机器学习视角看优化方法综述”，IEEE 控制论汇刊，第 50, 8, 3668 3681, 2019 3 卷，期，页码 – 。[43] H. Sakai 和 H. Iiduka，“黎曼自适应优化算法及其在自然语言处理中的应用”，IEEE 控制论汇刊，第 52, 8, 7328 7339, 2021 3 卷，期，页码 – 。B. T. Polyak，“加速迭代方法收敛的一些方法”，苏联计算数学与数学物理，第 4, 5, 1 17, 1964 3 卷，期，页码 – 。Y. [44] [45] Nesterov，《凸优化导论：基础课程》。Springer Science & Business Media, 2003, 87 3 卷。。[46] B. O’Donoghue 和 E. J. Candes，“加速梯度方案的自适应重启”，《计算数学基础》，第15卷，732，2015年 3，715 3期，页码 – 。[47] W. Liu, L. Chen, Y. Chen 和 W. Zhang，“通过动量梯度下降加速联邦学习”，IEEE 并行与分布式系统汇刊，第 31, 8, 1754 1766, 2020 3 卷，期，页码 – 。X. Luo, Z. Liu, S. Li, M. Shang 和 Z. Wang，“基于广义动量法的快速非负潜因子模型”，IEEE 系统、人类与控制论汇刊：系统，第51卷，第1期，页码 610 620, 2018 3 – 。M. Shang, Y. Yuan, X. Luo 和 M. Zhou，“一种 [48] [49] $\alpha$–$\beta$散度广义推荐模型，可高精度预测缺失的用户偏好”，IEEE 控制论汇刊，第52卷，第8期，页码8006–8018，2021年。3[50] T. Yang, Q. Lin 和 Z. Li，“凸优化和非凸优化中随机动量方法的统一收敛性分析”，arXiv 预印本 arXiv: *1604 03257*, 2016 3 。。S. S. Mannelli 和 P. Urbani，“范式 [51] 性高维非凸问题中基于动量的加速方法的解析研究”，载于《神经信息处理系统进展34：神经信息处理系统年会2021》，NeurIPS 2021，2021年12月6日至14日，线上，M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang 和 J. W. Vaughan 编，2021年，页码187–199。3[52] [53]X. Gao, M. G¨urb¨uzbalaban 和 L. Zhu，“非凸随机优化中随机梯度哈密顿蒙特卡洛的全局收敛：非渐近性能界与基于动量的加速”，《运筹学》，第70卷，第5期，页码2931–2947，2022年。3[54] I. Sutskever, J. Martens, G. Dahl 和 G. Hinton，“初始化和动量在深度学习中的重要性”，载于国际机器学习会议。PMLR, 2013, 1139 1147 3 页码 – 。[55] B. Can, M. Gurbuzbalaban 和 L. Zhu，“Wasserstein 距离下随机动量方法的加速线性收敛”，载于第36届国际机器学习会议论文集，ICML 2019，2019年6月9日至15日，美国加利福尼亚州长滩，机器学习研究论文集系列，K. Chaudhuri 和 R. Salakhutdinov 编，第 97 2019, 891 901 3 卷。PMLR，页码 – 。F. Huang, S. [56] Gao, J. Pei 和 H. Huang，“从小批量到极小极大优化的加速零阶和一阶动量方法”，《机器学习研究期刊》，第 23, 36 1 36 70, 2022 3 卷，页码：–：。R. [57] Bollapragada, D. Scieur 和 A. d’Aspremont，“动量算法与原始-对偶算法的非线性加速”，《数学规划》，页码 1 38, 2022 3 – 。B. O’ [58] donoghue 和 E. Candes，“加速梯度方案的自适应重启”，《计算数学基础》，第15卷，页码715–732，2015年。3[59] B. Wang, T. M. Nguyen, T. Sun, A. L. Bertozzi, R. G. Baraniuk 和 S. J. Osher，“用于加速随机梯度下降的定期重启动量”，《SIAM 影像科学期刊》，第15卷，第2期，页码738–761，2022年。3[60] B. Liu、L. Chai 和 J. Yi，“含一个和两个动量项的分布式梯度下降算法的收敛性分析”，*IEEE* 控制论汇刊，2022 3 。A. [61] Cutkosky 和 F. Orabona，“非凸随机梯度下降中的基于动量的方差缩减”，神经信息处理系统进展，卷 32, 2019 4 。F. Huang、J. Li 和 H. Huang，“ [62] Super-Adam：自适应梯度的更快通用框架”，神经信息处理系统进展，卷 34, 2021 4 。M. Li、D. G. Andersen、A. J. [63] Smola 和 K. Yu，“借助参数服务器实现通信高效的分布式机器学习”，神经信息处理系统进展，卷 27, 2014 4 。[64] M. Li、T. Zhang、Y. Chen 和 A. J. Smola，“随机优化的高效小批量训练”，发表于第*20*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，2014, 661 670 4 页码 – 。A. Wang、A. Singh、J. Michael、[65] F. Hill、O. Levy 和 S. Bowman，“GLUE：自然语言理解的多任务基准与分析平台”，发表于 *EMNLP*，2018 6 。Y. Liu、M. [66] Ott、N. Goyal、J. Du、M. Joshi、D. Chen、O. Levy、M. Lewis、L. Zettlemoyer 和 V. Stoyanov，“RoBERTa：一种稳健优化的 BERT 预训练方法”，*arXiv*，2019 7 。X. Li、B. Karimi 和 P. Li，“ [67] 关于采用梯度压缩的分布式自适应优化”，发表于 国际学习表征会议，2021年。7

在这份补充材料中，我们对本文进行进一步讨论。在附录 A 中，我们列出了详细的实验设置，例如一些超参数。在附录 B 中，我们先给出证明，再给出一些有助于证明主定理的引理。在附录 C 中，我们提供了额外的实验说明。

