arXiv:2603.02092v1 [cs.LG] 2 Mar 2026

Adam 无需修改更新规则即可收敛

Yushun Zhang[12], Bingran Li[1], Congliang Chen[1],Zhi-Quan Luo[12],Ruoyu Sun[12] †

[1]香港中文大学（深圳），中国[2]深圳大数据研究院
{yushunzhang,bingranli,congliangchen}@link.cuhk.edu.cn,
{sunruoyu,luozq}@cuhk.edu.cn

摘要

Adam 是训练神经网络（包括大语言模型（LLM））的默认算法。然而，Reddi 等人 [2018] 给出了一个 Adam 发散的示例，引发了人们对其在人工智能模型训练中应用的担忧。我们发现，该发散示例与实际应用之间存在一个关键差异：Reddi 等人 [2018] 先选定 Adam 的超参数，即(β_1,β_2)，再选择问题；而实际应用通常先固定问题，再调整(β_1,β_2)。在本文中，我们证明，采用适当且依赖于问题的超参数时，Adam 能够收敛。首先，我们证明，当 β_2 较大且 $\beta_1<\sqrt{\beta_2}$时，Adam 能够收敛。其次，当 β_2 较小时，我们指出存在一片(β_1,β_2) 组合区域，使 Adam 会发散至无穷大。我们的结果表明，改变(β_1,β_2) 组合会使 Adam 从发散转变为收敛，呈现出相变。据我们所知，这是文献中首次报道(β_1,β_2) 二维平面中的相变，为 Adam 优化器提供了严格的理论保证。我们进一步指出，临界边界(β_1^*,β_2^*) 依赖于问题，尤其依赖于批量大小。这为如何调整 β_1 和 β_2提供了建议：当 Adam 表现不佳时，我们建议根据批次大小反向调高 β_2，以超过阈值 β_2^*，然后尝试 $\beta_1<\sqrt{\beta_2}$。若干实证研究的报告支持了我们的建议；这些研究观察到，采用这些方法后，大语言模型训练性能有所提升。

1 引言

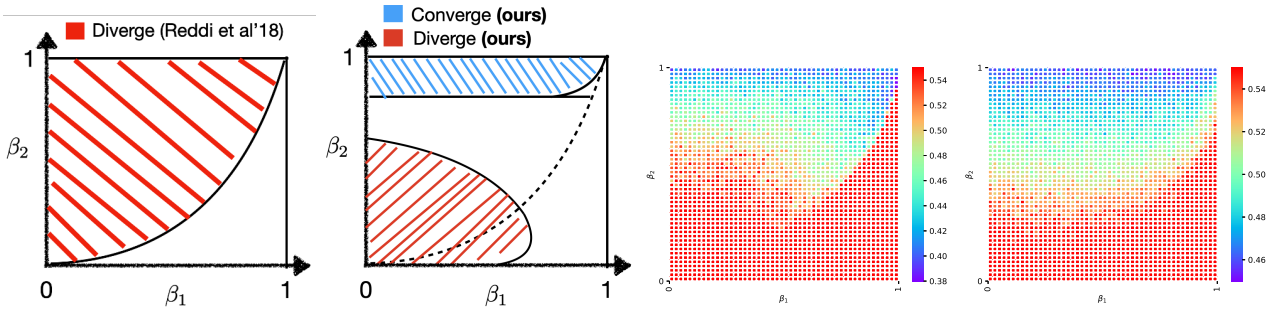
机器学习任务通常旨在求解以下经验风险最小化（ERM）问题。

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x):=\frac{1}{n}\sum_{i=0}^{n-1}f_i(x), \tag{1.1}$$

其中，$x\in\mathbb{R}^d$ 表示可训练参数，$n\in\mathbb{N}$ 表示将数据集划分为若干小批次的数量，$f_i(x)$ 表示第 i 个小批次数据上的损失。对于大小为 $\mathcal{D}$ 的固定数据集，每个小批次的批量大小为 $\mathcal{D}/n$。在深度学习中，Adam [Kingma 和 Ba, 2014] 是求解（1.1）最常用的算法之一。它已应用于自然语言处理（NLP）和计算机视觉（CV）等多个领域（例如， [Vaswani 等人, 2017年, Dosovitskiy 等人, 2021]）。截至2025年12月，其影响力也体现在超过23万次引用上，而且这一数字仍在快速增长 [Scholar, 2025]。

在大语言模型（例如 [OpenAI、 2022]）时代，Adam 在大规模训练中发挥着核心作用。据报道，许多主流大语言模型都使用 Adam 进行训练，包括 Llama 系列 [Touvron

†：通讯作者。本文是会议论文“Adam无需对更新规则做任何修改即可收敛”的扩展期刊版本，该论文发表于 NeurIPS 2022。本文给出了 Adam 在随机打乱抽样下的简化证明（基于会议论文进行简化），以及 Adam 在有放回抽样下的新证明（受会议论文启发）。



（a）Reddi 等人所声称的发散区域 [Reddi 等人， 2018]　（b）我们的贡献　（c）MNIST　（d）CIFAR-10

图 1：（**a**）：Reddi 等人所声称的 Adam 发散区域， [Reddi 等人， 2018]。他们先固定 (β_1, β_2)，然后选择一个问题来构造发散示例。（**b**）：我们在 (β_1, β_2) 相图中的贡献示意图。我们在选择 (β_1, β_2) 之前先固定问题。请注意，这与 （**a**）中的设定不同，因此并不矛盾。**红色和蓝色区域的两条边界都取决于批量大小（稍后展示）**。区域的形状遵循我们的解析条件所得的解。点线满足 $\beta_1 = \sqrt{\beta_2}$。（**c**）、（**d**）：MNIST 和 CIFAR-10 上的训练损失。我们在网格 $\{(k_1/50, k_2/50) | k_1 = 0, \cdots, 49, k_2 = 0, \cdots, 49\}$上遍历 β_1 和 β_2，共进行 2,500 次试验。Adam 的性能与我们在 （**b**）中的理论刻画相一致。

等人、 2023], 通义千问系列 [Bai 等人、 2023], 以及 DeepSeek 系列 [Liu 等人、 2024], 等等。显然，Adam 是推动人工智能进步的重要动力。它的影响力最近也得到了认可，获得了 *ICLR 2025* 时间检验奖 [ICLR， 2025]。

尽管其应用十分广泛，一篇有影响力的论文 [Reddi 等人， 2018] （*ICLR 2018* 最佳论文奖[ICLR， 2018]的获奖者）给出了一个示例，表明 Adam 在很大范围的超参数下会发散。一项主要结果在 [Reddi 等人， 2018] 指出：

[Reddi 等人， 2018]：对于任意 β_1， β_2 使得 $0 \leqq \beta_1 < \sqrt{\beta_2} < 1$，存在一个问题使得 *Adam* 发散。

这里， β_1 和 β_2 是用于控制 Adam 的一阶动量和二阶动量 m_k 和v_k的超参数。发散区域如图 1 (a)。这一发现引发了人们对将 Adam 部署于人工智能模型训练的严重关切，其中发散可能警示不可预测的训练失败。此后，许多新变体被设计出来。例如，AMSGrad [Reddi 等人， 2018] 强制v_k （稍后在算法1中定义）非递减；AdaBound [Luo 等人，2019a] 施加约束$v_k \in [C_l, C_u]$ 以确保有效步长有界。

另一方面，与直觉相反，原始 Adam 仍然异常流行。在不对更新规则做任何修改的情况下，Adam 在实践中表现良好。由于存在发散理论，这一点相当令人惊讶。更神秘的是，我们发现通常报告的超参数实际上满足前文所述的发散条件。例如，Kingma 和 Ba [2014] 断言$(\beta_1, \beta_2) = (0.9, 0.999)$ 是“针对所测试机器学习问题的良好选择”，并且它确实是深度学习库中的默认设置。GAN 研究人员（例如 Radford 等人 [2015], Isola 等人 [2017]）使用$(\beta_1, \beta_2) = (0.5, 0.999)$。对于 GPT-3、Llama 系列和 DeepSeek 系列等大语言模型 [Brown 等人，2020，Touvron 等人，2023年，Liu 等人， 2024], (β_1, β_2) 被选为 $(0.9, 0.95)$。所有这些超参数都位于发散区域 $\beta_1 < \sqrt{\beta_2}$。令人惊讶的是，这些超参数并未出现发散问题，反而取得了良好性能，并且实际上呈现出收敛的迹象。

尽管存在理论发散问题，为什么 Adam 仍然表现良好？深度学习问题与发散示例之间是否存在不匹配？我们仔细研究了发散示例，发现这种不匹配确实存在。具体来说，我们注意到发散示例的一个重要（但常被忽略）特征： [Reddi 等人先 2018] 选择(β_1, β_2)，然后再选择问题（或者准确地说，小批次数量n）。也就是说，为了构造发散示例，他们会针对不同的(β_1, β_2)改变n。例如，对于$(\beta_1, \beta_2) = (0,0.99)$，他们使用一个$n$ 来构造发散示例；对于$(\beta_1, \beta_2) = (0,0.9999)$，则使用另一个$n$ 来构造另一个发散示例。另一方面，在上面列出的 Adam 实际应用中，从业者会在问题（或n）确定之后再调整超参数(β_1, β_2)。因此存在一种

与.....之间的差距 理论与实践的设定：选择问题与(β_1, β_2) 的顺序不同 t.

鉴于 Adam 在实践中表现出色，我们推测，当问题固定时，Adam 能够收敛。遗憾的是，与其变体相比，原始 Adam 的行为远未得到充分研究（或许是因为它曾受到发散的批评）。为了验证这一猜想，我们在 MNIST 和 CIFAR-10 数据上运行 Adam 来解决分类问题，如图 1 (c) 和 (d) 所示。对于这两个问题，我们发现：首先，当 β_2 较大时，几乎在所有 β_1 取值下，优化误差都很小；其次，当 β_1 和 β_2 都较小时，会出现一个红色区域，其误差大 1.4 倍。

虽然 Adam 在红色区域的表现似乎不稳定，但我们发现 它在图 *1* 的顶部蓝色区域始终表现良好。 这似乎表明：当问题固定时，经过适当调节 β_1 和 β_2后，Adam 无需修改即可收敛。我们提出以下问题：

不对更新规则作任何修改，能否证明 *Adam* 可以收敛?

在这项工作中，我们从理论上刻画了 Adam 在不同(β_1, β_2)取值下的行为。我们证明，原始 Adam（未经算法修改）表现出两种性质迥异的状态：在安全区域(β_1, β_2) 内，它可被证明会收敛到临界点集合（可实现情形），或收敛到该集合的一个邻域（不可实现情形）；而在危险 区域内，它可能发散至无穷大。综合来看，这些结果揭示了(β_1, β_2) 平面中的发散—收敛相变。

我们的贡献如图 1(b) 所示，概括如下。

- **较大 β_2时的收敛**。 当 $0 \le \beta_1 < \sqrt{\beta_2} < 1$ 和 β_2大于依赖于问题的阈值时，Adam 在可实现情形下收敛到临界点，在不可实现情形下收敛到临界点的一个邻域。 β_2 的阈值取决于问题类，并且会随n 增大而增大（等价地，随批量大小增大而减小）。

- **当 β_2 较小时会发散**。对于任意$n \ge 3$ 以及任意问题类参数(L, D_0, D_1)，只要D_1足够大——其中L 是光滑常数（假设2.1），而(D_0, D_1) 用于量化梯度方差条件（假设2.2）——问题类中就存在一个实例，使得对于某个区域内的所有(β_1, β_2)，Adam 都会发散至无穷大。发散区域的边界随n 增大而扩展（或随批次大小增大而缩小）。

- **关键特征**。我们强调结果的以下几个方面。

  (1) 相变。发散结果考虑的设置与我们的收敛结果相同（固定的问题类），这表明存在一个临界边界(β_1^*, β_2^*)，它划分了从发散到收敛的相变。

  (2) 问题依赖的界。我们的收敛区域和发散区域(β_1, β_2) 取决于问题类参数。这与Reddi 等人 2018], 的做法截然不同，后者考虑的是问题无关的(β_1, β_2)。

  (3) 非渐近刻画。我们的“发散区域”(β_1, β_2) 会随着n 增大而扩展，并在n 趋于无穷大时收敛到整个区域 $[0, 1)^2$，这重现了（实际上强于） [Reddi 等人 2018]的问题无关发散结果。从这个意义上说，我们可以将 [Reddi 等人 2018] 的发散结果视为发散区域的渐近刻画（当$n \to \infty$时），而将我们的发散结果视为非渐近刻画（对于任何固定且有限的n）。

  (4) 不要求有界梯度假设。我们的收敛分析不需要文献中常用的有界梯度假设。去掉这一假设对于揭示相变至关重要：如果梯度先验有界，Adam 的梯度就不可能发散；而我们证明，在某些(β_1, β_2)条件下，这种情况确实可能发生。

- 收敛证明的主要挑战在于刻画随机非线性动力学的极限行为。我们通过考察 Adam 的 $1/\sqrt{v_k}$ 在 $1/\sqrt{\mathbb{E}}$ (v_k)附近的集中效应来处理这一问题。一个具体的难点是，v_k 可能任意接近 0，而在最坏情况下，$1/\sqrt{v_k}$ 的表现可能很差。我们发现，较大的 β_2 有助于集中并稳定动力系统。直观来说，较大的 β_2 会减缓 v_k的变化，使其行为变得可预测。

• **关于调节 (β_1, β_2)的建议**。 我们的正面和负面结果为调节 β_1 和β_2提供了指导：当 Adam 表现不佳时，我们建议随着批量大小的增加，相应调高 β_2，使其超过阈值 β_2^*，然后再尝试 $\beta_1 < \sqrt{\beta_2}$。这些建议得到了大语言模型预训练中的实证发现支持 [Porian 等人，2024年，Zhang 等人，2024a]。

我们相信，我们的结果推进了对 Adam 的理论理解。尽管 Reddi 等人 [2018] 指出“Adam 可能发散”，但我们的结果展现了另一面：当问题固定时，适当调参便能使 Adam 无需任何修改即可收敛。具体而言，(β_1, β_2)的收敛区域保证非空。这类似于在 L-光滑目标函数上进行梯度下降，其中总存在一个能保证收敛的步长（例如 $\eta < 2/L$）。总的来说，我们的结果为原始 Adam 提供了理论依据，并为实际应用中如何选择 (β_1, β_2) 提供了具体指导。

与会议版本的关系 [Zhang 等人，2022b]。 本文是我们发表于 NeurIPS 2022 的会议论文 [Zhang 等人，2022b]的扩展期刊版本。当前稿件给出了随机打乱条件下 Adam 的简化收敛证明（相较会议版本大幅简化），并建立了有放回抽样条件下的新收敛结果及其证明；会议版本未涉及这一问题。从技术上讲，这个期刊版本沿着 Adam 的轨迹推导出新的集中性结果，并提出了一套系统性流程，以便在存在动量和无界梯度的情况下应用这些集中性工具。这些改进将随机打乱证明的篇幅从约 44 页缩短至约 18 页，并使有放回抽样条件下的收敛分析成为可能。相较会议版本，这些证明技术都是新的，也可能广泛适用于 Adam 类算法的分析。

2 预备知识

2.1 Adam 概述

我们考虑问题 (1.1)。我们用 x 表示优化变量。我们将 ∇f_i 记作 f_i的梯度。我们考虑基于数据抽样策略的 Adam 的两种实现：有放回抽样下的 Adam（算法 1）和随机打乱下的 Adam（算法 2）。算法 1 在理论上更具针对性，而算法 2 则在 Adam 的实际部署中更为常用 [1]。

算法 1 有放回抽样下的 Adam

初始化 $\mathbf{m}_0, \mathbf{v}_0$, 和 $\mathbf{x}_1$。对于 $\mathbf{k} = 1 \to \infty$,
执行以下操作：从 $\{0, 1,$ 中均匀抽取 τ_k。 . . .
, $\mathbf{n} - 1\}$。 $\mathbf{m}_k = \beta_1 \mathbf{m}_{k-1} + (1-\beta_1)\nabla \mathbf{f}_{\tau_k}(\mathbf{x}_k)$
$\mathbf{v}_k = \beta_2 \mathbf{v}_{k-1} + (1-\beta_2)\nabla \mathbf{f}_{\tau_k}(\mathbf{x}_k) \circ \nabla \mathbf{f}_{\tau_k}(\mathbf{x}_k)$
$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\eta_k}{\sqrt{\mathbf{v}_k + \epsilon}} \circ \mathbf{m}_k$结束循环

在算法1 和2中，m 和v 分别表示一阶动量和二阶动量。乘积 $\circ$除法和平方根算子均按分量进行。对于算法1，我们将x_k、m_k、$v_k \in \mathbb{R}^d$分别记为第k次迭代时x、m、v 的值。对于算法2，我们将$x_{k,i}$、$m_{k,i}$、$v_{k,i} \in \mathbb{R}^d$分别记为第k次外循环（轮次）和第i次内循环（批次）时x、m、v 的值。我们选择$\eta_k = \frac{\eta_0}{\sqrt{k}}$ 作为步长。

在算法1 和2中，采用 ϵ 以避免这样的边界情况：沿着轨迹v_k始终为 0，此时 Adam 没有定义。为确保 Adam 定义良好，可以选择：(i) 使用较小的 ϵ （例如，默认 ϵ 为 10^{-8}）；或 (ii) 使用 $\epsilon = 0$，并采用非零初始化$v \succ 0$。我们的理论支持

[1]例如，GPT -3 技术报告指出：“训练期间，数据以无放回方式抽样”[Brown 等人， 2020]。

算法 2 随机打乱下的 Adam

初始化 m1、−1、v1、−1 和 x1,0。对于 k = 1 → ∞，执行以下操作：将 $\{\tau_{k,0}, \tau_{k,1}, \cdots, \tau_{k,n-1}\}$ 采样为 $\{0, 1, 2, \cdots, n-1\}$ 的随机排列。对于 i = 0 → n − 1，执行：$\mathbf{m}_{k,i} = \beta_1 \mathbf{m}_{k,i-1} + (1-\beta_1) \nabla \mathbf{f}_{\tau_k,i}(\mathbf{x}_{k,i})$；$\mathbf{v}_{k,i} = \beta_2 \mathbf{v}_{k,i-1} + (1-\beta_2) \nabla \mathbf{f}_{\tau_k,i}(\mathbf{x}_{k,i})$；$\circ \nabla \mathbf{f}_{\tau_k,i}(\mathbf{x}_{k,i})$；$\mathbf{x}_{k,i+1} = \mathbf{x}_{k,i} - \sqrt{\eta_k}\mathbf{v}_{,i+\epsilon} \circ \mathbf{m}_{k,i}$。结束 for。$\mathbf{x}_{k+1}$, $_0 = \mathbf{x}_{k,n}$；$\mathbf{v}_{k+1,-1} = \mathbf{v}_{k,n-1}$；$\mathbf{m}_{k+1,-1} = \mathbf{m}_{k,n-1}$。结束 for

这两种情况。在证明的主体部分，我们采用 (ii)，这样结果更简洁。我们还在附录 E 中给出了情况 (i) 的证明，即具有“非零 ϵ”的 Adam 的收敛性。无论哪种情况，我们最终的收敛结果都不依赖于 ϵ 或初始化 v，因此，无论 ϵ 或 v 如何取值，我们的结果都仍然是非平凡的。

在 [Kingma 和 Ba 的原始论文中， 2014], 作者引入了一个额外的“偏差校正”步骤，可以通过将步长 η_k 改为 $\hat{\eta}_k = \frac{\sqrt{1-\beta_2^k}}{1-\beta_1^k}\eta_k$ 来实现。注意，$\hat{\eta}_k \in [\sqrt{1-\beta_2}\eta_k, \frac{1}{1-\beta_1}\eta_k]$ 在 η_k 附近有界，因此 η_k 和 $\hat{\eta}_k$ 具有相同的收敛速率。在证明的主体部分，我们采用算法 1 和 2的形式，这样结果更简洁。为求完整，我们在附录 E 中补充了带有“偏差校正”步骤的 Adam 收敛性证明。

在我们的分析中，我们对 $f_i(x)$ 和 $f(x)$ 在经验风险最小化问题（1.1）作出以下温和假设。

假设 2.1。对于任意 $i \in [n]$ 以及任意 x、$y \in \mathbb{R}^d$, $\|\nabla f_i(x) - \nabla f_i(y)\|_2 \le L\|x-y\|_2$。此外，$f(x)$ 有一个有限常数 f^* 作为下界。

假设 2.2。 $\sum^{n-1}{}_{i=0} \|\nabla f_i(x)\|_2^2 \le D_1\|\nabla f(x)\|_2^2 + D_0, \forall x \in \mathbb{R}^d$，其中 $D_1, D_0 \ge 0$ 且二者不同时为零。