## 附录 A 实验设置

表 III：不同下游任务的实验设置和数据划分。值得注意的是，对于 GLUE 基准中的每项任务，我们分别提供微调期间的类别数（“classes”）、学习率（“lr”）、批量大小（“bsz”）、更新总步数（“total”）、预热更新步数（“warmup”）以及 GPU 数量（“GPUs”）。

| | MNLI | QNLI | QQP | RTE | SST-2 | MRPC | CoLA | STS-B |
|---|---|---|---|---|---|---|---|---|
| 不同下游任务的实验设置 | | | | | | | | |
| –类别 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| –lr | 1e-5 | 1e-5 | 1e-5 | 2e-5 | 1e-5 | 1e-5 | 1e-5 | 2e-5 |
| –bsz | 256 | 128 | 256 | 32 | 64 | 32 | 32 | 32 |
| –总计 | 15,484 | 8,278 | 14,453 | 1,018 | 10,467 | 1,148 | 2,668 | 1,799 |
| –预热 | 929 | 496 | 867 | 61 | 628 | 68 | 160 | 107 |
| –GPU | 4 | 4 | 8 | 2 | 2 | 2 | 2 | 2 |
| 每个数据集的数据划分 | | | | | | | | |
| 训练 | 392,720 | 104,743 | 363,870 | 2,491 | 67,350 | 5,801 | 8,551 | 5,749 |
| dev | 9,815 | 5,463 | 40,431 | 277 | 873 | 4,076 | 1,043 | 1,500 |
| test | 9,796 | 5,461 | 390,956 | 3,000 | 1,821 | 1,725 | 1,063 | 1,379 |

GLUE 基准包含 8 项任务，分别是 RTE、STS-B、CoLA、SST-2、MNLI、MRPC、QNLI 和 QQP。CoLA 是一项单句任务。每个句子都有一个标签 1 或 -1。1 表示该句子合乎语法，而 -1 表示该句子不合语法。马修斯相关系数，简称 **mcc** ，是我们使用的评估指标。STS-B 是一项相似度与释义任务。每个样本包含一对段落。人们根据两个段落之间的相似度，为样本标注 1 到 5 的分数。其指标是皮尔逊和斯皮尔曼相关系数，简称 **p/s** 。RTE 是一项推理任务。每个样本包含两个句子。如果两个句子之间存在蕴含关系，我们就将其视为正样本；否则，它们构成负样本。RTE 任务的指标是准确率，简称 **acc**。SST-2 是一项单句任务，其指标是准确率。MNLI 是一项句子级任务，包含 3 个类别，分别是蕴含、矛盾和中立。MRPC 是一项用于判断句子对是否等价的任务。QNLI 是一项问答任务。如果句子中包含问题的答案，则该样本为正样本。QQP 是一项社交问答任务，由 Quora 上的问题对组成，用于判断问题是否等价。MNLI、MRPC、QNLI 和 QQP 的指标均为准确率。

## APPENDIX B
## PROOF OF THE MAIN RESULTS

我们为 $t \geq 0$ 设置~~$z_t =$~~ $x_t + \frac{\beta_1}{1-\beta_1}(x_t - x_{t-1})$，并假设 $x_{-1} = 0$ 和 $m_{-1} = 0$。我们有

$$z_{t+1} - z_t = x_{t+1} + \frac{\beta_1}{1-\beta_1}(x_{t+1} - x_t) - x_t - \frac{\beta_1}{1-\beta_1}(x_t - x_{t-1}) \tag{21}$$

$$= \frac{1}{1-\beta_1}(x_{t+1} - x_t) - \frac{\beta_1}{1-\beta_1}(x_t - x_{t-1}) \tag{22}$$

$$= -\frac{1}{1-\beta_1}\gamma m_t \odot \eta_t + \frac{\beta_1}{1-\beta_1}(x_t - x_{t-1})\gamma m_{t-1} \odot \eta_{t-1} \tag{23}$$

$$= -\frac{1}{1-\beta_1}\gamma(\beta_1 m_{t-1} + (1-\beta_1)g_t) \odot \eta_t + \frac{\beta_1}{1-\beta_1}(x_t - x_{t-1})\gamma m_{t-1} \odot \eta_{t-1} \tag{24}$$

$$= \frac{\beta_1}{1-\beta_1}\gamma m_{t-1} \odot (\eta_{t-1} - \eta_t) - \gamma g_t \odot \eta_t \tag{25}$$

应用 L-光滑性可得

$$f(z_{t+1}) \leq f(z_t) + \langle \nabla f(z_t), z_{t+1} - z_t \rangle + \frac{L}{2}\|z_{t+1} - z_t\|^2 \tag{26}$$

我们重新整理后，得到

$$f(z_{t+1}) - f(z_t)$$
$$\leq \langle \nabla f(z_t), z_{t+1} - z_t \rangle + \frac{L}{2}\|z_{t+1} - z_t\|^2 \tag{27}$$

$$= \langle \nabla f(z_t), \frac{\gamma\beta_1}{1-\beta_1} m_{t-1} \odot (\eta_{t-1} - \eta_t) \rangle + \langle \nabla f(z_t), -\gamma g_t \odot \eta_t \rangle + \frac{L}{2} \|z_{t+1} - z_t\|^2 \tag{28}$$

$$\begin{aligned} &= \langle \nabla f(z_t), \frac{\gamma\beta_1}{1-\beta_1} m_{t-1} \odot (\eta_{t-1} - \eta_t) \rangle + \frac{L}{2} \|z_{t+1} - z_t\|^2 \\ &+ \langle \nabla f(z_t), \frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot (\eta_{t-1} - \eta_t) \rangle \\ &+ \langle \nabla f(z_t), -\frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot \eta_{t-1} \rangle \end{aligned} \tag{29}$$

$$\begin{aligned} &= \langle \nabla f(z_t), \frac{\gamma\beta_1}{1-\beta_1} m_{t-1} \odot (\eta_{t-1} - \eta_t) \rangle + \frac{L}{2} \|z_{t+1} - z_t\|^2 \\ &+ \langle \nabla f(z_t), \frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot (\eta_{t-1} - \eta_t) \rangle \\ &+ \langle \nabla f(z_t) - \nabla f(x_t), -\frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot \eta_{t-1} \rangle \\ &+ \langle \nabla f(x_t), -\frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot \eta_{t-1} \rangle \end{aligned} \tag{30}$$

$$\begin{aligned} &= \langle \nabla f(z_t), \frac{\gamma\beta_1}{1-\beta_1} m_{t-1} \odot (\eta_{t-1} - \eta_t) \rangle + \frac{L}{2} \|z_{t+1} - z_t\|^2 \\ &+ \langle \nabla f(z_t), \frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot (\eta_{t-1} - \eta_t) \rangle \\ &+ \langle \nabla f(z_t) - \nabla f(x_t), -\frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{\sum \nabla f_i(x_t)}{\|\sum \nabla f_i(x_t)\|}) \odot \eta_{t-1} \rangle \\ &+ \langle \nabla f(x_t), \frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) \odot \eta_{t-1} - \frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot \eta_{t-1} \rangle \\ &+ \langle \nabla f(x_t), -\frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) \odot \eta_{t-1} \rangle. \end{aligned} \tag{31}$$

由引理 5、引理6、引理7可得

$$\langle \nabla f(z_t), \frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot (\eta_{t-1} - \eta_t) \rangle \leq \gamma_t G^2 \|\eta_{t-1} - \eta_t\|_1, \tag{32}$$

$$\langle \nabla f(z_t), \frac{\gamma\beta_1}{1-\beta_1} m_{t-1} \odot (\eta_{t-1} - \eta_t) \rangle \leq \frac{\gamma\beta_1}{1-\beta_1} G^2 \|\eta_{t-1} - \eta_t\|_1, \tag{33}$$

$$\begin{aligned} &\langle \nabla f(x_t), \frac{\eta_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) \odot \eta_{t-1} - \frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot \eta_{t-1} \rangle \\ &\leq \frac{\gamma_t}{2\mu^2} \|\nabla f(x_t) \odot \sqrt{\eta_{t-1}}\|^2 + \frac{2\mu^2 \gamma_t L^2 \rho_t^2}{\epsilon}. \end{aligned} \tag{34}$$

取条件期望，得到

$$\mathbb{E} f(z_{t+1}) - f(z_t) \tag{35}$$

$$\begin{aligned} &\leq \mathbb{E} \langle \nabla f(x_t), -\frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) \odot \eta_{t-1} \rangle + \frac{L}{2} \mathbb{E} \|z_{t+1} - z_t\|^2 \\ &+ \frac{\gamma_t}{2\mu^2} \|\nabla f(x_t) \odot \sqrt{\eta_{t-1}}\|^2 + \frac{2\mu^2 \gamma_t L^2 \rho_t^2}{\epsilon} + \frac{\gamma}{1-\beta_1} G^2 \|\eta_{t-1} - \eta_t\|_1 \\ &+ \mathbb{E} \langle \nabla f(z_t) - \nabla f(x_t), -\frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot \eta_{t-1} \rangle \end{aligned} \tag{36}$$

其中 $\mu > 0$ 有待确定。

对于项

$$\mathbb{E} \langle \nabla f(x_t), -\frac{\gamma_t}{b} \sum_{i \in B} \nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) \odot \eta_{t-1} \rangle, \tag{37}$$

该项

$$\frac{L}{2} \mathbb{E} \|z_{t+1} - z_t\|^2, \tag{38}$$

以及该项

$$\mathbb{E}\langle \nabla f(z_t) - \nabla f(x_t), -\frac{\gamma_t}{b}\sum_{i\in B}\nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot \eta_{t-1}\rangle, \tag{39}$$

我们引入引理8、引理10 和引理9。我们对整个处理过程取期望，得到

$$\begin{aligned}
&\mathbb{E}f(z_{t+1}) - \mathbb{E}f(z_t) \\
&\leq \frac{\gamma_t}{2\mu^2}\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2 + \frac{2\mu^2\gamma_t L^2\rho_t^2}{\epsilon} + \frac{\gamma}{1-\beta_1}G^2\mathbb{E}\|\eta_{t-1}-\eta_t\|_1 \\
&- \gamma_t\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2 + \mathbb{E}\frac{\gamma_t}{2\alpha^2}\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2 + \frac{\gamma_t\alpha^2L^2\rho^2}{2\epsilon} + \frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \\
&+ \gamma_t^2 L(3\frac{1+\beta}{\beta\epsilon}(\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2 + \frac{L\rho_t^2}{\epsilon} + \frac{\sigma^2}{b\epsilon}) + (1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2) \\
&+ \frac{\gamma^3L^2\beta_1^2}{2\epsilon(1-\beta_1)^2}(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2})\frac{dG_\infty^2}{\epsilon^2} + \frac{\gamma\lambda_1^2}{2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2 + \frac{\gamma L^2\rho_t^2}{2\epsilon}(\lambda_2^2+4\lambda_3^2) \qquad (40) \\
&= -\gamma_t(1 - \frac{1}{2\mu^2} - \frac{1}{2\alpha^2} - \frac{3\gamma L(1+\beta)}{\beta\epsilon} - \frac{\lambda_1^2}{2})\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2 + \frac{2\mu^2\gamma_t L^2\rho_t^2}{\epsilon} + \frac{\gamma}{1-\beta_1}G^2\mathbb{E}\|\eta_{t-1}-\eta_t\|_1 \\
&+ \frac{\gamma_t\alpha^2L^2\rho^2}{2\epsilon} + \frac{3\gamma_t^2L(1+\beta)}{\beta\epsilon}(\frac{L\rho_t^2}{\epsilon}+\frac{\sigma^2}{b\epsilon}) + \gamma_t^2LG^2((\frac{\beta_1}{1-\beta_1})^2+1+\beta)\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \\
&+ \frac{\gamma^3L^2\beta_1^2}{2\epsilon(1-\beta_1)^2}(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2})\frac{dG_\infty^2}{\epsilon^2} + \frac{\gamma L^2\rho_t^2}{2\epsilon}(\lambda_2^2+4\lambda_3^2). \qquad (41)
\end{aligned}$$

我们设置 $\mu^2=\alpha^2=8$, $\beta=3$, $\lambda_1^2=\frac{1}{4}$, $\lambda_2^2=\lambda_3^2=1$ ， 并选择 $\frac{2\gamma_t L}{\epsilon}\leq\frac{1}{8}$。因此得到

$$\begin{aligned}
&\mathbb{E}f(x_{t+1}) - \mathbb{E}f(x_t) \\
&\leq -\frac{\gamma_t}{2}\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2 + \frac{16\gamma_t L^2\rho_t^2}{\epsilon} + \frac{\gamma}{1-\beta_1}G^2\mathbb{E}\|\eta_{t-1}-\eta_t\|_1 \\
&+ \frac{4\gamma_t L^2\rho^2}{\epsilon} + \frac{4\gamma_t^2L}{\epsilon}(\frac{L\rho_t^2}{\epsilon}+\frac{\sigma^2}{b\epsilon}) + (4+(\frac{\beta_1}{1-\beta_1})^2)\gamma_t^2LG^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \\
&+ \frac{3\gamma^3L^2\beta_1^2}{\epsilon(1-\beta_1)^2}\frac{dG_\infty^2}{\epsilon^2} + \frac{5\gamma L^2\rho_t^2}{2\epsilon} \qquad (42)
\end{aligned}$$