当 n、L、D_0, 和 D_1 预先固定时，我们定义相应的问题类

$$\mathcal{F}^n_{L,D_0,D_1}(\mathbb{R}^d) := \left\{ f(x) \,\middle|\, f(x) = \frac{1}{n}\sum_{i=0}^{n-1} f_i(x),\ x \in \mathbb{R}^d \text{ and Assumptions 2.1–2.2 hold with } (L, D_0, D_1) \right\}. \tag{2.1}$$

在假设2.1中，分量函数的 Lipschitz 条件是经验风险最小化问题中的标准条件（例如，[Bertsekas 和 Tsitsiklis, 1996, 2000, Bubeck, 2015, Schmidt 等人, 2017年, Allen-Zhu, 2018]）。

假设 2.2 涵盖了文献中的一类方差或增长条件。现在，我们讨论在对 D_1 和 D_0 作出不同选择时，假设 2.2 如何化简为不同的条件。

我们首先讨论 D_1。在对 $D_1 > n$ 施加稍严格一些的条件时，假设 2.2 就变成了式（2.2）中的“仿射方差”条件，它控制小批量梯度与全梯度之间的偏差。“仿射方差”条件最初由 [Bertsekas 和 Tsitsiklis提出, 2000] 后来由 [Bottou 等人推广， 2018]。

$$\frac{1}{n}\sum_{i=0}^{n-1}\|\nabla f_i(x) - \nabla f(x)\|_2^2 = \frac{1}{n}\left(\sum_{i=0}^{n-1}\|\nabla f_i(x)\|_2^2\right) - \|\nabla f(x)\|_2^2 \le \frac{(D_1-n)}{n}\|\nabla f(x)\|_2^2 + \frac{D_0}{n}, \forall x \in \mathbb{R}^d. \tag{2.2}$$

当 $D_1 = n$ 时，假设 2.2 和式（2.2）可化简为式（2.3）中的“有界方差”条件：

$$\frac{1}{n}\sum_{i=0}^{n-1}\|\nabla f_i(x) - \nabla f(x)\|_2^2 = \frac{1}{n}\left(\sum_{i=0}^{n-1}\|\nabla f_i(x)\|_2^2\right) - \|\nabla f(x)\|_2^2 \le \frac{D_0}{n}, \forall x \in \mathbb{R}^d. \tag{2.3}$$

“有界方差”条件被广泛用于分析随机梯度下降和自适应梯度方法（例如 [Polyak 和 Juditsky, 1992, Ghadimi 和 Lan, 2013, Ghadimi 等人, 2016, Zaheer 等人, 2018]）。假设2.2 允许方差随梯度范数增大，因此其严格弱于

有界方差条件（2.3）。这种放宽是有意义的：有界方差条件往往过于严格，而且很容易不成立。例如，考虑一个凸二次最小化问题 [Bottou 等人， 2018]:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x) = \frac{1}{2n}\|Ax-b\|_2^2 = \frac{1}{2n}\sum_{i=0}^{n-1}(a_i^\top x - b_i)^2 := \frac{1}{n}\sum_{i=0}^{n-1} f_i(x), \tag{2.4}$$

其中 A 的秩至少为 2。在这种情况下，(2.3) 左侧随 $\|x\|_2$呈二次增长，因此不存在有限的 D_0 能够满足 (2.3)。相比之下，假设2.2 在有限的 D_1 和 D_0下成立。我们在附录 A.2中提供了更详细的论证。

有界方差条件 (2.3) 的另一个缺点是，它排除了 SignSGD 的一些发散反例（SignSGD 等价于 $\beta_1 = \beta_2 = 0$ 的 Adam）。因此，基于 (2.3) 的分析可能无法揭示 Adam 的全貌。我们将在第 3.2 节提供更详细的解释。

还有一个更强的条件，要求 $D_1 = 0$，即 $\sum^{n-1}{}_{i=0} \|\nabla f_i(x)\|_2^2 \le D_0$。这个条件有时称为“有界二阶矩”（例如， [Nemirovski 等人， 2009]）。它严格强于 (2.3)，并蕴含有界梯度条件，因此限制更为严格。

因此，Bottou 等人建议采用2.2 中使用通用D_1 的假设 [Bottou 等人， 2018] 因为它“相对宽松”，而且“允许方差沿任意方向以二次速率增长。”

现在我们讨论 D_0。当 $D_0 = 0$ 时，该条件就成为“强增长条件”（SGC） [Solodov，1998年，Schmidt 和 Roux， 2013]。SGC 的一个推论是，当 $\|\nabla f(x)\| = 0$ 时，对于所有 i，都有 $\|\nabla f_i(x)\| = 0$。在过参数化情形下，神经网络可以拟合所有数据点，因此这一条件被认为是合理的 [Vaswani 等人，2019年，Shi 等人，2020]。我们将证明，当 SGC 成立时，Adam 会收敛到精确的临界点。

对于$D_0 > 0$的一般情形，Adam 不保证能够到达精确的临界点。相反，它只能收敛到临界点的一个邻域 [Zaheer 等人，2018，Shi 等人， 2020]。在实验中，即使采用递减步长，Adam 也确实会出现这种现象（见图 3(b)）。这种行为与随机梯度方法中的经典现象一致：已知恒定步长随机梯度下降会收敛到临界点的一个邻域，其大小随噪声水平而变化 [Luo，1991，Bertsekas，1997，Yan 等人，2018，Yu 等人，2019，Liu 等人，2020b]。

最后，我们强调，我们不会加入文献中常用的有界梯度假设 $\|\nabla f(x)\| \le G$。这对于揭示相变至关重要：如果梯度先验有界，Adam 的梯度就不会发散；而我们证明，在某些(β_1, β_2)条件下，这种情况确实可能发生。

2.2 Reddi 等人反例的简要回顾 [Reddi 等人， 2018]

在陈述我们的理论结果之前，我们先重述 [Reddi 等人， 2018] 给出的反例。为保持符号一致，我们将用本文的符号重述他们的结果。他们考虑以下一维凸问题：最小化 $\frac{1}{n}\sum^{n-1}{}_{i=0} f_i(x)$，其中 $x \in [-1,1]$，$n \ge 3$：

$$f_i(x) = \begin{cases} nx, & \text{for } i = 0 \\ -x, & \text{otherwise,} \end{cases} \tag{2.5}$$

注意，(2.5) 同时满足假设 2.1 和2.2 （取 $D_1 = n^4 + n^3 - n^2$ 且 $D_0 = 0$），因此我们的假设并未先验地排除这个反例。尽管如此，问题 (2.5) 是一个可行集为 $x \in [-1,1]$ 的约束问题，且其最优解为 $x^* = -1$。由于这是一个约束问题，此处“发散”一词实际上意味着迭代点会停留在一个范围巨大的区域内，其大小与整个可行集相当。由于这是一个约束问题，此处“发散”一词实际上意味着迭代点会停留在可行域的边界上，并且远离最优解 $x* = -1$。请注意，这两种情况描述的算法行为都与收敛相反，我们不强调二者之间的区别。

在 [Reddi 等人， 2018], 函数（2.5）被表述为“在线优化问题”（而非有限和问题）。我们将（2.5）改写为有限和形式，使其与我们在（1.1）中的经验风险最小化记法相符。我们将采用与（f_0, f_1, f_2; f_0, f_1, f_2; .....）相同的循环采样顺序，如 Reddi 等人 [2018]所用。为完整起见，我们在下文用我们的记法重述他们的主要断言。

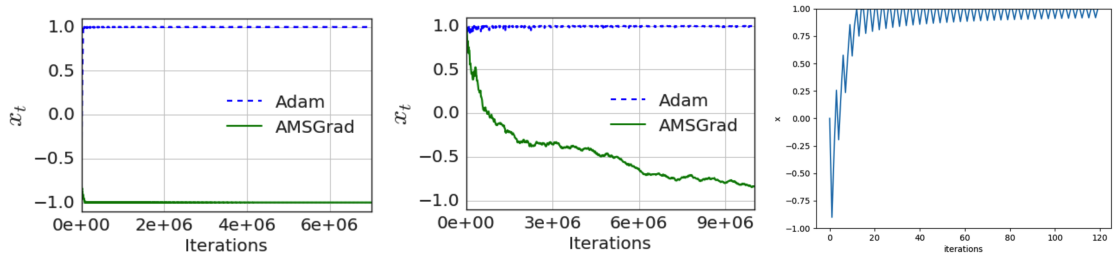
定理 2.3 （ [Reddi 等人的 2018]定理 2）。对于任意满足 $\beta_1 < \sqrt{\beta_2}$ 的固定（β_1，β_2），存在一个充分大的 n，使得对函数（ 2.5 ）应用 *Adam* （采用循环采样）会收敛到次优点 $x = 1$。

我们简要讨论该定理的发散条件。正如 [Reddi 等人 的附录 B 中公式（7）所述，对于每一组固定的 (β_1, β_2)，他们都需要一个“依赖于 β_1 和 β_2 的 n ”。因此，要使不同的 (β_1, β_2) 发生发散，他们需要不同的 n 。所考虑的问题类一直在变化。

为完整起见，我们进一步重述 [Reddi 等人 的定理 1 2018]。

定理 2.4 ([Reddi 等人 的定理 1 2018])。 对于函数 (2.5)，当 $\beta_1 = 0$ 和 $\beta_2 = 1/(n^2+1)$ 时，*Adam* 将收敛到一个次优点 $x = 1$。

该定理考虑在经过 n 次后选择 (β_1, β_2) 。然而，这一结果仅表明在单个超参数选择 $(\beta_1, \beta_2) = (0, 1/(n^2+1))$ 下会发生发散。该配置位于图1 (a) 左边界的某处。因此，除了这一点之外，它并未给出任何刻画：目前仍不清楚，沿哪些方向改变 (β_1, β_2) 能够保持发散，沿哪些方向会抑制发散或导致收敛。



(a) 循环更新顺序　(b) 随机更新顺序　(c) 递减步长

图 2：**(a, b)：**我们重新展示了 [Reddi 等人 2018] 中的图 1。Adam 在循环更新顺序和更新顺序下都会发生发散，因此随机化无法阻止发散。由于他们考虑的是约束问题，此处“发散”一词是指停滞在次优解 $x = 1$。图中，AMSGrad 是一种不同的方法，并非我们的关注重点。**(c)：**递减步长 $\eta_k = \frac{1}{\sqrt{k}}$ 无法阻止发散。

最后，正如 [Reddi 等人，2018，定理 3] 所证明的，当使用随机更新顺序而非循环顺序时，这些发散结果同样成立。因此，随机化无法阻止发散。此外， [Reddi 等人， 2018] 的证明使用了 $\eta_k \propto \frac{1}{\sqrt{k}}$，因此递减步长也无法阻止发散。这些断言得到了图 2 中数值证据的支持。

Reddi 等人 [2018] 写道：“这些结果有着重要影响，因为为了避免不良的收敛行为，必须使用依赖于问题的ϵ， β_1和 β_2 。对于高维问题，这通常意味着要针对每个维度分别使用不同的..... ϵ， β_1和 β_2 。然而，这违背了自适应方法的初衷，因为它需要调优大量参数。”接下来，我们将证明，在我们的设定下，无须更改算法或引入新的超参数，就能避免这种发散。关键在于以依赖于问题的方式选择(β_1, β_2) ，并使其特别依赖于小批次的数量n（等价地，也就是批量大小）。重要的是，这不要求对(β_1, β_2)进行逐坐标调优——Reddi 等人 [2018]在讨论中提到过这一方向。

2.3 相关工作

自 Reddi 等人 [2018] 指出发散问题以来，许多研究尝试设计新的 Adam变体。由于我们专注于理解未经修改的原始 Adam，因此本文不讨论这些变体。关于更多 Adam变体的介绍，我们将在附录 A.1中再行说明。与提出新的变体相比，原始 Adam 的收敛性研究远少于对其变体的研究（或许是由于对发散问题的批评）。具体讨论如下。

Shi 等人 [2020] 研究了 [Hinton 等人提出的随机打乱 RMSProp，即 2012] 带有$\beta_1 = 0$的算法 2 [2]。我们认为，研究 Adam 而非 RMSProp 十分重要：在数值上，Adam 的性能通常显著优于 RMSProp（例如 [Agarwal 等人， 2020]）。此外，所有主流大语言模型都使用 Adam 进行训练，据报道，RMSProp 的性能比 Adam $\geq$ 20低 % [Orvieto 和 Gower，2025年，表 1]。理论上，关于 RMSProp 的文献无法揭示这些超参数如何共同影响（或危及）Adam 在（β_1， β_2）二维平面中的收敛性。我们注意到，Shi 等人 [2020] 猜想 RMSProp 存在一个临界阈值 β_2^*。我们将这一猜想推广到 Adam 的（β_1， β_2）二维平面，并证实临界边界（β_1^*， β_2^*）的存在。另一个区别是，Shi 等人 [2020]仅研究了随机打乱 RMSProp，而我们则考虑了更一般的有放回 Adam 和随机打乱 Adam 情形。

Défossez 等人 [2022]分析了具有 $0 < \beta_2 < 1$ 和 $\beta_1 < \beta_2$的 Adam。然而，他们要求满足有界梯度假设$\|\nabla f(x)\| \leq G$，这从一开始就排除了潜在的发散。此外，他们的界与 $1/\epsilon$成正比，其中 ϵ 是稳定性超参数，通常较小（例如，默认值 ϵ 为 10^{-8}）。因此，他们要求非零 ϵ 足够大，才能使结果不至于空洞。这些假设给 Adam 人为地施加了有界性：$0 < \epsilon \leq \sqrt{v} + \epsilon \leq G$，因此 Adam 永远不会发散，并且实际上被改成了另一种算法 AdaBound [Luo 等人，2019b]。相反，我们不施加这些有界性条件，并揭示了不同 (β_1, β_2)下的双收敛相变。

2022年会议版本在线发布后，关于 Adam 的新证明。 本文的会议版本 [Zhang 等人，2022b]于2022年8月在线发布后，陆续有一系列工作为 Adam 提供了新的收敛证明。在此，我们总结这些工作与我们的假设有何不同。

近期的研究 [Xie 等人、2024年、Li 等人、2025a、b] 为 RMSProp 和 Adam 提供了新的证明。与我们的工作相比，他们的证明依赖于“逐坐标有界方差”条件，该条件要求每个坐标分别满足不等式（2.3）。请注意，这比（2.3）更强，因为后者只要求整体梯度满足该不等式。此外，既然已知（2.3）具有局限性——例如，它在简单的二次问题（2.4）中也不成立——那么其逐坐标变体必然限制更多。除了逐坐标有界方差外，他们还要求 T 相关的 (β_1, β_2)，其中 T 是预先确定的总迭代次数预算。在实践中，从业者很少使用或报告与 T 相关的 (β_1, β_2)。在理论上，我们认为这种对 T 的依赖是不必要的，可以去除。相比之下，我们的证明不需要这两个条件。

Wang 等人 [2023]在“逐坐标仿射方差”条件下，为 Adam 提供了一种新颖的收敛证明，并且 T-依赖的 (β_1, β_2)。类似地，逐坐标版本严格强于仿射方差条件（2.2）或假设2.2；而且 T-依赖性并非必要。 Wang 等人[2024] 进一步放宽了这两个条件，但他们的分析聚焦于随机打乱下的 Adam，未涵盖有放回抽样。

Li 等人 [2023], Peng 等人 [2025], Hong 和 Lin [2023, 2024], Zhang 等人 [2024b] 给出了更精细的 Adam 收敛证明。其局限之一是复杂度与 $1/\epsilon$ 成正比，其中 ϵ 是稳定性超参数。当 ϵ 趋近于 0 时，这些界会变得无意义；这令人担忧，因为 ϵ 在实践中通常很小（例如，默认的 ϵ 为 10^{-8}）。我们认为，对 ϵ 的依赖是人为的；相比之下，我们的结果独立于 ϵ，并且允许 ϵ 任意小，包括 0。除了依赖于 ϵ之外，Li 等人 [2023], Peng 等人 [2025] 还要求满足有界方差条件（2.3）或其更强版本；Hong 和 Lin [2023, 2024],Zhang 等人 [2024b] 则要求仿射方差条件（2.2）几乎必然成立，或满足其逐坐标版本。所有这些假设都强于假设 2.2。

另一系列新工作在有界梯度或有界迭代点，或有界v_k 假设下证明了 Adam 的收敛 [Ding等人，2023年，Jiang等人，2023年b，Xiao等人，2024年，Liang等人， 2025]，，而我们并不施加这些有界性条件。

除了假设上的差异外，我们还要强调我们的工作与上述文献之间一个重要的不同之处。上述工作着重于证明在某些条件下 Adam 的收敛上界，其中条件为

[2]我们注意到Shi 等人 [2020] 也分析了随机打乱 Adam，且 β_1 足够接近 0。然而，Zhang 等人 [2022a] 表明他们需要 $\beta_1 < 10^{-7}$。因此，他们的结果除随机打乱 RMSProp 之外并未提供太多额外信息。

特定的（β_1, β_2）取值。然而，良好的上界只是问题的一面。与之相反，我们揭示了在（β_1, β_2）的不同区域中，从发散到收敛的相变，从而呈现出更完整的图景。我们的工作指出了文献中遗漏的两个事实。

- **首先**，我们证明，调整 (β_1, β_2) 不仅足以保证收敛，而且也是收敛所必需的。具体而言，我们证明存在一个相变：随着 β_2 从 0 增至 1，Adam 的行为会从发散（定理 3.5中的较差下界）转变为收敛（定理 3.1 和3.3中的良好上界）。这表明，调整 β_2 不仅仅是为了优化收敛速率，对于防止根本性的发散风险也至关重要。据我们所知，这是文献中首次报告 Adam 在(β_1, β_2) 二维平面上出现相变。

- **其次**，我们证明，这一相变发生在一个依赖于问题的边界(β_1^*, β_2^*)处，该边界随批量大小的增加而反比缩小。这与先前的理论研究形成对比：那些研究认为 β_2 应随着总迭代预算T 的增加而增大，且未揭示其对所考虑问题的依赖性。

3 主要结果

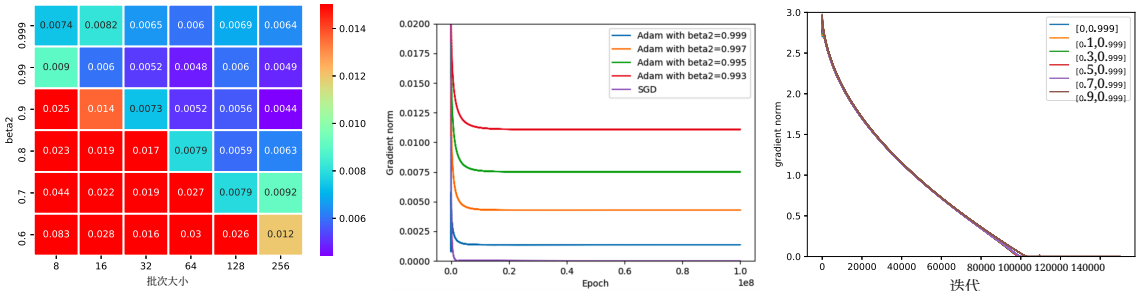
3.1 收敛结果

现在，我们证明在适当的、依赖于问题的超参数下，Adam 会收敛。我们先给出算法 1 的结果，然后给出算法 2 的结果。

定理 3.1。（算法 1的收敛结果）假设算法 *1* 满足：$0 \le \beta_1 < \sqrt{\beta_2} < 1$；$\beta_2$满足 $\beta_2 \ge \gamma_1(n) := 1 - \overline{\mathcal{O}}\left(\frac{1-\beta_1^n}{n^5}\right)$；且 $\eta_k = \frac{\eta_0}{\sqrt{k}}$。对于任意 $f(x) \in \mathcal{F}_{nL,D_0,D_1}(\mathbb{R}^d)$，当 $T \in \mathbb{N}$ 足够大时，我们有

$$\min_{k\in[1,T]} \mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_k)\|_2^2}{\sqrt{D_0}}, \frac{\|\nabla f(x_k)\|_2}{2\sqrt{dD_1}}\right\}\right] = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right) + \mathcal{O}(\delta(\beta_2)\sqrt{D_0}),$$

其中 $\delta(\beta_2)$ 是一个正的常数，当 β_2 趋近于 *1* 时趋近于 *0*（参见第 (6.46) 节中的6）。



（a）批量大小与 β_2　　(b) 大 β_2 且$D_0 > 0$　　(c) 大 β_2 且$D_0 = 0$

图 3：**(a):** 不同批量大小和 β_2下 Adam 在 MNIST 上的训练损失。趋势与我们的理论一致：当批量大小较小时，我们需要更大的 β_2。此处，我们使用默认的 $\beta_1 = 0.9$。**(b) (c):** 大-β_2 Adam 在 $D_0 > 0$ 时收敛到临界点的一个邻域，而在 $D_0 = 0$时收敛到精确的临界点。我们使用递减步长 $\eta_k = 0.1/\sqrt{k}$，与我们的理论一致。实验细节见附录 F。