我们重新整理后， $\eta_t$ 有界。我们有

$$\begin{aligned}
&\frac{\gamma_t}{2G}\mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{\gamma_t}{2}\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2 \qquad (43) \\
&\leq -\mathbb{E}f(x_{t+1}) + \mathbb{E}f(x_t) + \frac{45\gamma_t L^2\rho_t^2}{2\epsilon} + \frac{\gamma}{1-\beta_1}G^2\mathbb{E}\|\eta_{t-1}-\eta_t\|_1 \\
&+ \frac{4\gamma_t^2L}{\epsilon}(\frac{L\rho_t^2}{\epsilon}+\frac{\sigma^2}{b\epsilon}) + (4+(\frac{\beta_1}{1-\beta_1})^2)\gamma_t^2LG^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 + \frac{3\gamma^3L^2\beta_1^2}{(1-\beta_1)^2}\frac{dG_\infty^2}{\epsilon^3}. \qquad (44)
\end{aligned}$$

我们将其从 $t=0$ 总结到 $t=T-1$，并假设 $\gamma_t$ 为常数，得到

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|^2 \leq 2G\frac{\mathbb{E}f(x_0)-\mathbb{E}f(x_{t+1})}{\gamma_t T} + \frac{45GL^2\rho_t^2}{\epsilon} + \frac{2G^3}{(1-\beta_1)T}\mathbb{E}\sum_{t=0}^{T-1}\|\eta_{t-1}-\eta_t\|_1 \\
&+ \frac{8G\gamma_t L}{\epsilon}(\frac{L\rho_t^2}{\epsilon}+\frac{\sigma^2}{b\epsilon}) + \frac{2(4+(\frac{\beta_1}{1-\beta_1})^2)\gamma_t LG^3}{T}\mathbb{E}\sum_{t=0}^{T-1}\|\eta_t-\eta_{t-1}\|^2 + \frac{6\gamma^2L^2\beta_1^2}{(1-\beta_1)^2}\frac{dG^3}{\epsilon^3} \qquad (45) \\
&\leq \frac{2G(f(x_0)-f^*)}{\gamma_t T} + \frac{45GL^2\rho_t^2}{\epsilon} + \frac{2G^3}{(1-\beta_1)T}d(\frac{1}{\epsilon}-\frac{1}{G}) + \frac{8G\gamma_t L}{\epsilon}(\frac{L\rho_t^2}{\epsilon}+\frac{\sigma^2}{b\epsilon}) \\
&+ \frac{2(4+(\frac{\beta_1}{1-\beta_1})^2)\gamma_t LG^3}{T}d(\epsilon^{-2}-G^{-2}) + \frac{6\gamma^2L^2\beta_1^2}{(1-\beta_1)^2}\frac{dG^3}{\epsilon^3} \qquad (46) \\
&= \frac{2G(f(x_0)-f^*)}{\gamma_t T} + \frac{8G\gamma_t L}{\epsilon}\frac{\sigma^2}{b\epsilon} + \frac{45GL^2\rho_t^2}{\epsilon} + \frac{2G^3}{(1-\beta_1)T}d(\frac{1}{\epsilon}-\frac{1}{G}) + \frac{8G\gamma_t L}{\epsilon}\frac{L\rho_t^2}{\epsilon} \\
&+ \frac{2(4+(\frac{\beta_1}{1-\beta_1})^2)\gamma_t LG^3}{T}d(\epsilon^{-2}-G^{-2}) + \frac{6\gamma^2L^2\beta_1^2}{(1-\beta_1)^2}\frac{dG^3}{\epsilon^3}. \qquad (47)
\end{aligned}$$

*A.* 技术引理

**引理 1.** 给定两个向量 $a$、 $b \in \mathbb{R}^d$，对于参数 $\lambda$、 $\forall \lambda \in$，我们有 $\langle a, b\rangle \leq \frac{\lambda^2}{2}\|a\|^2 + \frac{1}{2\lambda^2}\|b\|^2$ $(1, +\infty)$。

证明。

$$RHS = \frac{\lambda^2}{2}\sum_{j=1}^{d}(a)_j^2 + \frac{1}{2\lambda^2}\sum_{j=1}^{d}(b)_j^2 \geq \sum_{j=1}^{d} 2\sqrt{\frac{\lambda^2}{2}(a)_j^2 \times \frac{1}{2\lambda^2}(b)_j^2} = \sum_{j=1}^{d}|(a)_j| \times |(b)_j| \geq LHS. \tag{48}$$

□

**引理 2.** 对于任意向量 $x$、$y \in \mathbb{R}^d$，有

$$\|x \odot y\|^2 \leq \|x\|^2 \times \|y\|_\infty^2 \leq \|x\|^2 \times \|y\|^2. \tag{49}$$

证明. 第一不等式可由 $\sum_{i=1}^{d}(x_i^2 y_i^2) \leq \sum_{i=1}^{d}(x_i^2\|y\|_\infty^2)$ 得出。第二不等式由$\|y\|_\infty^2 \leq \|y\|^2$得出。

□

**引理 3.** $\eta$ 有界，即 $\frac{1}{G_\infty} \leq (\eta_t)_j \leq \frac{1}{\epsilon}$。

证明. 由于梯度由 $G$ 和 $(\eta_t)^j = \frac{1}{\sqrt{(\hat{v}_t)^j}}$ 有界。根据更新规则，有 $\frac{1}{G_\infty} \leq (\eta_t)_j \leq \frac{1}{\epsilon}$。 □

**引理 4**。对于算法中定义的项，有

$$\frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}\|\eta_{t-1} - \eta_t\|^1 \leq \frac{d}{T}(\frac{1}{\epsilon} - \frac{1}{G}) \tag{50}$$

证明。$(\eta_t)_i$，随着 t 的增加， $\eta_t$ 的第 i 维递减。因此，有

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}\|\eta_{t-1} - \eta_t\|^1 = \mathbb{E}\frac{1}{T}\sum_{i=1}^{d}\sum_{t=0}^{T-1}|(\eta_{t-1})_i - (\eta_t)_i| \\
&\leq \mathbb{E}\frac{1}{T}\sum_{i=1}^{d}((\eta_{-1})_i - (\eta_{T-1})_i) \leq \mathbb{E}\frac{1}{T}\sum_{i=1}^{d}(\frac{1}{\epsilon} - \frac{1}{G}) = \frac{d}{T}(\frac{1}{\epsilon} - \frac{1}{G})
\end{aligned} \tag{51}$$

□

**引理 5**。对于算法中定义的项，有

$$\langle \nabla f(z_t), \frac{\gamma_t}{b}\sum_{i\in B}\nabla f_i(x_t + \rho_t\frac{s_t}{\|s_t\|}) \odot (\eta_{t-1} - \eta_t)\rangle \leq \gamma_t G^2\|\eta_{t-1} - \eta_t\|_1 \tag{52}$$

证明。

$$\begin{aligned}
&\langle \nabla f(z_t), \frac{\gamma_t}{b}\sum_{i\in B}\nabla f_i(x_t + \rho_t\frac{s_t}{\|s_t\|}) \odot (\eta_{t-1} - \eta_t)\rangle \\
&\leq \gamma_t\sum_{j=1}^{d}|(\nabla f(z_t))_{(j)}| \times |(\frac{1}{b}\sum_{i\in B}\nabla f_i(x_t + \rho_t\frac{\sum\nabla f_i(x_t)}{\|\sum\nabla f_i(x_t)\|}) \odot (\eta_{t-1} - \eta_t))_{(j)}| && (53)\\
&\leq \gamma_t G\sum_{j=1}^{d}|((\frac{1}{b}\sum_{i\in B}\nabla f_i(x_t + \rho_t\frac{\sum\nabla f_i(x_t)}{\|\sum\nabla f_i(x_t)\|}) \odot (\eta_{t-1} - \eta_t))_{(j)}| && (54)\\
&\leq \frac{\gamma_t G}{b}\sum_{j=1}^{d}\sum_{i\in B}|((\nabla f_i(x_t + \rho_t\frac{\sum\nabla f_i(x_t)}{\|\sum\nabla f_i(x_t)\|}) \odot (\eta_{t-1} - \eta_t))_{(j)}| && (55)\\
&= \frac{\gamma_t G}{b}\sum_{j=1}^{d}\sum_{i\in B}|(\nabla f_i(x_t + \rho_t\frac{\sum\nabla f_i(x_t)}{\|\sum\nabla f_i(x_t)\|})_{(j)} \times (\eta_{t-1} - \eta_t)_{(j)}| && (56)\\
&\leq \frac{\gamma_t G^2}{b}\sum_{j=1}^{d}\sum_{i\in B}|(\eta_{t-1} - \eta_t)_{(j)}| && (57)\\
&= \gamma_t G^2\|\eta_{t-1} - \eta_t\|_1 && (58)
\end{aligned}$$

□

**引理6**。 对于算法中定义的项，我们有

$$\langle \nabla f(z_t), \frac{\gamma\beta_1}{1-\beta_1} m_{t-1} \odot (\eta_{t-1} - \eta_t) \rangle \leq \frac{\gamma\beta_1}{1-\beta_1} G^2 \|\eta_{t-1} - \eta_t\|_1 \tag{59}$$

证明。

$$\begin{aligned}
&\langle \nabla f(z_t), \frac{\gamma\beta_1}{1-\beta_1} m_{t-1} \odot (\eta_{t-1} - \eta_t) \rangle \\
&\leq \frac{\gamma\beta_1}{1-\beta_1} \sum_{j=1}^{d} |(\nabla f(z_t))_{(j)}| \times |(m_{t-1} \odot (\eta_{t-1} - \eta_t))_{(j)}| && (60)\\
&\leq \frac{\gamma\beta_1}{1-\beta_1} G \sum_{j=1}^{d} |(m_{t-1} \odot (\eta_{t-1} - \eta_t))_{(j)}| && (61)\\
&= \frac{\gamma\beta_1}{1-\beta_1} \sum_{j=1}^{d} |(m_{t-1})_{(j)} \times (\eta_{t-1} - \eta_t)_{(j)}| && (62)\\
&\leq \frac{\gamma\beta_1}{1-\beta_1} G^2 \sum_{j=1}^{d} |(\eta_{t-1} - \eta_t)_{(j)}| && (63)\\
&= \frac{\gamma\beta_1}{1-\beta_1} G^2 \|\eta_{t-1} - \eta_t\|_1 && (64)
\end{aligned}$$

□

**引理7**。 对于算法中定义的项，我们有

$$\begin{aligned}
&\langle \nabla f(x_t), \frac{\gamma_t}{b} \sum_{i\in B} \nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) \odot \eta_{t-1} - \frac{\gamma_t}{b} \sum_{i\in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot \eta_{t-1} \rangle \\
&\leq \frac{\gamma_t}{2\mu^2} \|\nabla f(x_t) \odot \sqrt{\eta_{t-1}}\|^2 + \frac{2\mu^2\gamma_t L^2 \rho_t^2}{\epsilon}. && (65)
\end{aligned}$$