注释 1：选择 β_2。我们的理论表明，应使用较大的 β_2 以确保收敛。这一结论与我们在图1中的实验结果相符。我们强调，“较大的 β_2”这一要求是必要的，因为较小的 β_2 确实会导致可证明的发散（将在定理3.5中展示）。这里我们对阈值 $\gamma_1(n)$稍作说明。根据定理3.1的证明，尤其是第6.43节中的式6，β_2 需要

满足 $\beta_2 \geq 1 - \mathcal{O}\left(\frac{1-\beta_1^n}{n^5}\right)$。我们还要指出，$\gamma_1(n)$ 会随着 β_1缓慢增加。这一性质在图1 (b)中有所展示：当 β_1 增加时，蓝色区域的下边界会略微上升。请注意，我们给出的 β_2 阈值是收敛的充分条件，并未声称其幂次是紧的。进一步收紧 β_2 的阈值将是一个有趣的未来研究方向。

备注 2： β_2 **与批量大小**。我们发现，β_2 的条件：$\beta_2 \geq \gamma_1(n) = 1 - \mathcal{O}\left(\frac{1-\beta_1^n}{n^5}\right)$ 是与具体问题相关的，并且随 n 增大而增大。这一性质表明，我们需要更大的 β_2，当 n 较大时；或者等价地，我们需要更大的 β_2，当批量大小较小时 [3]。这与我们在图 3 (a) 中的实验一致：在 MNIST 上，较小的批量大小 $\mathcal{D}_n$（等价于较大的 n）需要更大的 β_2，才能达到较小的损失。

备注 3：关于选择 β_1。定理 3.1 要求 $\beta_1 < \sqrt{\beta_2}$。由于 β_2 建议取较大值，我们的收敛结果可以涵盖灵活选取的 $\beta_1 \in [0, 1)$。例如，$\beta_2 = 0.999$ 使得阈值达到$\beta_1 < 0.9995$，这基本上覆盖了所有实际选择 β_1（文献中所报告的），包括默认设置 $\beta_1 = 0.9$。我们的理论与我们在 MNIST 和 CIFAR-10 上的网格搜索实验相一致（图1（c,d），见第 1节），其中 Adam 在很宽的取值范围 β_1 上表现良好，当 β_2 较大时。

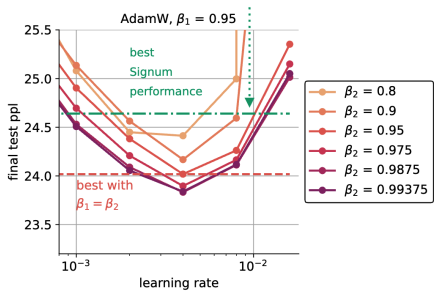
备注 4：对大语言模型预训练的指导。我们的理论表明，当 n 较大时，需要更大的 β_2。这等价地表明，当批量大小（等于 $\mathcal{D}_n$）较小时，需要更大的 β_2。我们随后的发散理论（将在后文给出）进一步表明，为避免发散，有必要使 β_2 依赖于批量大小。综合而言，这些结论可为大语言模型预训练中的超参数调优提供指导，并已得到多篇文献的证实。我们列出一些数值证据如下。

- Zhao 等人 [2024], Orvieto 和 Gower [2025] 报告称，若两者 (β_1, β_2) 都接近 0，则训练失败；而当使用 $\beta_1 = \beta_2 = 0.95$ 或 0.975 时，性能会显著提升。这一发现与我们的结果一致。我们在图 4 (a) 中重述了他们的部分数值结果。
- Sre´ckovi´c 等人 [2025] 指出*"Zhang 等人 [2022b] 表明，更高的 β_2 值会显著改善小批量训练，并且 Marek 等人 [2025] 强调了缩放 β_2 在该情形下的重要性。"* 我们在图 4 (b) 和 (c) 中重述了他们的部分数值结果。
- Zhang 等人 [2024强调] 指出，*"增大 β_2（从 0.95 提高到 0.99、0.999 或 0.9995）可以显著改善小批量大小下的训练……这与 [Zhang 等人、2022b]的发现一致。"*我们在图 4（d、e）中重现了他们的一些数值结果。
- Porian 等人 [2024] 报告称，*"在较小的批量大小下，增大 β_2（从 0.95 提高到 0.99 和 0.999）至关重要……这与 [Zhang 等人、2022b]的理论研究相符。"*

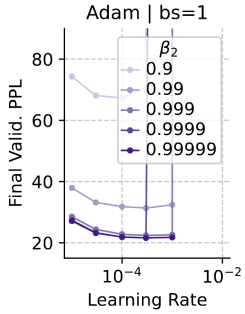
注 5：收敛到临界点的邻域。当$D_0 > 0$时，Adam 收敛到临界点的邻域，而不是精确的临界点。我们强调，这并非分析局限性所致；实践中也观察到了这一现象：即使对于具有$D_0 > 0$的简单凸二次函数，采用递减步长的 Adam 也无法达到梯度恰好为零（见图 3（b））。我们在附录 F中给出了该不可实现函数。

在不可实现情形（$D_0 > 0$）中，随机梯度方法收敛到临界点邻域的情况很常见，包括恒定步长 SGD [Luo, 1991, Bertsekas, 1997, Yan 等人, 2018, Yu 等人, 2019, Liu 等人, 2020b] 以及递减步长 RMSProp [Zaheer 等人, 2018, Shi 等人 , 2020]。这是因为：尽管 η_k 在减小，有效步长 $\eta_k/\sqrt{v_k}$ 却可能不会衰减。好消息是，邻域的大小 $\mathcal{O}(\delta(\beta_2)\sqrt{D_0})$ 会随着 β_2 趋近于 1 而趋于 0（理论分析和图 3(b) 中的实验均显示如此）。这一点可以从附录 6 中公式 (6.46) 的 $\delta(\beta_2)$表达式看出。邻域大小之所以会缩小至 0，是因为随着 β_2 增大，$\sqrt{v_k}$ 的变化幅度会减小。

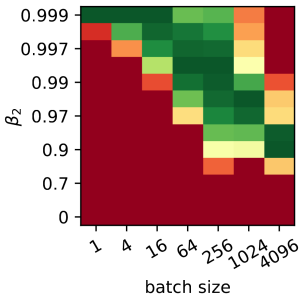
[3]这是因为批量大小等于 $\mathcal{D}_n$，其中$\mathcal{D}$ 表示样本总数，而n 表示数据集被划分成的小批次数量。



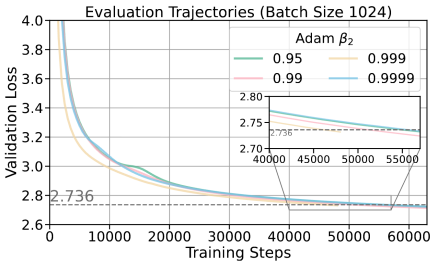
(a) 图 12 引自 [Orvieto 和 Gower， 2025]



(b) 图 5 引自 [Sre´ckovi´c 等人， 2025]



(c) 图4 引自 [Marek 等人， 2025]



(b) β_2 Ablation

(d) 图 8 (b) 引自 [Zhang 等人，2024a]

Batch Size	β_2	τ
1.2B		
64	0.999	0.9995
128	0.999	0.9995
256	0.995	0.9995
512	0.99	0.9995
1024	0.99	0.999
2048	0.95	0.998
4096	0.95	0.995
8192	0.95	0.99
16384	0.95	0.99

(e) 表 4 引自 [Zhang 等人，2024a]

图 4：近期文献中关于 β_2 对大语言模型预训练的影响。**(a,b,c)**：在不同 β_2 和批量大小下训练的大语言模型的最终验证损失。**(c):** 颜色越绿表示验证损失越低。**(d, e):** 在不同批量大小下训练一个拥有 1.2B 个参数的大语言模型时的最优 β_2。此处， τ 代表其他与我们的讨论无关的训练技巧。这些结果得出了一致的结论：较大的 β_2 有助于提升性能，并且在小批量大小情形下应调高。这些结果证实，我们的理论为大语言模型预训练中的超参数调优提供了有效指导。

作为定理3.1的推论，我们在可实现情形下（即$D_0 = 0$）得到以下结果。

推论 3.2. 在定理*3.1*的设定下。当 $D_0 = 0$对于假设2.2时，我们有

$$\min_{k\in[1,T]} \mathbb{E}\|\nabla f(x_k)\|_2 = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right).$$

在可实现情形下（即 $D_0 = 0$），推论 3.2 表明 Adam 可以收敛到临界点。这与我们在图 3 (c) 中的数值结果一致。定理3.1 和推论3.2 中的收敛速率与相同条件下的随机梯度下降相当，见 [Vaswani 等人， 2019]。

与定理 3.1 类似，我们现在给出 Adam 在随机打乱下的收敛结果（算法 2）。上述所有注释同样适用于随机打乱情形。

定理 3.3.（算法2的收敛结果） 假设算法2满足： $\beta_1 < \sqrt{\beta_2} < 1$； β_2 大于或等于阈值 $\gamma_2(n) := 1 - \mathcal{O}\left(\frac{1-\beta_1^n}{n^{5.5}}\right)$；且 $\eta_k = \frac{\eta_0}{\sqrt{k}}$。对于任意 $f(x) \in \mathcal{F}_{nL,D_0,D_1}(\mathbb{R}^d)$，当 $T \in \mathbb{N}$足够大时，我们有：

$$\min_{k\in[1,T]} \mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_{k,0})\|_2^2}{\sqrt{D_0}}, \frac{\|\nabla f(x_{k,0})\|_2}{2\sqrt{dD_1}}\right\}\right] = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right) + \mathcal{O}(\tilde{\delta}(\beta_2)\sqrt{D_0}),$$

其中 $\tilde{\delta}$（β_2） 是一个正数常量，当 β_2 趋近于 *1* 时，该常量趋近于 *0*。

作为一个推论 由定理3.3的推论，在可实现情形下（即$D_0 = 0$），我们有以下结果。

推论 3.4. 在定理 3.3 的设定下，当 $D_0 = 0$满足假设 2.2时，我们有

$$\min_{k\in[1,T]} \mathbb{E}\|\nabla f(x_{k,0})\|_2 = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right).$$

T 的证明　　定理 3.1 和定理3.3 分别见第 5 – 7 节及附录 D。　　y.

3.2 发散结果

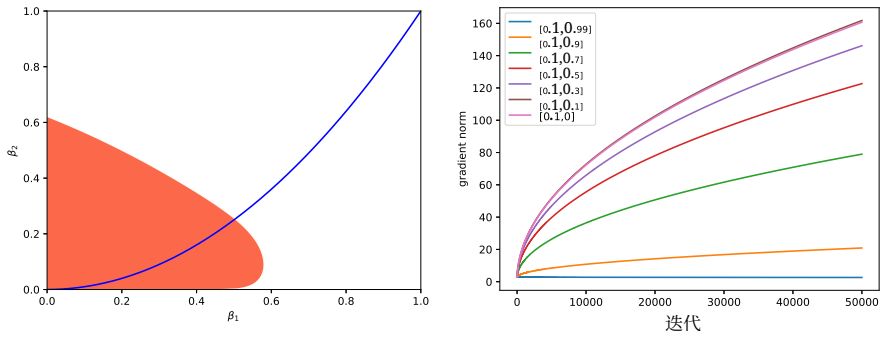
现在我们证明，小型β_2 Adam 可能发散，而且发散区域取决于问题。小型β_2 Adam 的发散表明，“较大的 β_2”是定理 3.1的必要条件。

我们在 $\mathcal{F}_{nL,D_0,D_1}(\mathbb{R}^d)$ 中构造一个反例。考虑 $f(x) = {}^1_n \sum^{n-1}_{i=0} f_i(x)$，对于 $n \geq 3$ 和 $x \in \mathbb{R}$，我们将 $f_i(x)$ 定义为：

$$f_0(x) = \begin{cases} (1+(n-1)a)\,x, & x \geq -1 \\ \frac{(1+(n-1)a)}{2}(x+2)^2 - \frac{3n}{2}, & x < -1 \end{cases}, \quad f_i(x) = \begin{cases} -ax, & x \geq -1 \\ -\frac{a}{2}(x+2)^2 + \frac{3}{2}, & x < -1 \end{cases} \quad \text{for } i > 0, \tag{3.1}$$

其中 a > 0。将所有 f_i(x) 求和后，可以看出：对于任意有限正数 a > 0，f (x) 属于$\mathcal{F}$nL,D_0,D_1($\mathbb{R}^d$)。例如，当 a $= 1/(n-1)^2$,时，我们有 $D_1 = 2n^2, D_0 = 0$，且 f (x) 有下界，最优解为 $x^* = -2$。此外，它满足假设 2.2，但不满足有界方差 (2.3)，后者将 D_1 限制为 n。函数 (3.1) 是在 [Reddi 等人 2018]的基础上修改而来，但它允许迭代点和梯度都发散到无穷大。我们在定理 3.5 中给出发散结果。我们注意到，符号随机梯度下降是 Adam 在 $\beta_1 = \beta_2 = 0$ 下的特殊情形，因此以下发散结果同样适用于符号随机梯度下降。

定理 3.5. 对于任意满足 $n \geq 3$、$D_1 \geq 2n^2$, 且 $D_0 \geq 0$ 的问题类 $\mathcal{F}nL,D_0D_1(\mathbb{R}^d)$，存在一个 $f(x) \in \mathcal{F}nL,D_0D_1(\mathbb{R}^d)$，使得当 (β_1, β_2) 满足第 8 节中的解析条件 *(8.3)*、*(8.4)*、*(8.5)* 时，*Adam* 的迭代点、迭代点的梯度以及迭代点的函数值都发散至无穷大。该区域的大小取决于 n，且随着 n 趋于无穷大，该区域会扩展至整个区域 $[0,1]^2$。

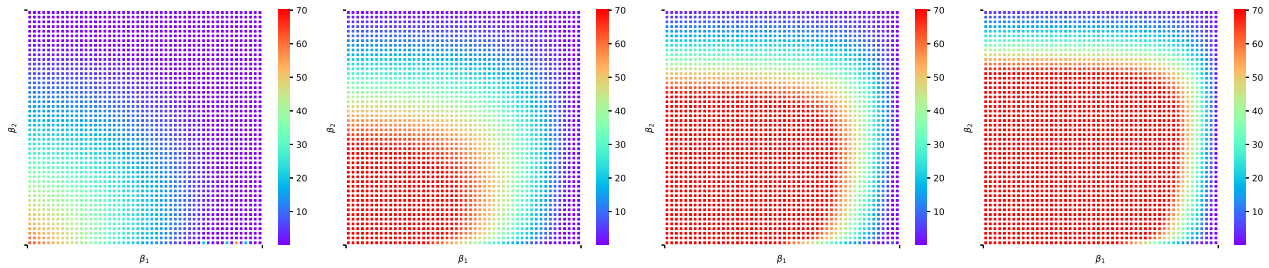


(a) 发散区域　　(b) 小 β_2 的影响

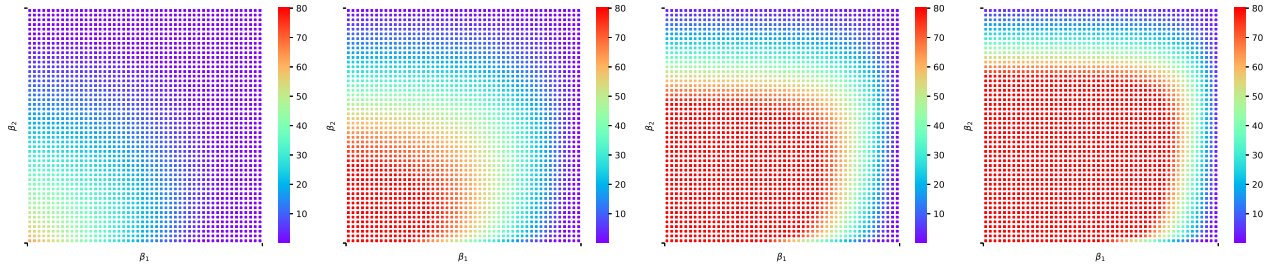
图 5：**(a):** 对于函数（3.1），当 $n = 20$ 且 $a = 1$ 时，Adam 在彩色区域内发散。该区域通过在 `NumPy` 中求解条件（8.3）、（8.4）、（8.5）绘制。蓝色曲线满足 $\beta_1 = \sqrt{\beta_2}$。**(b):** 当 β_2 较小时，Adam 会发散。我们使用函数（3.1），其初始化为 $x = -5$，且 $n = 20$。(b) 中的标签表示 $[\beta_1, \beta_2]$。

证明见第 8 节。我们发现，“发散区域”取决于具体问题。此外，它始终低于定理 $\gamma_1(n)$ 中的“收敛阈值” 3.1，因此这两个结果彼此一致（参见第 8 节的注释）。数值上也观察到了小型 β_2 Adam 的发散现象（见图 5(b) ）。这些结果从数值和理论两个方面刻画了 Adam 的发散行为。

在图 6 中，我们给出了 （β_1，β_2）在反例（3.1）上的网格搜索。我们发现，Adam 的行为与我们的理论预测一致：当初始化于函数（3.1）的线性侧时，即 $x \geq -1$，迭代点在 β_1 和 β_2 都较小的情况下会持续向右移动，从而导致发散。发散区域随 n 扩大。类似地，当初始化于二次区域时，即 $x < -1$，迭代点会持续向左移动，因此梯度范数也会发散（这也显示在图 5 (b) 中）。



(a) $n = 5$ (b) $n = 10$ (c) $n = 15$ (d) $n = 20$



(e) $n = 5$ (f) $n = 10$ (g) $n = 15$ (h) $n = 20$

图 6： **(a) - (d)：** 在 Adam 对（$|x - x^*|$ 3.1）运行 5 万次迭代后的最优性差距。我们使用初始化 $x = 1$。**(e-h)：** 在 Adam 对（3.1）运行 5 万次迭代后的梯度范数。我们使用初始化 $x = -5$。

最后，我们强调，图5(a) 中的橙色区域并未在 [Reddi 等人中讨论， 2018] 因为我们将n 视为固定，而他们允许n 变化。借助定理3.1与定理3.5，我们更清晰地刻画了(β_1, β_2) 与 Adam 定性行为之间的关系。

注释 7：关于 SignSGD 的发散。 定理3.5 意味着 SignSGD（带有 $\beta_1 = \beta_2 = 0$ 的 Adam）可能会发散。通过修改算法可以避免发散，例如添加动量 [Sun 等人、2023年、Jiang等人、 2025]；或引入更强的假设，例如（2.3）中的有界方差条件并增大批量大小 [Bernstein 等人、 2018]。事实上，仅有有界方差假设就足以将我们的反例排除在所考虑的函数类之外；在这种情况下，SignSGD 将收敛到临界点的邻域，收敛速率为 $\eta_k = \eta_0/\sqrt{k}$ [4]。然而，有界方差条件下的分析并未揭示 SignSGD 真正的发散行为。我们的反例满足假设2.2（其中$D_1 \geq 2n^2$），但不满足有界方差条件（2.3）（其中$D_1 = n$），这进一步促使我们放宽（2.3）。一个引人关注的开放问题是，是否还存在满足$D_1 \in (n、 2n^2)$的其他反例。我们将其留待未来研究。

注释 8：对临界边界进行更细致的刻画。综合上述讨论的发散理论与收敛理论，可以得出结论：至少存在一条临界边界(β_1^*, β_2^*) 用于划分双收敛相变。由于发散区域和收敛区域都取决于n （或者说，取决于批量大小 $\mathcal{D}_n$），临界边界(β_1^*, β_2^*) 也必然取决于批量大小，且位于图1 (b) 中的白色区域。需要注意的是，我们的理论表明临界边界存在，但尚未完全确定边界的确切数量和形状。尽管如此，图1 (c, d) 中的实验结果表明，可能只有一条边界，其形状很可能与图1 (b) 中蓝色区域的几何形状相似。我们在此仅指出其存在性，并将更精确的刻画留作未来研究方向。

[4]据我们所知，文献中尚未有在不增大批量大小的情况下证明 SignSGD 在有界方差条件（2.3）下收敛的简洁证明，但我们预计证明并不复杂。若将（2.3）放宽为我们的假设 2.2，则无法保证收敛。