证明。

$$\begin{aligned}
&\langle \nabla f(x_t), \frac{\gamma_t}{b} \sum_{i\in B} \nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) \odot \eta_{t-1} - \frac{\gamma_t}{b} \sum_{i\in B} \nabla f_i(x_t + \rho_t \frac{s_t}{\|s_t\|}) \odot \eta_{t-1} \rangle \\
&= \langle \nabla f(x_t) \odot \sqrt{\eta_{t-1}}, \frac{\gamma_t}{b} \sum_{i\in B} (\nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) - \nabla f_i(x_t + \rho_t \frac{\sum_{i\in B} \nabla f_i(x_t)}{\|\sum_{i\in B} \nabla f_i(x_t)\|})) \odot \sqrt{\eta_{t-1}} \rangle && (66)\\
&\leq \frac{\mu^2\gamma_t}{2b^2} \| \sum (\nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) - \nabla f_i(x_t + \rho_t \frac{\sum_{i\in B} \nabla f_i(x_t)}{\|\sum_{i\in B} \nabla f_i(x_t)\|})) \odot \sqrt{\eta_{t-1}} \|^2 \\
&+ \frac{\gamma_t}{2\mu^2} \|\nabla f(x_t) \odot \sqrt{\eta_{t-1}}\|^2 && (67)\\
&\leq + \frac{\mu^2\gamma_t}{2b} \sum \|\nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) - \nabla f_i(x_t + \rho_t \frac{\sum_{i\in B} \nabla f_i(x_t)}{\|\sum_{i\in B} \nabla f_i(x_t)\|}) \odot \sqrt{\eta_{t-1}} \|^2 \\
&+ \frac{\gamma_t}{2\mu^2} \|\nabla f(x_t) \odot \sqrt{\eta_{t-1}}\|^2 && (68)\\
&\leq + \frac{\mu^2\gamma_t}{2b} \sum \|\nabla f_i(x_t + \rho_t \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}) - \nabla f_i(x_t + \rho_t \frac{\sum_{i\in B} \nabla f_i(x_t)}{\|\sum_{i\in B} \nabla f_i(x_t)\|}) \|^2 \times \|\sqrt{\eta_{t-1}}\|_\infty^2 \\
&+ \frac{\gamma_t}{2\mu^2} \|\nabla f(x_t) \odot \sqrt{\eta_{t-1}}\|^2 && (69)\\
&\leq \frac{\gamma_t}{2\mu^2} \|\nabla f(x_t) \odot \sqrt{\eta_{t-1}}\|^2 + \frac{\mu^2\gamma_t L^2 \rho_t^2}{2b\epsilon} \sum \| \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|} - \frac{\sum_{i\in B} \nabla f_i(x_t)}{\|\sum_{i\in B} \nabla f_i(x_t)\|} \|^2 && (70)\\
&\leq \frac{\gamma_t}{2\mu^2} \|\nabla f(x_t) \odot \sqrt{\eta_{t-1}}\|^2 + \frac{2\mu^2\gamma_t L^2 \rho_t^2}{\epsilon}. && (71)
\end{aligned}$$

□

**引理 8**。 对于算法中定义的项，我们有

$$\mathbb{E}\langle\nabla f(x_t), -\frac{\gamma_t}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})\odot\eta_{t-1}\rangle$$
$$\leq -\gamma_t\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\mathbb{E}\frac{\gamma_t}{2\alpha^2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{\gamma_t\alpha^2L^2\rho_t^2}{2\epsilon} \tag{72}$$

证明。

$$\mathbb{E}\langle\nabla f(x_t), -\frac{\gamma_t}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})\odot\eta_{t-1}\rangle$$
$$= -\gamma_t\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\mathbb{E}\langle\nabla f(x_t), \frac{\gamma_t}{b}\sum_{i\in B}(\nabla f(x_t)-\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}))\odot\eta_{t-1}\rangle \tag{73}$$
$$= -\gamma_t\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\mathbb{E}\langle\nabla f(x_t), \frac{\gamma_t}{b}\sum_{i\in B}(\nabla f_i(x_t)-\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}))\odot\eta_{t-1}\rangle \tag{74}$$
$$\leq -\gamma_t\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\mathbb{E}\frac{\gamma_t}{2\alpha^2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2$$
$$+\frac{\gamma_t\alpha^2}{2}\mathbb{E}\|\frac{1}{b}\sum_{i\in B}(\nabla f_i(x_t)-\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}))\odot\sqrt{\eta_{t-1}}\|^2 \tag{75}$$
$$\leq -\gamma_t\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\mathbb{E}\frac{\gamma_t}{2\alpha^2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2$$
$$+\frac{\gamma_t\alpha^2}{2\epsilon}\mathbb{E}\|\frac{1}{b}\sum_{i\in B}(\nabla f_i(x_t)-\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}))\|^2 \tag{76}$$
$$\leq -\gamma_t\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\mathbb{E}\frac{\gamma_t}{2\alpha^2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2$$
$$+\frac{\gamma_t\alpha^2}{2b\epsilon}\mathbb{E}\sum_{i\in B}\|(\nabla f_i(x_t)-\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}))\|^2 \tag{77}$$
$$\leq -\gamma_t\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\mathbb{E}\frac{\gamma_t}{2\alpha^2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{\gamma_t\alpha^2L^2\rho_t^2}{2b\epsilon}\mathbb{E}\sum_{i\in B}\|\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}\|^2 \tag{78}$$
$$= -\gamma_t\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\mathbb{E}\frac{\gamma_t}{2\alpha^2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{\gamma_t\alpha^2L^2\rho_t^2}{2\epsilon} \tag{79}$$

□

**引理 9.** 对于算法中定义的项，我们有

$$\mathbb{E}\langle\nabla f(z_t)-\nabla f(x_t), -\frac{\gamma_t}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|})\odot\eta_{t-1}\rangle$$
$$\leq\frac{\gamma^3L^2\beta_1^2}{2\epsilon(1-\beta_1)^2}(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2})\frac{dG_\infty^2}{\epsilon^2}+\frac{\gamma\lambda_1^2}{2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{\gamma L^2\rho_t^2}{2\epsilon}(\lambda_2^2+4\lambda_3^2). \tag{80}$$

证明。

$$\mathbb{E}\langle\nabla f(z_t)-\nabla f(x_t), -\frac{\gamma_t}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|})\odot\eta_{t-1}\rangle \tag{81}$$
$$=\gamma\mathbb{E}\langle(\nabla f(x_t)-\nabla f(z_t))\odot\sqrt{\eta_{t-1}}, \frac{1}{b}\sum_{i\in B}\nabla f_i(x_t+\rho_t\frac{\sum_{i\in B}\nabla f_i(x_t)}{\|\sum_{i\in B}\nabla f_i(x_t)\|})\odot\sqrt{\eta_{t-1}}\rangle \tag{82}$$
$$=\gamma\mathbb{E}\langle(\nabla f(x_t)-\nabla f(z_t))\odot\sqrt{\eta_{t-1}}, \nabla f(x_t)\odot\sqrt{\eta_{t-1}}\rangle$$
$$+\gamma\mathbb{E}\langle(\nabla f(x_t)-\nabla f(z_t))\odot\sqrt{\eta_{t-1}}, \frac{1}{b}\sum_{i\in B}(\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})-\nabla f_i(x_t))\odot\sqrt{\eta_{t-1}}\rangle$$
$$+\gamma\mathbb{E}\langle(\nabla f(x_t)-\nabla f(z_t))\odot\sqrt{\eta_{t-1}}, \frac{1}{b}\sum_{i\in B}(\nabla f_i(x_t+\rho_t\frac{\sum_{i\in B}\nabla f_i(x_t)}{\|\sum_{i\in B}\nabla f_i(x_t)\|})-\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})\odot\sqrt{\eta_{t-1}}\rangle \tag{83}$$
$$\leq\frac{\gamma}{2}(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2})\mathbb{E}\|(\nabla f(x_t)-\nabla f(z_t))\odot\sqrt{\eta_{t-1}}\|^2+\frac{\gamma\lambda_1^2}{2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2$$
$$+\frac{\gamma\lambda_2^2}{2}\mathbb{E}\|\frac{1}{b}\sum_{i\in B}(\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})-\nabla f_i(x_t))\odot\sqrt{\eta_{t-1}}\|^2$$

$$+\frac{\gamma\lambda_3^2}{2}\mathbb{E}\|\frac{1}{b}\sum_{i\in B}(\nabla f_i(x_t+\rho_t\frac{\sum_{i\in B}\nabla f_i(x_t)}{\|\sum_{i\in B}\nabla f_i(x_t)\|})-\nabla f_i(x_t+\rho_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|})\odot\sqrt{\eta_{t-1}}\|^2 \tag{84}$$

$$\leq\frac{\gamma}{2}(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2})\mathbb{E}\|(\nabla f(x_t)-\nabla f(z_t))\odot\sqrt{\eta_{t-1}}\|^2+\frac{\gamma\lambda_1^2}{2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2$$
$$+\frac{\gamma\lambda_2^2L^2\rho_t^2}{2\epsilon}+\frac{2\lambda_3^2\gamma L^2\rho_t^2}{\epsilon} \tag{85}$$

$$\leq\frac{\gamma L^2}{2\epsilon}(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2})\mathbb{E}\|z_t-x_t\|^2+\frac{\gamma\lambda_1^2}{2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2$$
$$+\frac{\gamma\lambda_2^2L^2\rho_t^2}{2\epsilon}+\frac{2\lambda_3^2\gamma L^2\rho_t^2}{\epsilon} \tag{86}$$

$$=\frac{\gamma^3L^2\beta_1^2}{2\epsilon(1-\beta_1)^2}(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2})\|m_{t-1}\odot\eta t-1\|^2+\frac{\gamma\lambda_1^2}{2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2$$
$$+\frac{\gamma\lambda_2^2L^2\rho_t^2}{2\epsilon}+\frac{2\lambda_3^2\gamma L^2\rho_t^2}{\epsilon} \tag{87}$$

$$\leq\frac{\gamma^3L^2\beta_1^2}{2\epsilon(1-\beta_1)^2}(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2})\frac{dG_\infty^2}{\epsilon^2}+\frac{\gamma\lambda_1^2}{2}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{\gamma L^2\rho_t^2}{2\epsilon}(\lambda_2^2+4\lambda_3^2). \tag{88}$$

□

**引理10**。 对于算法中定义的项，我们有

$$\frac{L}{2}\mathbb{E}\|z_{t+1}-z_t\|^2\leq\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2$$
$$+\gamma_t^2L(3\frac{1+\beta}{\beta\epsilon}(\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{L\rho_t^2}{\epsilon}+\frac{\sigma^2}{b\epsilon})+(1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2) \tag{89}$$

证明。

$$\frac{L}{2}\mathbb{E}\|z_{t+1}-z_t\|^2$$
$$=\frac{L}{2}\mathbb{E}\|\frac{\gamma\beta_1}{1-\beta_1}m_{t-1}\odot(\eta_t-\eta_{t-1})-\gamma g_t\odot\eta_t\|^2 \tag{90}$$

$$\leq\frac{L\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|m_{t-1}\odot(\eta_t-\eta_{t-1})\|^2+L\mathbb{E}\|\frac{\gamma_t}{b}\sum(\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|}))\odot\eta_t\|^2 \tag{91}$$

$$\leq\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2+L\mathbb{E}\|\frac{\gamma_t}{b}\sum(\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|}))\odot\eta_t\|^2 \tag{92}$$

$$=\gamma_t^2L\mathbb{E}\|\frac{1}{b}\sum(\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|}))\odot\eta_{t-1}+\frac{1}{b}\sum(\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|}))\odot(\eta_t-\eta_{t-1})\|^2$$
$$+\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \tag{93}$$

$$\leq\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2+\gamma_t^2L((1+\frac{1}{\beta})\mathbb{E}\|\frac{1}{b}\sum(\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|}))\odot\eta_{t-1}\|^2$$
$$+(1+\beta)\mathbb{E}\|\frac{1}{b}\sum(\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|}))\odot(\eta_t-\eta_{t-1})\|^2) \tag{94}$$

$$\leq\gamma_t^2L((1+\frac{1}{\beta})\mathbb{E}\|\frac{1}{b}\sum(\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|}))\odot\eta_{t-1}\|^2+(1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2)$$
$$+\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \tag{95}$$

$$\leq\gamma_t^2L((1+\frac{1}{\beta})\mathbb{E}\|\frac{1}{b}\sum(\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|}))\odot\sqrt{\eta_{t-1}}\|^2\times\|\sqrt{\eta_{t-1}}\|_\infty^2$$
$$+(1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2)+\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \tag{96}$$

$$\leq\gamma_t^2L(\frac{1+\beta}{\beta\epsilon}\mathbb{E}\|\frac{1}{b}\sum(\nabla f_i(x_t+\rho_t\frac{s_t}{\|s_t\|}))\odot\sqrt{\eta_{t-1}}\|^2+(1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2)$$
$$+\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \tag{97}$$