4 定理3.1收敛结果的关键引理

此处，我们总结定理3.1证明中的关键难点。定理3.1 的证明见第7节。定理3.3 的证明采用类似思路，具体证明见附录D。

补充符号。对于一个 $\in \mathbb{R}$，我们用$\lceil a \rceil$ 表示向上取整（将 a 向上取整）。对于一个 $\in \mathbb{N}$，我们将 $[\ ]$ $\{1, \cdots, \}$ $1\ 1$ $[\ ] = a$ 记为指标集 $\{0, \cdots, -1\}$ ∂_l $\partial_{l\ i}$ ∇ $\nabla_{\ i}$。对于（.）中的函数索引 n，我们滥用记号 n n。我们分别用 f (x) 和 f (x) 表示 f (x) 和 f (x) 的第 l 个分量。我们用 $x_{l,k}$ 表示第 k 次迭代时 x 的第 l 个分量，即 x_k。$v_{l,k}$ 和 $m_{l,k}$ 也类似。我们用 $\mathbb{I}(\cdot)$ 表示示性函数。 $\mathbb{E}(\cdot)$ 表示对整个轨迹取期望。对于 $k \geq 1$, $\mathbb{E}_k[\cdot]$，表示给定截至第 $k-1$ 次迭代结束时的全部历史的条件期望（即给定 x_1, m_0, v_0 及所有过去的采样索引 $\tau_1, \ldots, \tau_{k-1}$），但不包括新鲜样本 τ_k。我们如下滥用 α 的记号：对于 $\partial_l f_\alpha(x)$，定义 $\alpha := \arg\max_i |\partial_l f_i(x)|$; 对于 $\partial_\alpha f(x)$，定义 $\alpha := \arg\max_l |\partial_l f(x)|$。$m_{\alpha,k}$ 和 $v_{\alpha,k}$ 也类似。我们用“i.i.d.”和“r.v.”分别作为“独立同分布”和“随机变量”的缩写。

证明的核心是刻画 Adam $\mathbb{E}\left(m\sqrt[k]{v_k}\right)$ 的期望更新方向何时构成下降方向，即何时该方向位于梯度 $\nabla f(x_k)$的对偶锥中。这里至少存在两方面的挑战。**首先**，v_k 是一个随机变量，而且出现在分母中。这使得整个系统成为一个随机非线性动力系统，通常难以分析。此外，v_k 可能会变为 0，这也带来了额外的困难。**其次**，m_k 包含大量历史信号，这些信号会使轨迹偏离梯度方向。

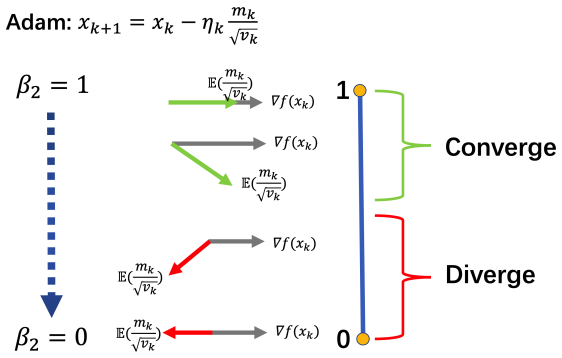


图 7: β_2 变化时 Adam 更新方向变化的示意图。

我们的关键见解如图 7所示：我们发现，当 $\mathbb{E}\left(m\sqrt[k]{v_k}\right)$ 较大时，$\nabla f(x_k)$ 较为接近，从而实现收敛；而当 β_2 较小时，更新方向开始偏向相反方向，导致发散。现在我们介绍收敛结果的证明。发散结果的证明将在后文第 3.2 节中给出。

步骤 I: $\sqrt[1]{v_k}$ 较大时的集中效应。~~我~~我们发现，当β_2 较大时，$\sqrt[1]{v_k}$ 集中在~~$\sqrt{}$~~$\sqrt[1]{\mathbb{E}_k(v_k)}$ 附近。具体而言，~~我们证~~明：

$$\frac{1}{\sqrt{v_k}} \approx \frac{1}{\sqrt{\mathbb{E}_k(v_k)}}, \text{ w.h.p., when } \beta_2 \text{ is large}$$

凭借这种集中性质，$\sqrt[1]{v_k}$ 得以稳定。~~**为什么较大的 β_2 会有帮助?** 直观地说，这是因为较大的~~β_2会减缓v_k的变化，使其行为变得相当可预测。因此，在这种情况下，$-\nabla_{f^{\tau_k}(x_k)\sqrt{v_k}}$会成为一个下降方向，即：

$$\mathbb{E}_k\left(\frac{\nabla f_{\tau_k}(x_k)}{\sqrt{v_k}}\right) \approx \frac{\mathbb{E}_k\left(\nabla f_{\tau_k}(x_k)\right)}{\sqrt{\mathbb{E}_k(v_k)}} = \frac{\nabla f(x_k)}{\sqrt{\mathbb{E}_k(v_k)}},$$

它位于梯度方向 $\nabla f_1(x_k)$的对偶锥中。

比较 $\frac{1}{\sqrt{v_k}}$ 与 $\sqrt{\mathbb{E}_k(v_k)}$需要付出相当大的努力。在统计理论中，通常会证明随机变量 x 集中于其均值 $\mathbb{E}(x)$附近，但刻画 $\frac{1}{\sqrt{x}}$ 偏离 $\frac{1}{\sqrt{\mathbb{E}(x)}}$ 的程度则少见得多，尤其是当 x 可以任意接近 0，且未施加额外的有界性条件时。若对随机变量 x没有任何特定性质，$\sqrt{x}$ 中的任何微小偏差都会在 $\frac{1}{\sqrt{x}}$ 中被放大，使其在零附近表现得极不稳定，并可能与 $\frac{1}{\sqrt{\mathbb{E}(x)}}$ [Durrett 相距任意远， 2019]。

在引理 4.1 中，我们证明，当 $\frac{1}{\sqrt{v_k}}$ 足够大时，它会集中于 $\frac{1}{\sqrt{\mathbb{E}_k(v_k)}}$ 附近 β_2，而且该结果不需要对随机梯度施加任何有界性条件。我们利用 Adam 的两个特殊性质证明了这一结果：1）随机梯度在v_k中具有“几何级数”结构。2）随机梯度的索引是从一个有限指标集中均匀抽取的。利用这两个特殊性质，我们发现，当 β_2 足够大时，$\frac{1}{\sqrt{v_k}}$的表现基本上是可预测的。

严格来说，集中性结果通过两个步骤得到：(i) 将$1/\sqrt{v_k}$ 的动态过程映射为一列独立同分布的伯努利随机变量的动态过程；该序列有界，因此更易于分析；(ii) 通过若干解耦步骤将动态过程映射回来。步骤 (i) 和 (ii) 使我们能够使用有界的伯努利代理变量来追踪可能无界的随机变量序列的动态过程。严格的集中效应如下所示于引理 4.1。引理4.1可用作分析 Adam 类算法的通用工具。

引理 *4.1.* 假设算法 *1* 满足：$\beta_2 \geq \frac{1}{2}$; $0 \leq \beta_1 < \sqrt{\beta_2} < 1$; β_2满足 $\frac{(1-\beta_2)^2}{\beta_2^n} < \frac{1}{8}n - \frac{\delta}{4}$; 且 $\eta_k = \frac{\eta_0}{\sqrt{k}}$。对于任意 $f(x) \in \mathcal{F}nL,D_0,D_1(\mathbb{R}^d)$，以及任意 $0 < \delta \leq \frac{1}{4}n$，令 $k \geq \lceil log(n\delta) log_{\beta_2} \rceil + n + 1$。定义阈值 $R_k := \frac{16\sqrt{2}\Delta_1}{\sqrt{k}\left(\lceil \frac{\log(n\delta)}{\log \beta_2} \rceil + n\right)}$，其中 Δ_1在 *(B.1)* 中定义。对于任意索引 $l \in [d]$，若 $max_i |\partial_l f_i(x_k)| \geq R_k$，则其概率至少为

$$1 - n\exp\left(-\frac{\delta^2}{(1-\beta_2)(\frac{28}{3n} + \frac{8}{3}\delta)}\right),$$

我们有如下集中界：

$$\frac{C_{lower}}{\sqrt{\mathbb{E}_k(v_{l,k})}} \leq \frac{1}{\sqrt{v_{l,k}}} \leq \frac{C_{upper}}{\sqrt{\mathbb{E}_k(v_{l,k})}}, \tag{4.1}$$

其中常数定义如下：

$$C_{lower} := 1 - (1-\beta_2)\frac{4n}{(1-2n\delta)\beta_2^n}, \quad C_{upper} := \left(1 - (1-\beta_2)\frac{8n}{(1-2n\delta)\beta_2^n}\right)^{-1/2}.$$

引理 4.1 的证明见第5节。根据引理 4.1，我们有 $\frac{1}{\sqrt{v_k}} \approx \frac{1}{\sqrt{\mathbb{E}_k(v_k)}}$，其中 $\beta_2 \to 1$。

步骤 II：用于处理 m_k。注意，仅凭 $\frac{1}{\sqrt{v_k}} \approx \frac{1}{\sqrt{\mathbb{E}_k(v_k)}}$ 的集中效应，尚不足以证明 Adam 的收敛。这是因为m_k还包含大量历史梯度信号，它们会进一步使更新方向偏离梯度方向。换言之，尚不清楚 $\frac{\mathbb{E}_k(m_k)}{\sqrt{\mathbb{E}_k(v_k)}}$ 是否接近 $\nabla f(x_k)$。

我们使用势函数 $f(z_k)$ 来抵消m_k的影响，其中$f(\cdot)$ 是问题（1.1）中的原函数，且$z_k := \frac{x_k - \beta_1^n x_{1-\beta_1^n}^{k-n}}{}$。辅助序列$z_k$ 有助于抵消截至$(k-n)$次迭代的所有历史信号，从而使收敛分析更加便捷。

$$\begin{aligned} z_k - z_{k+1} &= \frac{x_k - x_{k+1} - \beta_1^n(x_{k-n} - x_{k+1-n})}{1-\beta_1^n} = \frac{1}{1-\beta_1^n} \cdot \left(\eta_k \frac{m_k}{\sqrt{v_k}} - \beta_1^n \eta_{k-n} \frac{m_{k-n}}{\sqrt{v_{k-n}}}\right) \\ &\overset{\text{Lemma 4.1}}{\approx} \frac{\eta_k}{1-\beta_1^n} \cdot \frac{\nabla f_{\tau_k}(x_k) + \beta_1 \nabla f_{\tau_{k-1}}(x_{k-1}) + \cdots + \beta_1^{n-1} \nabla f_{\tau_{k-n}}(x_{k-n})}{\sqrt{\mathbb{E}_k(v_k)}} \end{aligned}$$

注意，p 势函数的灵感来自以下L_k 随机梯度下降分析 [Liu 等人，2020b]:

$$L_k = (f(\widetilde{z}_k) - f^\star) + \sum_{i=1}^{k-1} c_i \|x_{k+1-i} - x_{k-i}\|^2, \quad \widetilde{z}_k = \frac{x_k - \beta_1 x_{k-1}}{1-\beta_1}, \quad c_i > 0$$

我们发现，构造L 和 $\widetilde{z}_k$ 这一思路对 Adam 分析很有用，因此我们做出如下修改。

- **修改 I：** 对于 Adam，我们发现 L_k 可以通过设置 $c_i = 0$来简化。这是因为引入 c_i 的目的是帮助界定更新幅度 $m_k/\sqrt{v_k}$，而对于 Adam，该值本身有界（稍后将在附录B.2 引理中证明6）。

- **修改 II：** 我们将 $\widetilde{z}_k$ 替换为 z_k，区别在于保留了$(n-1)$ 个额外的历史梯度信号。此修改扩大了(β_1, β_2)图中的收敛区域：使用 $\widetilde{z}_k$时，收敛要求 $\beta_2 \geq 1 - \mathcal{O}\left(\frac{1-\beta_1}{n^5}\right)$；而使用 z_k时，要求放宽为 $\beta_2 \geq 1 - \mathcal{O}\left(\frac{1-\beta_1}{n^5}\right)$，这定义了一个严格更大的区域。

借助引理 4.1 和势函数 $f(z_k)$，我们证明 Adam 的更新方向接近$\nabla f(x_k)$ 当 β_2 足够大时，从而可以导致收敛。我们发现，由于动量的存在、可能的无界梯度以及多种耦合效应，将引理 4.1 中的集中性结果应用于收敛证明相当非平凡。我们在引理 4.2中提出了一种系统化的方法来克服这些问题。

引理 4.2. 考虑与引理 *4.1* 相同的条件；当 $k \in \mathbb{N}$ 满足 $k \geq \lceil \log_{1/4} \log \beta_2 \rceil + 2n + 1$ 且 $C\sqrt{k} \leq \frac{1}{4} nd\sqrt{5D_1 n}$ 且 $\beta_1^{k-n} \leq \frac{1}{\sqrt{k}}$ 时，我们有

$$\mathbb{E}\langle \nabla f(z_k), z_k - z_{k+1} \rangle \geq \mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_k)\|_2^2}{nd\sqrt{5D_0 nd}}, \frac{\|\nabla f(x_k)\|_2}{2nd^2\sqrt{5D_1 n}}\right\}\right] - \delta(\beta_2)\sqrt{D_0} - \frac{C}{\sqrt{k}},$$

其中常数 $C > 0$ 和 $\delta(\beta_2)$ 是一个常数，当 β_2 趋近于 *1* 时它趋近于 *0*（见(6.46) 附录 *6*）。引理 *4.2* 的证明见第 *6* 节，定理*3.1* 的完整证明见第 *7*节。

引理 5 的证明4.1

证明分三步进行。在步骤 I 中，我们通过比值 $\max_i \partial_{l\ i\ k}^{\ 2}(f(x))(v,)$，推导出将 $\frac{1}{\sqrt{v_{l,k}}}$ 与 $\frac{1}{\sqrt{\mathbb{E}_k(v_{l,k})}}$ 联系起来的确定性代数界。在步骤 II 中，我们利用 Adam 的几何和结构，建立 $\mathbb{E}_k(v_{l,k})$ 的下界，从而得到该比值的上界。在步骤 III 中，我们应用伯恩斯坦不等式证明几何级数的下界以高概率成立，由此建立集中性结果。

步骤 I：确定性代数界，联系 $1/\sqrt{v_{l,k}}$ 与 $1/\sqrt{\mathbb{E}_k(v_{l,k})}$。 我们首先建立 $\frac{1}{\sqrt{v_{l,k}}}$ 与 $\frac{1}{\sqrt{\mathbb{E}_k(v_{l,k})}}$ 之间的关系。

对于下界，我们有

$$\begin{aligned}
\frac{1}{\sqrt{v_{l,k}}} &= \frac{1}{\sqrt{\mathbb{E}_k(v_{l,k}) + (v_{l,k} - \mathbb{E}_k(v_{l,k}))}} \\
&\geq \frac{1}{\sqrt{\mathbb{E}_k(v_{l,k})}} \cdot \frac{1}{\sqrt{1 + \frac{|v_{l,k} - \mathbb{E}_k(v_{l,k})|}{\mathbb{E}_k(v_{l,k})}}} \\
&\geq \frac{1}{\sqrt{\mathbb{E}_k(v_{l,k})}} \left(1 - \frac{|v_{l,k} - \mathbb{E}_k(v_{l,k})|}{2\mathbb{E}_k(v_{l,k})}\right),
\end{aligned} \tag{5.1}$$

其中最后一个不等式使用了 $\frac{1}{\sqrt{1+x}} \geq 1 - x/2$ 当 $x \geq 0$。

接下来，我们对 $|v_{l,k} - \mathbb{E}_k(v_{l,k})|$ 进行界定。根据 $v_{l,k}$ 的更新规则，我们有

$$|v_{l,k} - \mathbb{E}_k(v_{l,k})| = (1-\beta_2)\left|(\partial_l f_{\tau_k}(x_k))^2 - \mathbb{E}_k\left[(\partial_l f_{\tau_k}(x_k))^2\right]\right| \leq 2(1-\beta_2)\max_i (\partial_l f_i(x_k))^2. \tag{5.2}$$

将 (5.2) 代入 (5.1)，得到

$$\frac{1}{\sqrt{v_{l,k}}} \geq \frac{1}{\sqrt{\mathbb{E}_k(v_{l,k})}}\left(1 - (1-\beta_2)\frac{\max_i (\partial_l f_i(x_k))^2}{\mathbb{E}_k(v_{l,k})}\right). \tag{5.3}$$

对于上界，我们采用类似的方法：

$$\begin{aligned}\frac{1}{\sqrt{v_{l,k}}} &= \frac{1}{\sqrt{v_{l,k} - \mathbb{E}_k(v_{l,k}) + \mathbb{E}_k(v_{l,k})}} \\ &= \frac{1}{\sqrt{\mathbb{E}_k(v_{l,k})}} \cdot \frac{1}{\sqrt{1 + \frac{v_{l,k} - \mathbb{E}_k(v_{l,k})}{\mathbb{E}_k(v_{l,k})}}},\end{aligned} \tag{5.4}$$

为了从 (5.4) 得到有意义的上界，我们需要确保分母始终为正且远离零。特别是，我们希望满足

$$\left|\frac{v_{l,k} - \mathbb{E}_k(v_{l,k})}{\mathbb{E}_k(v_{l,k})}\right| < 1. \tag{5.5}$$

根据 (5.2)，若满足以下条件，则条件 (5.5) 成立

$$2(1-\beta_2)\frac{\max_i (\partial_l f_i(x_k))^2}{\mathbb{E}_k(v_{l,k})} < 1. \tag{5.6}$$

假设 (5.6) 成立，则有 $1 + \frac{v_{l,k} - \mathbb{E}_k(v_{l,k})}{\mathbb{E}_k(v_{l,k})} \geq 1 - \left|\frac{v_{l,k} - \mathbb{E}_k(v_{l,k})}{\mathbb{E}_k(v_{l,k})}\right| > 0$。这使我们能够对倒数平方根进行界定：

$$\begin{aligned}\frac{1}{\sqrt{1 + \frac{v_{l,k} - \mathbb{E}_k(v_{l,k})}{\mathbb{E}_k(v_{l,k})}}} &\leq \frac{1}{\sqrt{1 - \left|\frac{v_{l,k} - \mathbb{E}_k(v_{l,k})}{\mathbb{E}_k(v_{l,k})}\right|}} \\ &\overset{(5.2)}{\leq} \frac{1}{\sqrt{1 - 2(1-\beta_2)\frac{\max_i (\partial_l f_i(x_k))^2}{\mathbb{E}_k(v_{l,k})}}}.\end{aligned} \tag{5.7}$$

结合（5.4）和（5.7），我们得到如下结果（假设（5.6）成立）：

$$\frac{1}{\sqrt{v_{l,k}}} \leq \frac{1}{\sqrt{\mathbb{E}_k(v_{l,k})}} \cdot \frac{1}{\sqrt{1 - 2(1-\beta_2)\frac{\max_i (\partial_l f_i(x_k))^2}{\mathbb{E}_k(v_{l,k})}}}. \tag{5.8}$$

只有在能够界定比值 5.3）和（5.8) are useful only if we can bound the ratio $\frac{\max_i \partial_l f_i(x_k))^2}{\mathbb{E}_k(v_{l,k})}$ 并验证条件（5.6）。平凡界可立即得到：

$$\frac{\max_i (\partial_l f_i(x_k))^2}{\mathbb{E}_k(v_{l,k})} \leq \frac{n\mathbb{E}_k\left[(\partial_l f_{\tau_k}(x_k))^2\right]}{\mathbb{E}_k(v_{l,k})} \leq \frac{n\mathbb{E}_k\left[(\partial_l f_{\tau_k}(x_k))^2\right]}{(1-\beta_2)\mathbb{E}_k\left[(\partial_l f_{\tau_k}(x_k))^2\right]} = \frac{n}{1-\beta_2}, \tag{5.9}$$

它随 $\beta_2 \to 1^-$ 发散。将其代入（5.6）将要求 $(1-\beta_2) \cdot 2n_{1-\beta_2} = 2n < 1$，而这对任何 $n \geq 1$ 都不成立。因此，平凡界不够用。

我们的目标是推导出一个精细界，使得

$$(1-\beta_2)\frac{\max_i (\partial_l f_i(x_k))^2}{\mathbb{E}_k(v_{l,k})} \to 0 \quad \text{as } \beta_2 \to 1^-, \tag{5.10}$$

条件（5.6）对 β_2 充分接近 1 时成立，并且（5.3）和（5.8）仍然是紧的。要建立这样的界需要付出大量努力，我们将在接下来的步骤 II 和 III 中完成这项工作。