$$\leq \gamma_t^2 L(3\frac{1+\beta}{\beta\epsilon}\mathbb{E}(\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\|(\frac{1}{b}\sum\nabla f_i(x_t)-\nabla f(x_t))\odot\sqrt{\eta_{t-1}}\|^2$$
$$+\|\frac{1}{b}\sum(\nabla f_i(x_t+\rho_t\frac{\sum_{i\in B}\nabla f_i(x_t)}{\|\sum_{i\in B}\nabla f_i(x_t)\|})-\nabla f_i(x_t))\odot\sqrt{\eta_{t-1}}\|^2)+(1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2)$$
$$+\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \tag{98}$$
$$\leq \gamma_t^2 L(3\frac{1+\beta}{\beta\epsilon}(\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\mathbb{E}\|\frac{1}{b}\sum(\nabla f_i(x_t+\rho_t\frac{\sum_{i\in B}\nabla f_i(x_t)}{\|\sum_{i\in B}\nabla f_i(x_t)\|})-\nabla f_i(x_t))\odot\sqrt{\eta_{t-1}}\|^2$$
$$+\frac{\sigma^2}{b\epsilon})+(1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2)+\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \tag{99}$$
$$\leq \gamma_t^2 L(3\frac{1+\beta}{\beta\epsilon}(\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{1}{\epsilon}\mathbb{E}\|\frac{1}{b}\sum(\nabla f_i(x_t+\rho_t\frac{\sum_{i\in B}\nabla f_i(x_t)}{\|\sum_{i\in B}\nabla f_i(x_t)\|})-\nabla f_i(x_t))\|^2$$
$$+\frac{\sigma^2}{b\epsilon})+(1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2)+\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \tag{100}$$
$$\leq \gamma_t^2 L(3\frac{1+\beta}{\beta\epsilon}(\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{1}{\epsilon b}\mathbb{E}\sum\|\nabla f_i(x_t+\rho_t\frac{\sum_{i\in B}\nabla f_i(x_t)}{\|\sum_{i\in B}\nabla f_i(x_t)\|})-\nabla f_i(x_t)\|^2$$
$$+\frac{\sigma^2}{b\epsilon})+(1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2)+\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2 \tag{101}$$
$$\leq \gamma_t^2 L(3\frac{1+\beta}{\beta\epsilon}(\mathbb{E}\|\nabla f(x_t)\odot\sqrt{\eta_{t-1}}\|^2+\frac{L\rho_t^2}{\epsilon}+\frac{\sigma^2}{b\epsilon})+(1+\beta)G^2\mathbb{E}\|\eta_t-\eta_{t-1}\|^2)$$
$$+\frac{LG^2\gamma^2\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\|\eta_t-\eta_{t-1}\|^2. \tag{102}$$

□

## 附录C 附加实验图示

### *A.* 实验图示

在消融实验中，我们分别使用 AdaSAM、AMSGrad、SAM 和随机梯度下降在 GLUE 基准上进行实验。这些优化器不包含动量部分（$\beta_1 = 0$）。作为表 II的补充，图 3 展示了训练过程中损失和评估指标随步数变化的详细曲线。在所有任务中，AdaSAM 的损失曲线都比 SAM 和随机梯度下降下降得更快，且下降速度与 AMSGrad 相近。AdaSAM 和 AMSGrad 的指标曲线表明，自适应学习率方法优于随机梯度下降和 SAM。而在所有任务中，AdaSAM 的下降速度与 AMSGrad 一样快。

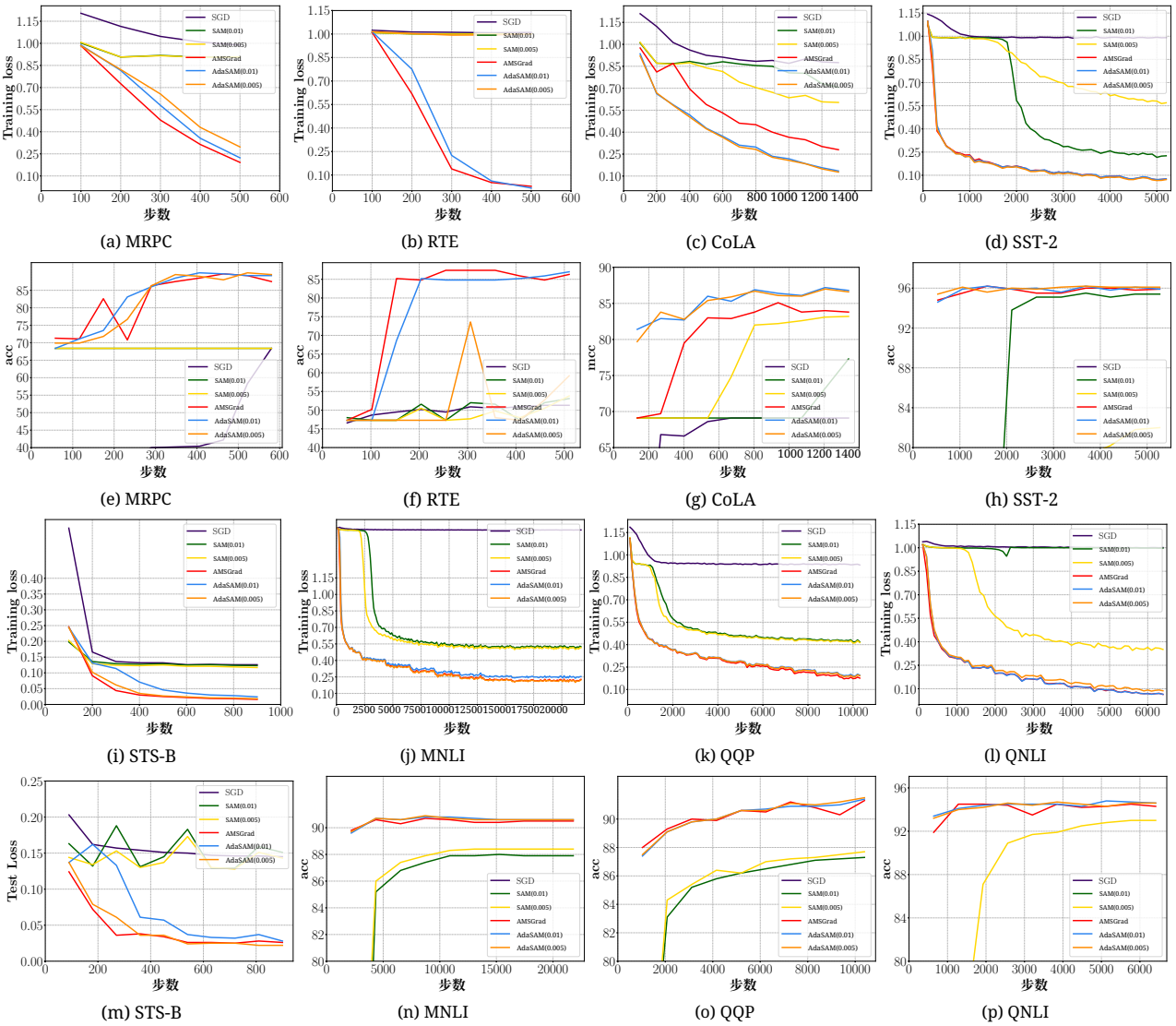



图 3：MRPC、RTE、CoLA、SST-2、STS-B、MNLI、QQP 和 QNLI 上的损失和评估指标随步数的变化。($\beta_1 = 0$)