步骤 II：利用几何和结构给出下界 $\mathbb{E}_k(v_{l,k})$。**定义 :arg maxf**$_i(x_k)|$。展开 $\mathbb{E}_k\left(v_{l,k}\right)$，可得

$$\begin{aligned}\mathbb{E}_k\left(v_{l,k}\right) &= (1-\beta_2)\left(\mathbb{E}_k\left[\left(\partial_l f_{\tau_k}(x_k)\right)^2\right]+\sum_{j=1}^{k-1}\beta_2^j\left(\partial_l f_{\tau_{k-j}}(x_{k-j})\right)^2\right)\\ &\geq (1-\beta_2)\sum_{j=1}^{k-1}\beta_2^j\left(\partial_l f_{\tau_{k-j}}(x_{k-j})\right)^2\\ &\geq (1-\beta_2)\sum_{j=1}^{k-1}\beta_2^j\left(\partial_l f_{\alpha}(x_{k-j})\right)^2\mathbb{I}\left(\tau_{k-j}=\alpha\right).\end{aligned} \tag{5.11}$$

注意，当 $k \geq \lceil \frac{\log(n\delta)}{\log \beta_2} \rceil + n + 1$时，引理 4.1 的条件蕴含引理 B.3 对$j \leq \lceil \frac{\log(n\delta)}{\log \beta_2} \rceil + n$的条件。应用引理 B.3，得到

$$\begin{aligned}\mathbb{E}_k\left(v_{l,k}\right) &\geq \frac{\left(\partial_l f_\alpha(x_k)\right)^2}{2}(1-\beta_2)\sum_{j=1}^{\lceil \frac{\log(n\delta)}{\log\beta_2}\rceil+n}\beta_2^j\mathbb{I}\left(\tau_{k-j}=\alpha\right)\\ &\geq \frac{\left(\partial_l f_\alpha(x_k)\right)^2}{2}(1-\beta_2)\beta_2^n\sum_{t=1}^{\lceil \frac{\log(n\delta)}{\log\beta_2}\rceil}\beta_2^t\mathbb{I}\left(\tau_{k-n-t}=\alpha\right).\end{aligned}$$

对于 $k^* < k - n$，定义辅助量

$$Y_{k^*,k}(i) := (1-\beta_2)\sum_{j=1}^{k^*}\beta_2^j\mathbb{I}\left(\tau_{k-n-j}=i\right), \quad \forall i \in \{0,1,\ldots,n-1\}, \tag{5.12}$$

它表示在第 $k-n$ 步之前的 k^* 步历史窗口内，索引 i 被采样次数的加权计数。由于 τ_{k-n-j} 是从 $\{0,1,\ldots,n-1\}$中均匀采样的，序列 $\{\mathbb{I}(\tau_{k-n-j}=i)\}_{j=1}^{k^*}$构成一列成功概率为 $\frac{1}{n}$ 的独立同分布的伯努利随机变量。

步骤 III：通过伯恩斯坦不等式及 k^* 的选取，对 $Y_{k^*,k}(i)$ 进行集中分析。我们的目标是证明，对于任意 $0 < \delta \leq \frac{1}{4n}$，随机变量 $Y_{k^*,k}(\alpha)$ 满足 $Y_{k^*,k}(\alpha) \geq \frac{1}{2n} - \delta$，且当 $\beta_2 \to 1^-$ 时，该结论成立的概率趋近于 1，其中 $\alpha := \arg\max_i |\partial_l f_i(x_k)|$。几何加权结构使集中分析成为可能：$Y_{k^*,k}(\alpha)$ 中的每一项都乘以系数 $(1-\beta_2)\beta_2^j$，其大小受 $(1-\beta_2)$ 控制。当 $\beta_2 \to 1^-$ 时，该系数趋于零，因此我们可以应用伯恩斯坦不等式，得到精确的概率界。然而，直接分析较为复杂，因为索引 $\alpha = \arg\max_i |\partial_l f_i(x_k)|$ 取决于随机迭代点 x_k。因此，α 是一个与优化轨迹相关联的随机变量，使得 $\mathbb{I}\left(\tau_{k-n-j}=\alpha\right)$ 的分布难以刻画。为此，我们首先对任意固定索引 $i \in \{0,1,\ldots,n-1\}$ 建立集中性，然后应用并集界。

定义这些事件

$$\mathcal{E}_k(i) := \left\{Y_{k^*,k}(i) \geq \frac{1}{2n} - \delta\right\}, \quad \forall i \in \{0,1,\ldots,n-1\}, \tag{5.13}$$

以及全局失败事件

$$\mathcal{U}_k := \bigcup_{i=0}^{n-1}\left\{Y_{k^*,k}(i) < \frac{1}{2n} - \delta\right\} = \bigcup_{i=0}^{n-1}\mathcal{E}_k(i)^c. \tag{5.14}$$

我们的目标是界定概率 $\mathbb{P}(\mathcal{E}_k(\alpha)^c)$。由于 α 取值于 $[n]$，，我们观察到失败

事件 α 包含在全局失败事件中：

$$\mathcal{E}_k(\alpha)^c = \left\{ Y_{k^*,k}(\alpha) < \frac{1}{2n} - \delta \right\} \subseteq \mathcal{U}_k.$$

根据并集界，

$$\mathbb{P}\left(\mathcal{E}_k(\alpha)^c\right) \leq \mathbb{P}\left(\mathcal{U}_k\right) \leq \sum_{i=0}^{n-1} \mathbb{P}\left(Y_{k^*,k}(i) < \frac{1}{2n} - \delta \right) \leq \sum_{i=0}^{n-1} \mathbb{P}\left(Y_{k^*,k}(i) \leq \frac{1}{2n} - \delta \right). \tag{5.15}$$

因此，只需对任意固定的i给出 $\mathbb{P}\left(Y_{k^*,k}(i) \leq \frac{1}{2n} - \delta\right)$ 的上界。

为便于表述，我们通过固定任意索引$i \in [n]$,来简化记号，去掉变量k，并将下标 τ_{k-n-j} 替换为 τ_j，重新定义

$$Y_{k^*} := (1-\beta_2) \sum_{j=1}^{k^*} \beta_2^j \mathbb{I}(\tau_j = i).$$

为应用伯恩斯坦不等式，我们引入中心化随机变量。定义

$$X_j := (1-\beta_2)\beta_2^j \left[\mathbb{I}(\tau_j = i) - \frac{1}{n} \right]$$

对于$j = 1, \ldots, k^*$。那么，对于所有j，$\mathbb{E}[X_j] = 0$ 且 $|X_j| \leq (1-\beta_2)\beta_2^j$。我们可以表示为

$$Y_{k^*} - \mathbb{E}\left[Y_{k^*}\right] = \sum_{j=1}^{k^*} X_j.$$

在应用伯恩斯坦不等式之前，我们先给Y_{k^*} 的期望一个下界，并给出其方差的上界。当 $\beta_2 \geq \frac{1}{2}$ 且k^* 足够大，使得 $\beta_2^{k^*} \leq n\delta$时，我们有

$$\mathbb{E}\left[Y_{k^*}\right] = (1-\beta_2) \sum_{j=1}^{k^*} \beta_2^j \mathbb{E}\left[\mathbb{I}(\tau_j = i)\right] = \frac{\beta_2(1-\beta_2^{k^*})}{n} \geq \frac{1}{2n} - \frac{\delta}{2}. \tag{5.16}$$

对于方差，由于X_j 是独立的，我们有

$$\begin{aligned} \operatorname{Var}(Y_{k^*}) &= \sum_{j=1}^{k^*} \operatorname{Var}(X_j) = (1-\beta_2)^2 \sum_{j=1}^{k^*} \beta_2^{2j} \operatorname{Var}\left(\mathbb{I}(\tau_j = i)\right) \\ &\leq \frac{(1-\beta_2)^2\beta_2^2}{1-\beta_2^2} \frac{1}{n}\left(1 - \frac{1}{n}\right) \leq \frac{(1-\beta_2)^2}{n(1-\beta_2^2)} \leq \frac{1-\beta_2}{n}. \end{aligned} \tag{5.17}$$

我们现在应用伯恩斯坦不等式。注意，对所有$j \in \{1, \ldots, k^*\}$，随机变量X_j 满足一致界$|X_j| \leq (1-\beta_2)\beta_2^j \leq 1-\beta_2$。应用伯恩斯坦不等式并利用方差界（5.17），对任意$a > 0$，我们得到：

$$\begin{aligned} \mathbb{P}\left(Y_{k^*} - \mathbb{E}\left[Y_{k^*}\right] \leq -a\right) &\leq \exp\left(-\frac{a^2/2}{\operatorname{Var}(Y_{k^*}) + \frac{1}{3}(1-\beta_2)a} \right) \\ &\leq \exp\left(-\frac{a^2/2}{\frac{1-\beta_2}{n} + \frac{1}{3}(1-\beta_2)a} \right). \end{aligned}$$

令 $a = \mathbb{E}\left[Y_{k^*}\right] - \left(\frac{1}{2n} - \delta\right)$，并注意到 $a \leq \frac{1}{2n} + \delta$。当 $a \geq \frac{\delta}{2}$ 时，对于 $\beta_2 \geq \frac{1}{2}$ 和 $k^* \geq \lceil \log(n\delta)/\log \beta_2 \rceil$，

我们有

$$\begin{aligned}\mathbb{P}\left(Y_{k^*} \le \frac{1}{2n} - \delta\right) &\le \exp\left(-\frac{a^2/2}{\frac{1-\beta_2}{n} + \frac{1}{3}(1-\beta_2)a}\right) \\ &\le \exp\left(-\frac{(\delta/2)^2/2}{\frac{1-\beta_2}{n} + \frac{1}{3}(1-\beta_2)\left(\frac{1}{2n}+\delta\right)}\right) \\ &= \exp\left(-\frac{\delta^2}{(1-\beta_2)\left(\frac{28}{3n}+\frac{8}{3}\delta\right)}\right). \end{aligned} \tag{5.18}$$

该集中界对任意固定索引$i \in [n]$均成立。将（5.18）代入并集界（5.15），可得

$$\mathbb{P}(\mathcal{E}_k(\alpha)^c) \le n \exp\left(-\frac{\delta^2}{(1-\beta_2)\left(\frac{28}{3n}+\frac{8}{3}\delta\right)}\right).$$

我们得到，以至少为 $1 - n\exp\left(-\frac{\delta^2}{1-\beta_2}\frac{()}{(\frac{28}{3n}+\frac{8}{3}\delta)}\right)$ 的概率，不等式 $Y_{k^*,\ k}(\alpha) \ge \frac{1}{2n} - \delta$ 成立。

令$k^* = \lceil \frac{\log(n\delta)}{\log \beta_2} \rceil$，并将该结果代入 $\mathbb{E}_k\left(v_{l\text{的下界}k}\right)$（5.12），我们得出，以至少为 $1 - n\exp\left(-\frac{\delta^2}{1-\beta_2}\frac{()}{(\frac{28}{3n}+\frac{8}{3}\delta)}\right)$ 的概率，

$$\mathbb{E}_k\left(v_{l,k}\right) \ge \frac{(\partial_l f_\alpha(x_k))^2}{2} \cdot \beta_2^n \cdot \left(\frac{1}{2n} - \delta\right). \tag{5.19}$$

回顾 $0 < \delta \le \frac{1}{4n}$，我们等价地得到

$$\frac{\max_i (\partial_l f_i(x_k))^2}{\mathbb{E}_k\left(v_{l,k}\right)} \equiv \frac{(\partial_l f_\alpha(x_k))^2}{\mathbb{E}_k\left(v_{l,k}\right)} \le \frac{4n}{(1-2n\delta)\beta_2^n}. \tag{5.20}$$

最后，我们验证当5.6中的条件对于 β_2 充分接近 1 时成立。将界（5.20）代入（5.6），我们要求

$$(1-\beta_2)2\frac{\max_i (\partial_l f_i(x_k))^2}{\mathbb{E}_k\left(v_{l,k}\right)} \le (1-\beta_2)\frac{8n}{(1-2n\delta)\beta_2^n} < 1. \tag{5.21}$$

整理（5.21），得到充分条件

$$\frac{1-\beta_2}{\beta_2^n} < \frac{1-2n\delta}{8n} = \frac{1}{8n} - \frac{\delta}{4}. \tag{5.22}$$

由于 $0 < \delta \le \frac{1}{4n}$，式（5.22）的右侧至少为 $\frac{1}{16n} > 0$。左侧 $\frac{1-\beta_2}{\beta_2^n}$ 关于 β_2 严格递减，并且当 $\beta_2 \to 1^-$ 时趋于零。因此，当 β_2 充分接近 1 时，式（$1/n^{1.5}$）成立（例如，取 $1-\beta_2 = \mathcal{O}(1/n^{1.5})$），从而验证了条件（5.6）和上界（5.8）。

将（5.20）代入集中界，我们以至少 $1 - n\exp\left(-\frac{\delta^2}{1-\beta_2}\frac{()}{(\frac{28}{3n}+\frac{8}{3}\delta)}\right)$ 的概率得到，

$$\frac{1}{\sqrt{\mathbb{E}_k\left(v_{l,k}\right)}}\left(1 - (1-\beta_2)\frac{4n}{(1-2n\delta)\beta_2^n}\right) \le \frac{1}{\sqrt{v_{l,k}}} \le \frac{1}{\sqrt{\mathbb{E}_k\left(v_{l,k}\right)}}\left(\frac{1}{\sqrt{1-(1-\beta_2)\frac{8n}{(1-2n\delta)\beta_2^n}}}\right).$$

由于 l 的选择是任意的，因此该界对所有 $l \in [d]$均成立。引理 4.1的证明到此完成。 □

6 引理 4.2的证明

为了证明引理 4.2，我们需要证明 Adam 的更新方向位于负梯度方向的对偶锥中，也就是说，它是一个下降方向。回顾 $z_k := \frac{x_k - \beta_1^n x_1}{1-\beta_1^n}^{k-n}$。我们旨在建立一个下界

关于 $\mathbb{E}\langle \nabla f(z_k),\ z_k - z_{k+1}\rangle$，并将其分解如下：

$$\mathbb{E}\underbrace{\langle \nabla f(z_k) - \nabla f(x_k), z_k - z_{k+1}\rangle}_{:=(a)} + \mathbb{E}\underbrace{\langle \nabla f(x_k), z_k - z_{k+1}\rangle}_{:=(b)}.$$

现在我们对这两项分别进行界定。 $\mathbb{E}(a)$ 可直接利用引理 B.2进行界定。

$$\begin{aligned}
\mathbb{E}(a) &\overset{\text{Cauchy-Schwarz}}{\geq} -\frac{1}{1-\beta_1^n}\|\nabla f(z_k) - \nabla f(x_k)\|_2 \|x_k - x_{k+1} - \beta_1^n(x_{k-n} - x_{k-n+1})\|_2 \\
&\overset{\text{Assumption 2.1}}{\geq} -\frac{L}{1-\beta_1^n}\left\|\frac{x_k - \beta_1^n x_{k-n}}{1-\beta_1^n} - x_k\right\|_2 (\|x_k - x_{k+1}\|_2 + \beta_1^n\|x_{k-n} - x_{k-n+1}\|_2) \\
&\geq -\frac{L\beta_1^n}{1-\beta_1^n}\left(\|x_k - x_{k-1}\|_2 + \cdots + \|x_{k-n+1} - x_{k-n}\|_2\right)(\|x_k - x_{k+1}\|_2 + \beta_1^n\|x_{k-n} - x_{k-n+1}\|_2) \\
&\overset{(B.5)}{\geq} -\frac{L\beta_1^n}{(1-\beta_1^n)^2}\left(\frac{d(1-\beta_1)}{\sqrt{1-\beta_2}}\frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)^2 (\eta_{k-1} + \eta_{k-2} + \cdots + \eta_{k-n})(\eta_k + \beta_1^n\eta_{k-n}) \\
&\overset{k\geq n+1}{\geq} -\frac{L\beta_1^n}{(1-\beta_1^n)^2}\left(\frac{d(1-\beta_1)}{\sqrt{1-\beta_2}}\frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)^2 n\sqrt{n+1}\eta_k\left(\eta_k + \beta_1^n\sqrt{n+1}\eta_k\right) \\
&\geq -\eta_k\frac{1}{\sqrt{k}}\frac{L\eta_0\beta_1^n d^2(1-\beta_1)^2(1+\beta_1^n\sqrt{n+1})n\sqrt{n+1}}{(1-\beta_1^n)^2(1-\beta_2)(1-\frac{\beta_1}{\sqrt{\beta_2}})^2} := -\eta_k\frac{C_1}{\sqrt{k}}.
\end{aligned} \tag{6.1}$$

为了处理项(b)，我们首先针对每个维度l将分析分为两种情况：

- **情况 1：** 有界偏梯度事件 偏梯度事件 $B_{l,k}$，定义为 $\max_i |\partial_l f_i(x_k)| \leq Q_k$；常数$Q_k$ 的定义见下文的（B.1）。

- **情况 2：** 情况 1 的补集，即无界偏梯度事件$B_{cl,k}$。

请注意，由于有界梯度条件，前一种情况较为简单。后一种情况则更复杂，我们将使用引理4.1来处理。我们首先陈述两个观察结果，它们有助于后续分析。

- **观察 1：** 当应用于$\delta = \frac{1}{4n}$时，引理4.1 保证，对于整数 $\bar{k} \in [k-n, k]$，，在满足条件 $\max_i |\partial_l f_i(x_{\bar{k}})| \geq R_{\bar{k}}$ 的前提下，以下两个关键不等式（（6.2）和（6.3））以高概率成立：

$$\left(1 - (1-\beta_2)\frac{8n}{\beta_2^n}\right)\frac{1}{\sqrt{\mathbb{E}_{\bar{k}}(v_{l,\bar{k}})}} \leq \frac{1}{\sqrt{v_{l,\bar{k}}}} \leq \left(\frac{1}{\sqrt{1-(1-\beta_2)\frac{16n}{\beta_2^n}}}\right)\frac{1}{\sqrt{\mathbb{E}_{\bar{k}}(v_{l,\bar{k}})}}. \tag{6.2}$$

$$\frac{\max_i(\partial_l f_i(x_{\bar{k}}))^2}{\mathbb{E}_{\bar{k}}(v_{l,\bar{k}})} \leq \frac{4n}{(1-2n\delta)\beta_2^n} = \frac{8n}{\beta_2^n}, \tag{6.3}$$

- **观察 2：** 当$B_{cl,k}$ 发生时，式（6.2）和（6.3）对所有 $\bar{k} \in [k-n, k]$ 均成立，且其概率至少为$1 - n(n+1)\exp\left(-\frac{1}{160}(1-\beta_2)n\right)$。

观察 1 可以直接从引理 4.1 的陈述及其证明（5.20）看出。观察 2 可以如下证明 ws:

首先，根据引理 B.2 和Q_k 的定义，很容易看出：在$B_{l,k}^c$事件发生时，引理 4.1 的条件对所有整数 $\bar{k} \in [k-n, k]$都自动满足。

其次，在引理 4.1 的条件下，“对给定的 $\bar{k}$，不等式（6.2）和（6.3）成立”这一事件等价于事件 $\mathcal{E}_{\bar{k}}(\alpha(\bar{k})) := \{Y_{k^*,\bar{k}}(\alpha(\bar{k})) \geq \frac{1}{4n}\}$。从引理 4.1 的证明中可以很容易看出这一点。因此，要证明观察 2，需要计算事件 $\bigcap_{\bar{k}=k-n}^{k}\mathcal{E}_{\bar{k}}(\alpha(\bar{k}))$ 的概率。

第三，我们不直接处理 $\bigcap_{\bar{k}=k-n} \mathcal{E}_{\bar{k}}(\alpha(\bar{k}))$，而是考虑一个限制稍多的事件：

$$A_k := \bigcap_{i=0}^{n-1} \bigcap_{\bar{k}=k-n}^{k} \left\{ Y_{k^*,\bar{k}}(i) \geq \tfrac{1}{4n} \right\}. \tag{6.4}$$

定义 A_k 避免事件依赖于 $\alpha(\bar{k})$，这将有助于后续分析。尾部事件 A_k^c 的概率可以通过应用并集界和（5.18）轻松计算，进而得到 $\mathbb{P}(A_k^c) \leq n(n+1)\exp\left(-\frac{1}{160}(1-\beta_2)n\right)$。

最后，事件 A_k 蕴含事件 $\bigcap_{\bar{k}=k-n} \mathcal{E}_{\bar{k}}(\alpha(\bar{k}))$，在 $B_{l,k}^c$ 条件下，这又等价于 (6.2) 和 (6.3)。因此，在 $B_{l,k}^c$ 条件下，对于所有 $\bar{k} \in [k-n, k]$，(6.2) 和 (6.3) 以至少 $1 - n(n+1)\exp\left(-\frac{1}{160}(1-\beta_2)n\right)$ 的概率成立。至此，观察 2 的证明完成。

后续分析围绕观察 1 和观察 2 展开。首先，我们将项 (b) 中的求和分解为三个不同部分，以便分析其期望：

$$\begin{aligned}
(b) &= \frac{1}{1-\beta_1^n} \sum_{l=1}^{d} \left[\eta_k \partial_l f(x_k) \frac{m_{l,k}}{\sqrt{v_{l,k}}} - \beta_1^n \eta_{k-n} \partial_l f(x_k) \frac{m_{l,k-n}}{\sqrt{v_{l,k-n}}} \right] \\
&= \underbrace{\frac{1}{1-\beta_1^n} \sum_{l=1}^{d} \left[\left(\eta_k \partial_l f(x_k) \frac{m_{l,k}}{\sqrt{v_{l,k}}} - \beta_1^n \eta_{k-n} \partial_l f(x_k) \frac{m_{l,k-n}}{\sqrt{v_{l,k-n}}} \right) \mathbb{I}(B_{l,k}) \right]}_{(c)} \\
&\quad + \underbrace{\frac{1}{1-\beta_1^n} \sum_{l=1}^{d} \left[\left(\eta_k \partial_l f(x_k) \frac{m_{l,k}}{\sqrt{v_{l,k}}} - \beta_1^n \eta_{k-n} \partial_l f(x_k) \frac{m_{l,k-n}}{\sqrt{v_{l,k-n}}} \right) \mathbb{I}(A_k \cap B_{l,k}^c) \right]}_{(d)} \\
&\quad + \underbrace{\frac{1}{1-\beta_1^n} \sum_{l=1}^{d} \left[\left(\eta_k \partial_l f(x_k) \frac{m_{l,k}}{\sqrt{v_{l,k}}} - \beta_1^n \eta_{k-n} \partial_l f(x_k) \frac{m_{l,k-n}}{\sqrt{v_{l,k-n}}} \right) \mathbb{I}(A_k^c \cap B_{l,k}^c) \right]}_{(e)},
\end{aligned} \tag{6.5}$$

其中 $B_{l,k}$，A_k 及其补集 $B_{l,k}^c$，A_k^c 在前文和（6.4 ）.

证明的其余部分概述。我们将为每一项给出一个下界。我们将证明，项 (d) 构成下降方向，而 (c) 和 (e) 要么小于 (d)，要么随迭代 k 增大而趋于零。在这三项中，项 (c) 和 (e) 更容易处理：项 (c) 针对有界梯度情形，项 (e) 则考虑尾部事件 A_k^c。项 (d) 的处理会更复杂，因此将进一步细分为若干子项，并借助引理 4.1。

下界 $\mathbb{E}(c)$。我们从 $\mathbb{E}(c)$开始。根据引理 B.2的证明，我们有

$$\begin{aligned}
\mathbb{E}(c) &\geq -\frac{1}{1-\beta_1^n} \mathbb{E}\left[\sum_{l=1}^{d} \left(\eta_k |\partial_l f(x_k)| \frac{|m_{l,k}|}{\sqrt{v_{l,k}}} + \beta_1^n \eta_{k-n} |\partial_l f(x_k)| \frac{|m_{l,k-n}|}{\sqrt{v_{l,k-n}}} \right) \mathbb{I}(B_{l,k}) \right] \\
&\overset{(B.5)}{\geq} -\frac{1}{1-\beta_1^n} \mathbb{E}\left[\sum_{l=1}^{d} \left(\eta_k Q_k \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} + \beta_1^n \eta_{k-n} Q_k \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} \right) \mathbb{I}(B_{l,k}) \right] \\
&\geq -\frac{1}{1-\beta_1^n} d \eta_k Q_k \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} \left(1 + \beta_1^n \frac{\sqrt{k}}{\sqrt{k-n}} \right) \\
&\overset{k \geq n+1}{\geq} -\frac{\eta_k}{1-\beta_1^n} d Q_k \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} \left(1 + \sqrt{n+1} \beta_1^n \right) := -\eta_k \frac{C_2}{\sqrt{k}}.
\end{aligned} \tag{6.6}$$

下界 $\mathbb{E}(e)$。 现在我们推导项 $\mathbb{E}(e)$的一个下界。应用 (B.5) 并使用$\mathbb{I}(A_k^c \cap B_{l,k}^c) \leq \mathbb{I}(A_k^c)$，我们得到

$$\begin{aligned}
\mathbb{E}(e) &= \frac{1}{1-\beta_1^n}\mathbb{E}\left[\sum_{l=1}^{d}\left(\eta_k \partial_l f(x_k)\frac{m_{l,k}}{\sqrt{v_{l,k}}} - \beta_1^n \eta_{k-n}\partial_l f(x_k)\frac{m_{l,k-n}}{\sqrt{v_{l,k-n}}}\right)\mathbb{I}(A_k^c \cap B_{l,k}^c)\right] \\
&\overset{(B.5)}{\geq} -\frac{1}{1-\beta_1^n}\mathbb{E}\left[\sum_{l=1}^{d}\left(\eta_k |\partial_l f(x_k)|\frac{1-\beta_1}{\sqrt{1-\beta_2}}\frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} + \beta_1^n \eta_{k-n}|\partial_l f(x_k)|\frac{1-\beta_1}{\sqrt{1-\beta_2}}\frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)\mathbb{I}(A_k^c \cap B_{l,k}^c)\right] \\
&\overset{k\geq n+1}{\geq} -\eta_k \frac{1}{1-\beta_1^n}\left(1+\beta_1^n\sqrt{n+1}\right)\frac{1}{\sqrt{1-\beta_2}}\frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\mathbb{E}\left[\sum_{l=1}^{d}|\partial_l f(x_k)|\mathbb{I}(A_k^c \cap B_{l,k}^c)\right] \\
&\geq -\eta_k \frac{1}{1-\beta_1^n}\left(1+\beta_1^n\sqrt{n+1}\right)\frac{1}{\sqrt{1-\beta_2}}\frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\mathbb{E}\left[\sum_{l=1}^{d}|\partial_l f(x_k)|\mathbb{I}(A_k^c)\right]. \qquad (6.7)
\end{aligned}$$

这里，倒数第二个不等式成立是因为 $k \geq n+1$，这确保了 $\frac{\eta_{k-n}}{\eta_k} = \sqrt{\frac{k}{k-n}} \leq \sqrt{n+1}$。

为了给 $\mathbb{E}\sum_{l=1}^{d}|\partial_l f(x_k)|\mathbb{I}(A_k^c)$ 求上界，我们注意到，事件 A_k^c 依赖于历史窗口 $\{\tau_j\}_{j=k-n^*}^{k-n-1}$ 内的采样索引。这些索引通过算法的更新递归与 x_k 相关联，从而导致 $|\partial_l f(x_k)|$ 与 $\mathbb{I}(A_k^c)$ 之间存在依赖关系。为了得到所需的上界，我们执行以下解耦步骤。首先，引入一个滞后参数

$$h := 2n + k^*, \qquad \text{where} \quad k^* = \left\lceil \frac{\log(1/4)}{\log \beta_2} \right\rceil. \qquad (6.8)$$

现在，我们从 x_k 转到更早的迭代点 x_{k-h}。根据引理 4.2 的条件，$k-h$ 大于 0，因此它是一个有效的索引。此外，事件 A_k^c 依赖于这些索引

$$\{\tau_j\}_{j=k-2n-k^*}^{k-n-1} = \{\tau_j\}_{j=k-h}^{k-n-1},$$

这些变量是在x_{k-h}之后独立采样的，因此与 $\mathcal{F}_{k-h}$相互独立。因此，x_{k-h} 和 $\mathbb{I}(A_k^c)$是独立随机变量，从而得到

$$\mathbb{E}\left[|\partial_l f(x_{k-h})|\mathbb{I}(A_k^c)\right] = \mathbb{E}\left[|\partial_l f(x_{k-h})|\right]\mathbb{P}(A_k^c). \qquad (6.9)$$

利用引理B.2 关联x_k 和x_{k-h}处的梯度，我们得到

$$\begin{aligned}
\mathbb{E}\left[\sum_{l=1}^{d}|\partial_l f(x_k)|\mathbb{I}(A_k^c)\right] &\overset{\text{Lemma B.2}}{\leq} \mathbb{E}\left[\sum_{l=1}^{d}\left(|\partial_l f(x_{k-h})| + h\Delta_{k-h}\right)\mathbb{I}(A_k^c)\right] \\
&\overset{(6.9)}{=} \mathbb{E}\left[\sum_{l=1}^{d}\left(|\partial_l f(x_{k-h})| + h\Delta_{k-h}\right)\right]\cdot\mathbb{P}(A_k^c) \\
&\overset{\text{Lemma B.2}}{\leq} \mathbb{E}\left[\sum_{l=1}^{d}|\partial_l f(x_k)| + 2dh\,\Delta_{k-h}\right]\cdot\mathbb{P}(A_k^c), \qquad (6.10)
\end{aligned}$$

其中，在最后一个不等式中，我们再次应用引理B.2，将$\mathbb{E}\sum_{l=1}^{d}|\partial_l f(x_{k-h})|$ 转换回当前迭代点x_k，并产生大小为$h\Delta_{k-h}$的加性误差。

结合（6.7）和（6.10），并使用 $\sum_{l=1}^{d}|\partial_l f(x_k)| \leq d\,|\partial_\alpha f(x_k)|$，其中 $\partial_\alpha f(x) := \max_l |\partial_l f(x)|$，再结合 $\Delta_{k-h} \leq \sqrt{h+1}\Delta_1/\sqrt{k}$，我们得到

$$\begin{aligned}
\mathbb{E}(e) &\geq -\eta_k \frac{1}{1-\beta_1^n}\left(1+\beta_1^n\sqrt{n+1}\right)\frac{\mathbb{P}(A_k^c)}{\sqrt{1-\beta_2}}\frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\mathbb{E}\left[d|\partial_\alpha f(x_k)| + 2dh\,\Delta_{k-h}\right] \\
&\geq -\eta_k\,\delta_1(\beta_2)\,\mathbb{E}\left[|\partial_\alpha f(x_k)|\right] - \eta_k \frac{C_3}{\sqrt{k}}, \qquad (6.11)
\end{aligned}$$

其中我们定义

$$\delta_1(\beta_2) := \frac{d\left(1+\beta_1^n\sqrt{n+1}\right)}{1-\beta_1^n}\frac{\mathbb{P}(A_k^c)}{\sqrt{1-\beta_2}}\frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}, \quad C_3 := \frac{2\,dh\sqrt{h+1}\left(1+\beta_1^n\sqrt{n+1}\right)}{1-\beta_1^n}\frac{\mathbb{P}(A_k^c)}{\sqrt{1-\beta_2}}\frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\Delta_1. \tag{6.12}$$

由并集界以及（5.18）可知

$$\mathbb{P}(A_k^c) \le n(n+1)\exp\left(-\frac{1}{160(1-\beta_2)n}\right).$$

随着 $\beta_2 \to 1^-$，概率 $\mathbb{P}(A_k^c)$ 以速率 $O\left(\exp\left(-\frac{1}{160(1-\beta_2)n}\right)\right)$ 指数级地衰减，这主导了 $\frac{1}{\sqrt{1-\beta_2}}$ 的多项式增长。因此，因子 $\mathbb{P}(A_k^c)\sqrt{1-\beta_2}$ 随着 $\beta_2 \to 1^-$ 而趋于零，从而确保 $\delta_1(\beta_2) \to 0$ 随着 $\beta_2 \to 1^-$。至此完成了 $\mathbb{E}(e)$ 的下界。

$\mathbb{E}(d)$ 的下界。由于动量和无界梯度的存在，(d) 的下界更为复杂。我们首先将所有 $v_{l,k-n}$ 转换为 $v_{l,k}$。我们讨论两种情形。

情形（1）：当 $\partial_l f(x_k) m_{l,k-n} \le 0$ 时。回顾 $\frac{1}{\sqrt{v_{l,k-n}}} \ge \frac{1}{\sqrt{v_{l,k}}}\sqrt{\beta_2^n}$，我们有

$$-\partial_l f(x_k)\frac{m_{l,k-n}}{\sqrt{v_{l,k-n}}} \ge -\partial_l f(x_k)\frac{m_{l,k-n}}{\sqrt{v_{l,k}}}\sqrt{\beta_2^n}.$$

情形（2）：当 $\partial_l f(x_k) m_{l,k-n} \ge 0$。在这种情况下，我们首先给出 $v_{l,\bar{k}-1}$ 的一个下界，其中整数 $\bar{k} \in [k-n+1, k]$：

$$\begin{aligned} v_{l,\bar{k}-1} &\ge v_{l,\bar{k}}\left(1-\frac{|v_{l,\bar{k}-1}-v_{l,\bar{k}}|}{v_{l,\bar{k}}}\right) \\ &\ge v_{l,\bar{k}}\left(1-(1-\beta_2)\frac{\max_i|\partial_l f_i(x_{\bar{k}})|^2+v_{l,\bar{k}-1}}{v_{l,\bar{k}}}\right) \\ &\overset{(6.2)(6.3)}{\ge} v_{l,\bar{k}}\left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right). \end{aligned} \tag{6.13}$$

于是我们有

$$-\partial_l f(x_k)\frac{m_{l,k-n}}{\sqrt{v_{l,k-n}}} \overset{(6.13)}{\ge} -\partial_l f(x_k)\frac{m_{l,k-n}}{\sqrt{v_{l,k}}}\left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right)^{-\frac{n}{2}}.$$

结合这两种情况，我们得到

$$-\partial_l f(x_k)\frac{m_{l,k-n}}{\sqrt{v_{l,k-n}}} \ge -\partial_l f(x_k)\frac{m_{l,k-n}}{\sqrt{v_{l,k}}} - \delta_2(\beta_2)|\partial_l f(x_k)|\frac{|m_{l,k-n}|}{\sqrt{v_{l,k}}}, \tag{6.14}$$

其中

$$\delta_2(\beta_2) := \left|\sqrt{\beta_2^n}-1\right| + \left|\left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right)^{-\frac{n}{2}}-1\right|. \tag{6.15}$$

我们简要讨论一下 $\delta_2(\beta_2)$ 相对于n的阶。假设我们使用 $1-\beta_2 = \mathcal{O}(n^{-a})$，且$a > 0$，那么通过基本计算可以证明，第一项为 $\mathcal{O}(n^{-(a-1)})$，第二项为 $\mathcal{O}(n^{-(a-2)})$。

利用上面的不等式（6.14），我们已将项 (d) 中的所有 $v_{l,k-n}$ 都转换成了 $v_{l,k}$。现在我们进一步

将 (d) 分解为三个项 (d_1)、(d_2) 和 (d_3)：

$$
\begin{aligned}
(d) \quad &\geq \quad \frac{1}{1-\beta_1^n}\sum_{l=1}^{d}\left[\left(\eta_k \partial_l f(x_k)\frac{m_{l,k}}{\sqrt{v_{l,k}}} - \beta_1^n \eta_{k-n}\partial_l f(x_k)\frac{m_{l,k-n}}{\sqrt{v_{l,k}}} - \beta_1^n \eta_{k-n}\delta_2(\beta_2)|\partial_l f(x_k)|\frac{|m_{l,k-n}|}{\sqrt{v_{l,k}}}\right)\mathbb{I}(A_k \cap B_{l,k}^c)\right] \\
&= \quad \frac{1}{1-\beta_1^n}\sum_{l=1}^{d}\left[\left(\eta_k \frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}}(m_{l,k} - \frac{\sqrt{k}}{\sqrt{k-n}}\beta_1^n m_{l,k-n}) - \beta_1^n \eta_{k-n}\delta_2(\beta_2)|\partial_l f(x_k)|\frac{|m_{l,k-n}|}{\sqrt{v_{l,k}}}\right)\mathbb{I}(A_k \cap B_{l,k}^c)\right] \\
&= \quad \frac{1}{1-\beta_1^n}\sum_{l=1}^{d}\left[\left(\eta_k \frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}}\left((m_{l,k} - \beta_1^n m_{l,k-n}) + (1 - \frac{\sqrt{k}}{\sqrt{k-n}})\beta_1^n m_{l,k-n}\right)\right) - \beta_1^n \eta_{k-n}\delta_2(\beta_2)|\partial_l f(x_k)|\frac{|m_{l,k-n}|}{\sqrt{v_{l,k}}}\right)\mathbb{I}(A_k \cap B_{l,k}^c)\right] \\
&\overset{\text{(i)}}{\geq} \quad \eta_k \frac{(1-\beta_1)}{1-\beta_1^n}\sum_{l=1}^{d}\left(\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}}\partial_l f_{\tau_k}(x_k) + \underbrace{\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}}}_{:=F_k}\underbrace{\left(\beta_1 \partial_l f_{\tau_{k-1}}(x_{k-1}) + \cdots + \beta_1^{n-1}\partial_l f_{\tau_{k-n+1}}(x_{k-n+1})\right)}_{:=G}\right)\mathbb{I}(A_k \cap B_{l,k}^c) \\
&\qquad \underbrace{-\left(\eta_k \frac{n+1}{\sqrt{k}}\frac{\beta_1^n}{1-\beta_1^n} + \eta_k \frac{\sqrt{n+1}\beta_1^n}{1-\beta_1^n}\delta_2(\beta_2)\right)\sum_{l=1}^{d}\frac{|\partial_l f(x_k)||m_{l,k-n}|}{\sqrt{v_{l,k}}}\mathbb{I}(A_k \cap B_{l,k}^c)}_{:=(d_3)} \\
&= \quad \underbrace{\eta_k \frac{(1-\beta_1)}{1-\beta_1^n}\sum_{l=1}^{d}\frac{\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)}{\sqrt{v_{l,k}}}\cdot\mathbb{I}(A_k \cap B_{l,k}^c)}_{:=(d_1)} + \underbrace{\eta_k \frac{(1-\beta_1)}{1-\beta_1^n}\sum_{l=1}^{d}F_k \cdot G \cdot \mathbb{I}(A_k \cap B_{l,k}^c)}_{:=(d_2)} + (d_3). \qquad (6.16)
\end{aligned}
$$

其中 (i) 是由 $|1 - \sqrt{k}/\sqrt{k-n}| \leq \frac{n+1}{n\sqrt{k}}$ 得出的。

接下来，我们将说明 (d_1) 构成下降方向，而 (d_2) 和 (d_3) 是误差项。

下界 $\mathbb{E}(d_1)$。我们首先对 $\mathbb{E}(d_1)$进行界定。我们首先推导条件期望$\mathbb{E}_k(d_1)$ 的下界，随后取外层期望。根据引理 4.1，我们通过 $\frac{1}{\sqrt{v_{l,k}}}$ 经由 $\frac{1}{\sqrt{\mathbb{E}_k(v_{l,k})}}$控制该量，从而得到：

$$
\begin{aligned}
\mathbb{E}_k(d_1) \quad &= \quad \eta_k \frac{(1-\beta_1)}{1-\beta_1^n}\mathbb{E}_k\left(\sum_{l=1}^{d}\frac{\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)}{\sqrt{v_{l,k}}}\mathbb{I}(A_k \cap B_{l,k}^c)\right) \\
&\overset{\text{(i)}}{\geq} \quad \eta_k\left[\frac{(1-\beta_1)}{1-\beta_1^n}\mathbb{E}_k\left(\sum_{l=1}^{d}\frac{\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)}{\sqrt{\mathbb{E}_k(v_{l,k})}}\mathbb{I}(A_k \cap B_{l,k}^c)\right) - \delta_3(\beta_2)\,|\partial_\alpha f(x_k)|\right] \\
&= \quad \eta_k\left[\frac{(1-\beta_1)}{1-\beta_1^n}\mathbb{E}_k\left(\sum_{l=1}^{d}\frac{\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)}{\sqrt{\mathbb{E}_k(v_{l,k})}}\mathbb{I}(A_k \cap B_{l,k}^c)\right) - \delta_3(\beta_2)\,|\partial_\alpha f(x_k)|\right. \\
&\qquad + \frac{(1-\beta_1)}{1-\beta_1^n}\mathbb{E}_k\left(\sum_{l=1}^{d}\frac{\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)}{\sqrt{\mathbb{E}_k(v_{l,k})}}\mathbb{I}(A_k \cap B_{l,k})\right) \\
&\qquad \left. - \frac{(1-\beta_1)}{1-\beta_1^n}\mathbb{E}_k\left(\sum_{l=1}^{d}\frac{\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)}{\sqrt{\mathbb{E}_k(v_{l,k})}}\mathbb{I}(A_k \cap B_{l,k})\right)\right] \\
&\overset{\text{(ii)}}{\geq} \quad \eta_k\left[\frac{(1-\beta_1)}{1-\beta_1^n}\mathbb{E}_k\left(\sum_{l=1}^{d}\frac{\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)}{\sqrt{\mathbb{E}_k(v_{l,k})}}\mathbb{I}(A_k)\right) - \delta_3(\beta_2)\,|\partial_\alpha f(x_k)| - \frac{(1-\beta_1)}{1-\beta_1^n}\frac{2\sqrt{2n}d}{\sqrt{\beta_2^n}}Q_k\right] \\
&\overset{\text{(iii)}}{\geq} \quad \eta_k\left[\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}} - \sum_{l=1}^{d}\frac{(\partial_l f(x_k))^2}{n\sqrt{\mathbb{E}_k\left(v_{l,k}\right)}}\mathbb{I}(A_k^c) - \delta_3(\beta_2)\,|\partial_\alpha f(x_k)| - \frac{(1-\beta_1)}{1-\beta_1^n}\frac{2\sqrt{2n}d}{\sqrt{\beta_2^n}}Q_k\right], \qquad (6.17)
\end{aligned}
$$

其中 $\alpha := \arg\max_l |\partial_l f(x_k)|$ 且 $v_{\alpha,k} := \max_l v_{l,k}$。步骤 (i)、(ii)、(iii) 的依据如下：

(6.17) 中 (i) 的依据(6.17)：我们应用引理 4.1 中的集中不等式（6.2），根据 $\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)$ 的符号将该项拆分为两部分：

- 当 $\partial_l f(xk)\partial_l f\ _{\tau_k}(xk) \geq 0$时，我们使用（6.2）中的下界：

$$\frac{1}{\sqrt{v_{l,k}}} \geq \left(1-(1-\beta_2)\frac{8n}{\beta_2^n}\right)\frac{1}{\sqrt{\mathbb{E}_k(v_{l,k})}}.$$

- 当 $\partial_l f(xk)\partial_l f\ _{\tau_k}(xk) < 0$时，我们使用（6.2）中的上界：

$$\frac{1}{\sqrt{v_{l,k}}} \leq \left(\frac{1}{\sqrt{1-(1-\beta_2)\frac{16n}{\beta_2^n}}}\right)\frac{1}{\sqrt{\mathbb{E}_k(v_{l,k})}}.$$

合并　　这两种情况并汇总误差项后，我们得到系数 $\delta_3(\beta_2)$，其定义如下：

$$\delta_3(\beta_2) = \frac{(1-\beta_1)}{1-\beta_1^n}\left(\frac{(1-\beta_2)16\sqrt{2n}nd}{\sqrt{\beta_2^{3n}}} + \left(\frac{1}{\sqrt{1-(1-\beta_2)\frac{16n}{\beta_2^n}}}-1\right)\frac{2\sqrt{2n}d}{\sqrt{\beta_2^n}}\right).$$

证明采用与（6.14）中类似的策略，即根据乘积的符号对该项进行分解。

对（6.17）中（ii）的证明：在事件 $B_{l下},\ k$，有 $\partial_l f(x_k) \leq \max_i |\partial_l f_i(x_k)| \leq Q_k$。此外，在事件 A_k下，引理 4.1 （特别是不等式（6.3））保证：

$$\frac{\partial_l f_{\tau_k}(x_k)}{\sqrt{\mathbb{E}_k(v_{l,k})}} \leq \frac{2\sqrt{2n}}{\sqrt{\beta_2^n}}.$$

因此，在$A_k \cap B_{l,k}$条件下，我们有：

$$\left|\frac{\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)}{\sqrt{\mathbb{E}_k(v_{l,k})}}\right| \leq Q_k \cdot \frac{2\sqrt{2n}}{\sqrt{\beta_2^n}} = \frac{2\sqrt{2n}Q_k}{\sqrt{\beta_2^n}}.$$

对(6.17)中（iii）的论证：这一步利用了两个关键事实：

(i) 事实 1：根据不等式 $\frac{1-\beta_1}{1-\beta_1^n} \geq \frac{1}{n}$，我们可以为该系数给出下界。

(ii) 事实 2：事件 A_k 仅取决于截至步骤$k-n$的历史，因此在给定 $\mathcal{F}_k$ 的条件下，它与当前梯度样本 $\partial_l f_{\tau_k}(x_k)$ 独立。因此：

$$\begin{aligned}
\mathbb{E}_k\left[\sum_{l=1}^d \frac{\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)}{\sqrt{\mathbb{E}_k(v_{l,k})}}\mathbb{I}(A_k)\right] &= \sum_{l=1}^d \frac{\mathbb{E}_k\left(\partial_l f(x_k)\partial_l f_{\tau_k}(x_k)\right)}{\sqrt{\mathbb{E}_k(v_{l,k})}}\mathbb{I}(A_k) \\
&= \sum_{l=1}^d \frac{(\partial_l f(x_k))^2}{\sqrt{\mathbb{E}_k(v_{l,k})}}\left(1-\mathbb{I}(A_k^c)\right) \\
&\geq \frac{(\partial_\alpha f(x_k))^2}{\sqrt{\mathbb{E}_k(v_{\alpha,k})}} - \sum_{l=1}^d \frac{(\partial_l f(x_k))^2}{\sqrt{\mathbb{E}_k(v_{l,k})}}\mathbb{I}(A_k^c). \qquad (6.18)
\end{aligned}$$

至此，我们完成了对 $(\mathbb{E}_k(d_1)$ 在 (6.17) 中下界的推导。
现在我们来处理（6.17）。对（6.17）取期望后，我们现在对 $\mathbb{E}\left[\frac{(\partial_l f(x_k))^2 n}{\sqrt{\mathbb{E}_k(v_{l,k})}}\mathbb{I}(A_k^c)\right]$取上界。我们将使用上界（5.9）以及（6.11）中的相同去耦策略：

$$\begin{aligned}\mathbb{E}\left[\sum_{l=1}^{d}\frac{(\partial_l f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{l,k})}}\mathbb{I}(A_k^c)\right] &\leq \mathbb{E}\left[\sum_{l=1}^{d}\frac{(\max_i|\partial_l f_i(x_k)|)}{n\sqrt{\mathbb{E}_k(v_{l,k})}}|\partial_l f(x_k)|\mathbb{I}(A_k^c)\right]\\ &\overset{(5.9)}{\leq} \mathbb{E}\left[\sum_{l=1}^{d}\frac{1}{n}\cdot\frac{\sqrt{n}}{\sqrt{(1-\beta_2)}}|\partial_l f(x_k)|\mathbb{I}(A_k^c)\right]\\ &\leq \frac{\mathbb{P}(A_k^c)dh}{\sqrt{n(1-\beta_2)}}\mathbb{E}|\partial_\alpha f(x_k)|+\frac{\mathbb{P}(A_k^c)2dh}{\sqrt{n(1-\beta_2)}}\sqrt{h+1}\frac{\Delta_1}{\sqrt{k}}.\end{aligned}$$

结合上述结果，我们得到

$$\mathbb{E}(d_1)=\mathbb{E}\mathbb{E}_k(d_1) \geq \eta_k\left[\mathbb{E}\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}}-\delta_4(\beta_2)\mathbb{E}\left|\partial_\alpha f(x_k)\right|-\frac{C_4}{\sqrt{k}}\right], \tag{6.19}$$

其中

$$\delta_4(\beta_2) = \frac{\mathbb{P}(A_k^c)dh}{\sqrt{n(1-\beta_2)}}+\delta_3(\beta_2), \tag{6.20}$$

$$C_4 = \frac{\mathbb{P}(A_k^c)2dh}{\sqrt{n(1-\beta_2)}}\sqrt{h+1}\Delta_1+\frac{(1-\beta_1)}{1-\beta_1^n}\frac{2\sqrt{2n}d}{\sqrt{\beta_2^n}}Q_1. \tag{6.21}$$

这就完成了对 $\mathbb{E}(d_1)$ 在 (6.16) 中的下界。

(d_2) 的下界。为了界定 $\mathbb{E}(d_2)$ 在 (6.16) 中的项，我们将所有变量 x 都转换为 x_{k-n}，然后对截至 $\mathcal{F}_{k-n}$ 的历史条件下取期望。我们将目标项记为

$$F_{k-n}:=\frac{\partial_l f(x_{k-n})}{\sqrt{v_{l,k-n}}} \quad \text{and} \quad G_{k-n}:=\beta_1\partial_l f_{\tau_{k-1}}(x_{k-n})+\cdots+\beta_1^{n-1}\partial_l f_{\tau_{k-n+1}}(x_{k-n}). \tag{6.22}$$

现在，我们利用以下引理，将 F_k 在 (6.16) 中转换为 F_{k-n}。引理 6.1。

引理 6.1。考虑算法 *1*。假设 $k \geq \lceil^{\log(1/4)}{\log\beta_2}\rceil+2n+1$，且对于所有整数 $\bar{k}\in[k-n,k]$，$\max_i|\partial_l f_i(x_{\bar{k}})|\geq R_{\bar{k}}$，并且联合事件 $A_k\cap B_{l,k}^c$ 发生。则对于任意 $f\in\mathcal{F}_{nL,D_0,D_1}(\mathbb{R}^d)$，我们有

$$\left|\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}}-\frac{\partial_l f(x_{k-n})}{\sqrt{v_{l,k-n}}}\right|\leq\frac{n}{\sqrt{\beta_2^n}}\frac{\Delta_{k-n}}{\sqrt{v_{l,k-n}}}+n\delta_5(\beta_2), \tag{6.23}$$

其中，Δ_{k-n} 的定义与引理*B.2*中的相同，而 $\delta_5(\beta_2)$ 的定义如下。

$$\delta_5(\beta_2)=\delta_{5,1}(\beta_2)\delta_{5,2}(\beta_2)\delta_{5,3}(\beta_2), \tag{6.24}$$

with $\delta_{5,1}(\beta_2)=\left(\left[1-\left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right)^{\frac{1}{2}}\right]+\frac{1}{\sqrt{\beta_2}}-1\right)$,

$\delta_{5,2}(\beta_2)=\left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right)^{-\frac{1}{2}}$ *and* $\delta_{5,3}(\beta_2)=\frac{2\sqrt{2n}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}$.

引理6.1的证明见附录C。

结合引理6.1 以及引理B.2，我们有以下结果（G，F_{k-n}，G_{k-n} 定义见（6.16）并且(6.22)).

$$|F_k-F_{k-n}|\overset{\text{Lemma 6.1}}{\leq}\frac{n}{\sqrt{\beta_2^n}}\frac{\Delta_{k-n}}{\sqrt{v_{l,k-n}}}+n\delta_5(\beta_2):=C_{F,k},$$

$$|G_{k-n}-G|\overset{\text{Lemma B.2}}{\leq}\sum_{m=1}^{n-1}\beta_1^m\left(\sum_{\bar{k}=m}^{n-1}\Delta_{k-\bar{k}}\right)\leq n^2\Delta_{k-n}\leq n^2\Delta_k\sqrt{n+1}:=C_{G,k}.$$

那么，

$$
\begin{aligned}
(d_2) &= \sum_{l=1}^{d} \eta_k \frac{(1-\beta_1)}{1-\beta_1^n} (F_k \cdot G) \cdot \mathbb{I}(A_k \cap B_{l,k}^c) \\
&= \sum_{l=1}^{d} \eta_k \frac{(1-\beta_1)}{1-\beta_1^n} F_k \cdot [G_{k-n} - (G_{k-n} - G)] \cdot \mathbb{I}(A_k \cap B_{l,k}^c) \\
&\geq \sum_{l=1}^{d} \eta_k \frac{(1-\beta_1)}{1-\beta_1^n} [F_k \cdot G_{k-n} - |F_k| \cdot |G_{k-n} - G|] \cdot \mathbb{I}(A_k \cap B_{l,k}^c) \\
&\geq \sum_{l=1}^{d} \eta_k \frac{(1-\beta_1)}{1-\beta_1^n} [F_{k-n} \cdot G_{k-n} - |F_{k-n} - F_k| \cdot |G_{k-n}| - |F_k| \cdot |G_{k-n} - G|] \cdot \mathbb{I}(A_k \cap B_{l,k}^c) \\
&\geq \underbrace{\sum_{l=1}^{d} \eta_k \frac{(1-\beta_1)}{1-\beta_1^n} \left[\frac{\partial_l f(x_{k-n})}{\sqrt{v_{l,k-n}}} \left(\beta_1 \partial_l f_{\tau_{k-1}}(x_{k-n}) + \cdots + \beta_1^{n-1} \partial_l f_{\tau_{k-n+1}}(x_{k-n}) \right) \right] \cdot \mathbb{I}(A_k \cap B_{l,k}^c)}_{:=(d_{2,1})} \\
&\quad - \sum_{l=1}^{d} \eta_k \frac{(1-\beta_1)}{1-\beta_1^n} \left[|C_{F,k} \cdot G_{k-n}| + |F_k \cdot C_{G,k}| \right] \cdot \mathbb{I}(A_k \cap B_{l,k}^c). \qquad (6.25)
\end{aligned}
$$

我们可以通过采用与推导 $\mathbb{E}(d_{2,1})$ 的下界 $\mathbb{E}(d_1)$ 时相同的方法，对照（6.17）和（6.19），来界定该式。首先注意以下事实（与事实（6.18）类似）：由于事件 A_k 仅取决于截至第$(k-n)$次迭代的历史，并且对于每个 $\bar{k} \in [k-n,\ \tau_{\bar{k}}$与 $\mathcal{F}_{k-n}$ 独立，$k]$，我们有

$$
\mathbb{E}_{k-n}\left[\partial_l f(x_{k-n}) \partial_l f_{\tau_{\bar{k}}}(x_{k-n}) \mathbb{I}(A_k)\right] = \partial_l f(x_{k-n}) \mathbb{E}_{k-n}[\partial_l f_{\tau_{\bar{k}}}(x_{k-n})] \mathbb{I}(A_k) = (\partial_l f(x_{k-n}))^2 \mathbb{I}(A_k). \qquad (6.26)
$$

照前面的方法，利用以下事实可得如下界：$\mathbb{I}(A_k) = \mathbb{I}(A_k \cap Bl,k) + \mathbb{I}(A_k \cap B\overset{c}{l,k})$：

$$
\begin{aligned}
\mathbb{E}(d_{2,1}) = \mathbb{E}\mathbb{E}_{k-n}(d_{2,1}) &\geq \sum_{i=1}^{n-1} \beta_1^i \eta_k \left[\mathbb{E} \frac{(\partial_\alpha f(x_{k-n}))^2}{n\sqrt{\mathbb{E}_{k-n}(v_{\alpha,k-n})}} - \delta_4(\beta_2) \mathbb{E} \left|\partial_\alpha f(x_{k-n})\right| - \frac{C_4}{\sqrt{k-n}} \right] \\
&\geq -\eta_k \frac{1}{1-\beta_1} \left[\delta_4(\beta_2) \mathbb{E} \left|\partial_\alpha f(x_{k-n})\right| + \frac{C_4\sqrt{n+1}}{\sqrt{k}} \right] \\
&\geq -\eta_k \left[\delta_6(\beta_2) \mathbb{E} \left|\partial_\alpha f\left(x_k\right)\right| + \delta_7(\beta_2)\sqrt{D_0} + \frac{C_5}{\sqrt{k}} \right], \qquad (6.27)
\end{aligned}
$$

其中

$$
\delta_6(\beta_2) = \frac{\delta_4(\beta_2)}{1-\beta_1} nd\sqrt{D_1}\sqrt{n}d, \quad \delta_7(\beta_2) = \frac{\delta_4(\beta_2)}{1-\beta_1} nd\sqrt{n}\sqrt{d}, \quad C_5 = \frac{\delta_4(\beta_2)}{1-\beta_1} n^2 d\Delta_1\sqrt{n+1} + \frac{C_4\sqrt{n+1}}{1-\beta_1}. \qquad (6.28)
$$

W我们以类似的方式处理（6.25）中的其余项，使用引理 B.4、（6.2）和（6.3）　　.

$$
\begin{aligned}
|C_{F,k} \cdot G_{k-n}| \mathbb{I}(A_k \cap B_{l,k}^c) &= \left(\frac{n}{\sqrt{\beta_2^n}} \frac{\Delta_{k-n}}{\sqrt{v_{l,k-n}}} + n\delta_5(\beta_2) \right) \left(\beta_1 |\partial_l f_{\tau_{k-1}}(x_{k-n})| + \cdots + \beta_1^{n-1} |\partial_l f_{\tau_{k-n+1}}(x_{k-n})| \right) \mathbb{I}(A_k \cap B_{l,k}^c) \\
&\leq \frac{1}{\sqrt{k}} \frac{(1-\beta_1^n)}{1-\beta_1} \frac{\Delta_1 2\sqrt{2n} n\sqrt{n+1}}{\sqrt{\beta_2^n}\sqrt{\beta_2^n - (1-\beta_2)16n}} + \delta_5(\beta_2) \frac{(1-\beta_1^n)}{1-\beta_1} n\sqrt{D_1}\sqrt{n}d |\partial_\alpha f(x_k)| \\
&\quad + \frac{1}{\sqrt{k}} \delta_5(\beta_2) \frac{(1-\beta_1^n)}{1-\beta_1} n^2\sqrt{n+1}\Delta_1 + \delta_5(\beta_2) \frac{(1-\beta_1^n)}{1-\beta_1} n\sqrt{n}\sqrt{d}\sqrt{D_0}. \qquad (6.29)
\end{aligned}
$$

And

$$
|F_k \cdot C_{G,k}| \mathbb{I}(A_k \cap B_{l,k}^c) = \frac{|\partial_l f(x_k)|}{\sqrt{v_{l,k}}} n^2 \Delta_k \sqrt{n+1} \mathbb{I}(A_k \cap B_{l,k}^c) \overset{(6.2)(6.3)}{\leq} \frac{2\sqrt{2n}}{\sqrt{\beta_2^n - (1-\beta_2)16n}} n^2 \Delta_k \sqrt{n+1}. \qquad (6.30)
$$

将（6.27）、（6.29）、（6.30）代入（6.25），并取期望，

$$\begin{aligned}\mathbb{E}(d_2) &\overset{(6.25)}{\geq} \mathbb{E}(d_{2,1}) - \sum_{l=1}^{d} \eta_k \frac{(1-\beta_1)}{1-\beta_1^n}\left[\,\left|C_{F,k}\cdot G_{k-n}\right| + \left|F_k \cdot C_{G,k}\right|\,\right]\mathbb{I}(A_k \cap B_{l,k}^c)\\ &\overset{(6.27)(6.29)(6.30)}{\geq} -\eta_k\left[\delta_8(\beta_2)\mathbb{E}\left[|\partial_\alpha f(x_k)|\right] + \delta_9(\beta_2)\sqrt{D_0} + \frac{C_6}{\sqrt{k}}\right],\end{aligned}$$

其中常数项定义如下（$\delta_5(\beta_2)$， $\delta_6(\beta_2)$ 和 $\delta_7(\beta_2)$ 在（6.24）和（6.28）中定义）：

$$\delta_8(\beta_2) = \delta_6(\beta_2) + d\delta_5(\beta_2)n\sqrt{D_1}\sqrt{n}d, \tag{6.31}$$

$$\delta_9(\beta_2) = \delta_7(\beta_2) + d\delta_5(\beta_2)n\sqrt{n}\sqrt{d}, \tag{6.32}$$

$$C_6 = C_5 + d\frac{\Delta_1 2\sqrt{2n}n\sqrt{n+1}}{\sqrt{\beta_2^n}\sqrt{\beta_2^n-(1-\beta_2)16n}} + d\delta_5(\beta_2)\frac{(1-\beta_1^n)}{1-\beta_1}n^2\sqrt{n+1}\Delta_1 + \frac{(1-\beta_1)}{1-\beta_1^n}d\frac{\Delta_1 2\sqrt{2n}n^2\sqrt{n+1}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}. \tag{6.33}$$

的下界 $\mathbb{E}(d_3)$。 现在我们给出 (6.16) 中 $\mathbb{E}(d_3)$ 的一个下界。由于 $|\partial_l f(x_k)| \leq n\max_i|\partial_l f_i(x_k)|$，我们有

$$-\sum_{l=1}^{d}\frac{|\partial_l f(x_k)||m_{l,k-n}|}{\sqrt{v_{l,k}}}\mathbb{I}(A_k\cap B_{l,k}^c) \overset{(6.2)(6.3)}{\geq} -\frac{2\sqrt{2n}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}\sum_{l=1}^{d}|m_{l,k-n}|. \tag{6.34}$$

我们进一步对 $\sum_{l=1}^{d}|m_{l,k-n}|$，进行界定，

$$\begin{aligned}\sum_{l=1}^{d}|m_{l,k-n}| &\leq (1-\beta_1)\sum_{l=1}^{d}[|\partial_l f_{\tau_{k-n}}(x_{k-n})| + \beta_1|\partial_l f_{\tau_{k-n-1}}(x_{k-n-1})| + \cdots] + \beta_1^{k-n}\sum_{l=1}^{d}|m_{l,0}|\\ &\overset{\text{Lemma B.4}}{\leq} (1-\beta_1)\sum_{q=1}^{k-n}\beta_1^{q-1}\left[\sqrt{D_1}\sqrt{n}d\left(|\partial_\alpha f(x_k)| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{2\sqrt{n+1}qd\Delta_1}{\sqrt{k}}\right] + \beta_1^{k-n}\|m_0\|_1\\ &\overset{\text{Lemma B.1}}{\leq} \sqrt{D_1}\sqrt{n}d\left(|\partial_\alpha f(x_k)| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{1}{1-\beta_1}\frac{2\sqrt{n+1}d\Delta_1}{\sqrt{k}} + \beta_1^{k-n}\|m_0\|_1\\ &\overset{\text{(i)}}{\leq} \sqrt{D_1}\sqrt{n}d\left(|\partial_\alpha f(x_k)| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{1}{1-\beta_1}\frac{2\sqrt{n+1}d\Delta_1}{\sqrt{k}} + \frac{1}{\sqrt{k}}\|m_0\|_1,\end{aligned} \tag{6.35}$$

其中，当 k 足够大且满足 $\beta_1^{k-n} \leq \frac{1}{\sqrt{k}}$ 时，(i) 成立。将 (6.35) 代入 (6.34)，

$$\begin{aligned}&-\sum_{l=1}^{d}\frac{|\partial_l f(x_k)||m_{l,k-n}|}{\sqrt{v_{l,k}}}\mathbb{I}(A_k\cap B_{l,k}^c)\\ \overset{(6.34)}{\geq}\ &-\frac{2\sqrt{2n}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}\sum_{l=1}^{d}|m_{l,k-n}|\\ \overset{(6.35)}{\geq}\ &-\frac{2\sqrt{2n}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}\left(\sqrt{D_1}\sqrt{n}d\left(|\partial_\alpha f(x_k)| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{2\sqrt{n+1}d\Delta_1}{(1-\beta_1)\sqrt{k}} + \frac{\|m_0\|_1}{\sqrt{k}}\right).\end{aligned} \tag{6.36}$$

将不等式（6.36）代入（(d_3) 中的项（6.16），并取期望，

$$\mathbb{E}(d_3) \geq -\eta_k\left[\left(\delta_{10}(\beta_2) + \frac{C_7}{\sqrt{k}}\right)\mathbb{E}|\partial_\alpha f(x_k)| + \delta_{11}(\beta_2)\sqrt{D_0} + \frac{C_8}{\sqrt{k}}\right], \tag{6.37}$$

其中常数项定义如下（$\delta_2(\beta_2)$ 定义见（6.15）。：

$$\begin{aligned}
\delta_{10}(\beta_2) &= \frac{\sqrt{n+1}\beta_1^n}{1-\beta_1^n}\delta_2(\beta_2)\frac{2\sqrt{2n}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}\sqrt{D_1}\sqrt{n}d, && (6.38)\\
\delta_{11}(\beta_2) &= \frac{\sqrt{n+1}\beta_1^n}{1-\beta_1^n}\delta_2(\beta_2)\frac{2\sqrt{2n}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}\sqrt{n}\sqrt{d}, && (6.39)\\
C_7 &= (n+1)\frac{\beta_1^n}{1-\beta_1^n}\frac{2\sqrt{2n}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}\sqrt{D_1}\sqrt{n}d, && (6.40)\\
C_8 &= (n+1)\frac{\beta_1^n}{1-\beta_1^n}\frac{2\sqrt{2n}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}\sqrt{n}\sqrt{D_0 d}\\
&\quad+\frac{\sqrt{n+1}\beta_1^n}{1-\beta_1^n}\delta_2(\beta_2)\frac{2\sqrt{2n}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}\left(\frac{2\sqrt{n+1}d\Delta_1}{(1-\beta_1)}+\|m_0\|_1\right). && (6.41)
\end{aligned}$$

上述分析总结。现在，我们总结上述分析，并通过对（$\partial_\alpha \mathrm{f}(\mathrm{x}_k)$)^2、 $\partial_\alpha \mathrm{f}(\mathrm{x}_k)$ 和与梯度独立的误差中的各项进行分组，给出（6.16）中$\mathbb{E}(\mathrm{b}) \geq \mathbb{E}(\mathrm{c}) + \mathbb{E}(\mathrm{d}_1) + \mathbb{E}(\mathrm{d}_2) + \mathbb{E}(\mathrm{d}_3) + \mathbb{E}(\mathrm{e})$的下界。

$$\begin{aligned}
\mathbb{E}(b) &\geq \eta_k\left(\mathbb{E}\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}}-\mathbb{E}|\partial_\alpha f(x_k)|\cdot\left(\delta_4(\beta_2)+\delta_8(\beta_2)+\delta_{10}(\beta_2)+\delta_1(\beta_2)+\frac{C_7}{\sqrt{k}}\right)\right.\\
&\quad\left.-(\delta_9(\beta_2)+\delta_{11}(\beta_2))\sqrt{D_0}-\frac{C_2+C_3+C_4+C_6+C_8}{\sqrt{k}}\right)\\
&:= \eta_k\left\{\mathbb{E}\left[\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}}-\left(\delta_{12}(\beta_2)+\frac{C_7}{\sqrt{k}}\right)|\partial_\alpha f(x_k)|\right]-\delta_{11}(\beta_2)\sqrt{D_0}-\frac{C_9}{\sqrt{k}}\right\}, && (6.42)
\end{aligned}$$

其中

$$\begin{aligned}
\delta_{12}(\beta_2) &= \delta_4(\beta_2)+\delta_8(\beta_2)+\delta_{10}(\beta_2)+\delta_1(\beta_2),\\
C_9 &= C_2+C_3+C_4+C_6+C_8, && (6.43)
\end{aligned}$$

并且， $\delta_4(\beta_2)$、 $\delta_8(\beta_2)$、 $\delta_{10}(\beta_2)$ 和 $\delta_1(\beta_2)$ 分别定义于（6.20）、（6.31）、（6.38）、（6.12）中。注意，当$\beta_2 \to 1$ 和 $k \to \infty$时， $\delta_{12}(\beta_2) + C\sqrt{k}$ 趋于零。C_2,C_3,C_4,C_6,C_8 分别定义于（6.6）、（6.12）、（6.21）、（6.33）、（6.41）中。

基于（6.42）和（6.44），β_2 足够大以满足 $\sqrt{\delta_{12}(\beta_2)} \leq \frac{1}{\sqrt{5}}\sqrt{\frac{1}{ndD_n}}$，且（iii）$k$ 足够大以满足 $Ck \geq ndD_n$时，以下（6.44）成立。由于C_{11} 是关于n的多项式，因此可以通过设置 $1-\beta_2 = \mathcal{O}(1/n^5)$来满足条件（i）。

$$\mathbb{E}\left[\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}}-\left(\delta_{12}(\beta_2)+\frac{C_7}{\sqrt{k}}\right)|\partial_\alpha f(x_k)|\right] \geq \mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_k)\|_2^2}{nd\sqrt{5D_0nd}}, \frac{\|\nabla f(x_k)\|_2}{2nd^2\sqrt{5D_1n}}\right\}\right]-\delta_{12}(\beta_2)\sqrt{\frac{D_0}{D_1 d}}-\frac{C_{11}}{\sqrt{k}}, \quad (6.44)$$

其中 $C_{11} = \frac{2\sqrt{2}\Delta_1}{(1-\beta_2)D_1n^3d^2\sqrt{5}} + \frac{8\sqrt{2}\Delta_1}{(\delta_{12}(\beta_2)+C_7)(1-\beta_2)\sqrt{D_1nd}} + C_7\sqrt{\frac{D_0}{D_1 d}}$ 在（6.48）中定义。

（6.44）的证明将在近期给出。最后，我们结合项（a）（6.1）和项（b）（6.42）的下界，得到：

$$\mathbb{E}\langle\nabla f(z_k), z_k - z_{k+1}\rangle = \mathbb{E}[(a)+(b)] \overset{(6.1)(6.42)}{\geq} \eta_k\left\{\mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_k)\|_2^2}{nd\sqrt{5D_0nd}}, \frac{\|\nabla f(x_k)\|_2}{2nd^2\sqrt{5D_1n}}\right\}\right]-\delta(\beta_2)\sqrt{D_0}-\frac{C}{\sqrt{k}}\right\}, \quad (6.45)$$

其中

$$\delta(\beta_2) := \delta_{12}(\beta_2)\frac{1}{\sqrt{D_1 d}}+\delta_{11}(\beta_2), \quad C = C_1+C_9+C_{11}, \quad (6.46)$$

并且 $\delta(\beta_2)$ 在 β_2 趋近于 1 时趋近于 0。常数 $\delta_{12}(\beta_2)$ 和 $\delta_{11}(\beta_2)$ 分别定义于式 (6.43) 和 (6.39) 中；常数 C_1 和 C_9 分别定义于式 (6.1) 和 (6.43) 中；C_{11} 稍后在式 (6.48) 中定义。引理4.2 的证明至此完成。

现在我们证明 (6.44)。我们将为 $\left(\partial_\alpha f(x_k)\right)^2 n\sqrt{\mathbb{E}_k(v_{\alpha,k})} - \left(\delta_{12}(\beta_2) + C\sqrt{k}\right)|\partial_\alpha f(x_k)|$ 推导一个下界，其中 $\delta_{12}(\beta_2)$ 和 C_7 的定义如前文 (6.43) 所述。注意，当 β_2 趋近于 1 时，$\delta_{12}(\beta_2)$ 趋近于 0；而随着 k 增大，$C\sqrt{k}$ 趋于 0。我们分两种情况讨论：

情形（a）：当 $|\partial_\alpha f(x_k)| \geq \frac{4\sqrt{2}\Delta_k}{(1-\beta_2)\sqrt{D_1 nd}}$： 在这种情况下，我们有

$$
\begin{aligned}
\frac{(\partial_\alpha f(x_k))^2}{\sqrt{\mathbb{E}_k(v_{\alpha,k})}} &= \frac{(\partial_\alpha f(x_k))^2}{\sqrt{(1-\beta_2)\left(\frac{1}{n}\sum_{i=0}^{n-1}(\partial_\alpha f_i(x_k))^2 + \sum_{j=1}^{k-1}\left(\partial_\alpha f_{\tau_{k-j}}(x_{k-j})\right)^2\beta_2^j\right)}} \\
&\overset{(B.11)}{\geq} \frac{(\partial_\alpha f(x_k))^2}{\sqrt{(1-\beta_2)\left(\left(\sqrt{|\partial_\alpha f(x_k)|^2 + \frac{D_0}{D_1 d}}\sqrt{D_1 nd}\right)^2 + \sum_{j=1}^{k-1}\left(\sqrt{|\partial_\alpha f(x_k)|^2 + \frac{D_0}{D_1 d}}\sqrt{D_1 nd} + \sum_{t=1}^{j}\Delta_{k-t}\right)^2\beta_2^j\right)}} \\
&\geq \frac{(\partial_\alpha f(x_k))^2}{\sqrt{(1-\beta_2)\left(\sum_{j=0}^{\infty}\beta_2^j\left(\left(|\partial_\alpha f(x_k)|^2 + \frac{D_0}{D_1 d}\right)D_1 nd^2 + 4\sqrt{2}j\Delta_k\sqrt{|\partial_\alpha f(x_k)|^2 + \frac{D_0}{D_1 d}}\sqrt{D_1 nd} + 8j^2\Delta_k^2\right)\right)}} \\
&\overset{\text{Lemma B.1}}{\geq} \frac{(\partial_\alpha f(x_k))^2}{\sqrt{D_1 nd^2\left(\left(|\partial_\alpha f(x_k)|^2 + \frac{D_0}{D_1 d}\right) + \sqrt{|\partial_\alpha f(x_k)|^2 + \frac{D_0}{D_1 d}}\frac{4\sqrt{2}\Delta_k}{(1-\beta_2)\sqrt{D_1 nd}} + \frac{16\Delta_k^2}{(1-\beta_2)^2 D_1 nd^2}\right)}} \\
&\overset{\text{Case (a)}}{\geq} \frac{(\partial_\alpha f(x_k))^2}{\sqrt{\frac{5}{2}D_1 nd^2\left(|\partial_\alpha f(x_k)|^2 + \frac{D_0}{D_1 d}\right)}}.
\end{aligned}
$$

根据 $(\partial_\alpha f(x_k))^2$ 的相对大小，我们分两种子情形讨论：

子情形 1：当 $(\partial_\alpha f(x_k))^2 \leq D_0 D_1 d$ 时：

$$
\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}} - \left(\delta_{12}(\beta_2) + \frac{C_7}{\sqrt{k}}\right)|\partial_\alpha f(x_k)| \geq \frac{(\partial_\alpha f(x_k))^2}{n\sqrt{5D_0 nd}} - \left(\delta_{12}(\beta_2) + \frac{C_7}{\sqrt{k}}\right)\sqrt{\frac{D_0}{D_1 d}}.
$$

子情形 2：当 $(\partial_\alpha f(x_k))^2 > D_0 D_1 d$：

$$
\begin{aligned}
\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}} - \left(\delta_{12}(\beta_2) + \frac{C_7}{\sqrt{k}}\right)|\partial_\alpha f(x_k)| &\geq \frac{|\partial_\alpha f(x_k)|}{n\sqrt{5D_1 nd^2}} - \left(\delta_{12}(\beta_2) + \frac{C_7}{\sqrt{k}}\right)|\partial_\alpha f(x_k)| \\
&= |\partial_\alpha f(x_k)|\left(\frac{1}{nd\sqrt{5D_1 n}} - \delta_{12}(\beta_2) - \frac{C_7}{\sqrt{k}}\right).
\end{aligned}
$$

结合这些子情况，当 β_2 足够大，使得 $\delta_{12}(\beta_2) \leq \frac{1}{4\sqrt{5ndD_1 n}}$，并且 k 足够大，使得 $C\sqrt{k} \leq \frac{1}{4\sqrt{5ndD_1 n}}$时，我们有

$$
\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}} - \left(\delta_{12}(\beta_2) + \frac{C_7}{\sqrt{k}}\right)|\partial_\alpha f(x_k)| \geq \min\left\{\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{5D_0 nd}}, \frac{|\partial_\alpha f(x_k)|}{2nd\sqrt{5D_1 n}}\right\} - \delta_{12}(\beta_2)\sqrt{\frac{D_0}{D_1 d}} - \frac{C_7}{\sqrt{k}}\sqrt{\frac{D_0}{D_1 d}}.
$$

情况 (b)：当 $|\partial_\alpha f(x_k)| < \frac{4\sqrt{2}\Delta_k}{(1-\beta_2)\sqrt{D_1 nd}}$ **时：** 在这种情况下：

$$\begin{aligned}\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}} - \left(\delta_{12}(\beta_2) + \frac{C_7}{\sqrt{k}}\right)|\partial_\alpha f(x_k)| &\geq \frac{|\partial_\alpha f(x_k)|}{2nd\sqrt{5D_1 n}} - \frac{2\sqrt{2}\Delta_1}{(1-\beta_2)D_1 n^2 d^2\sqrt{5}}\frac{1}{\sqrt{k}} \\ &\quad - \frac{8\sqrt{2}\Delta_1\delta_{12}(\beta_2)}{(1-\beta_2)\sqrt{D_1 nd}}\frac{1}{\sqrt{k}} - \frac{8\sqrt{2}\Delta_1 C_7}{(1-\beta_2)\sqrt{D_1 nd}}\frac{1}{k} \\ &\geq \min\left\{\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{5D_0 nd}}, \frac{|\partial_\alpha f(x_k)|}{2nd\sqrt{5D_1 n}}\right\} - \frac{C_{10}}{\sqrt{k}},\end{aligned}$$

其中我们定义：

$$C_{10} = \frac{2\sqrt{2}\Delta_1}{(1-\beta_2)D_1 n^2 d^2\sqrt{5}} + \frac{8\sqrt{2}\Delta_1(\delta_{12}(\beta_2)+C_7)}{(1-\beta_2)\sqrt{D_1 nd}}. \tag{6.47}$$

取期望并合并情形 (a) 与 (b)，我们得到

$$\mathbb{E}\left[\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}} - \left(\delta_{12}(\beta_2) + \frac{C_7}{\sqrt{k}}\right)|\partial_\alpha f(x_k)|\right] \geq \mathbb{E}\left[\min\left\{\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{5D_0 nd}}, \frac{|\partial_\alpha f(x_k)|}{2nd\sqrt{5D_1 n}}\right\}\right] - \delta_{12}(\beta_2)\sqrt{\frac{D_0}{D_1 d}} - \frac{C_{11}}{\sqrt{k}},$$

其中

$$C_{11} = C_{10} + C_7\sqrt{\frac{D_0}{D_1 d}}, \tag{6.48}$$

且 C_{10}、C_7 分别定义于（6.47）、（6.40）。利用该关系 $d(\partial_\alpha f(x_k))^2 \geq \|\nabla f(x_k)\|_2^2$ 与 $d|\partial_\alpha f(x_k)| \geq \|\nabla f(x_k)\|_1 \geq \|\nabla f(x_k)\|_2$，我们有

$$\mathbb{E}\left[\frac{(\partial_\alpha f(x_k))^2}{n\sqrt{\mathbb{E}_k(v_{\alpha,k})}} - \left(\delta_{12}(\beta_2) + \frac{C_7}{\sqrt{k}}\right)|\partial_\alpha f(x_k)|\right] \geq \mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_k)\|_2^2}{nd\sqrt{5D_0 nd}}, \frac{\|\nabla f(x_k)\|_2}{2nd^2\sqrt{5D_1 n}}\right\}\right] - \delta_{12}(\beta_2)\sqrt{\frac{D_0}{D_1 d}} - \frac{C_{11}}{\sqrt{k}}.$$

至此完成对（6.44）的证明。最后，当 $D_1 = 0$ 时，即当有界二阶矩条件成立时，只需对常数项作一些改动，便可得到类似的结论。在 $D_1 = 0$ 下的证明严格比当前证明更简单，因为它可归结为有界梯度情形。我们完成了引理4.2的全部证明。

□

7 定理3.1中的收敛结果的证明

现在我们给出定理 3.1 的证明。根据下降引理和望远镜求和，由引理 4.2 可得以下关系。设 $T_{\min}$ 足够大且满足引理4.2的条件，则有：

$$\begin{aligned}\sum_{k=T_{\min}}^{T}\eta_k\left\{\mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_k)\|_2^2}{nd\sqrt{5D_0 nd}}, \frac{\|\nabla f(x_k)\|_2}{2nd^2\sqrt{5D_1 n}}\right\}\right] - \delta(\beta_2)\sqrt{D_0} - \frac{C}{\sqrt{k}}\right\} &\leq \sum_{k=T_{\min}}^{T}\mathbb{E}\langle\nabla f(z_k), z_k - z_{k+1}\rangle \\ &\leq \mathbb{E}f(z_{T_{\min}}) - \mathbb{E}f(z_{T+1}) + \sum_{k=T_{\min}}^{T}\frac{L}{2}\mathbb{E}\|z_{k+1}-z_k\|_2^2 \\ &\leq \mathbb{E}f(z_{T_{\min}}) - f^* + \sum_{k=T_{\min}}^{T}\frac{L}{2}\mathbb{E}\|z_{k+1}-z_k\|_2^2. \end{aligned} \tag{7.1}$$

其中 C 在式（6.46）中定义。

由式（B.5）以及关系 $z_{k+1}-z_k = \frac{x_{k+1}-x_k-\beta_1^n(x_{k-n+1}-x_{k-n})}{1-\beta_1^n}$，可得

$$\begin{aligned}
\|z_{k+1}-z_k\|_2^2 &\le \frac{1}{(1-\beta_1^n)^2}\left(\|x_{k+1}-x_k\|_2+\beta_1^n\|x_{k-n+1}-x_{k-n}\|_2\right)^2\\
&\le \frac{1}{(1-\beta_1^n)^2}\cdot\left(\frac{d(1-\beta_1)}{\sqrt{1-\beta_2}(1-\frac{\beta_1}{\sqrt{\beta_2}})}\right)^2\cdot\left(\eta_k^2+\beta_1^{2n}\eta_{k-n}^2+2\beta_1^n\eta_k\eta_{k-n}\right)\\
&\le \frac{\eta_k^2}{(1-\beta_1^n)^2}\cdot\frac{d^2(1-\beta_1)^2}{(1-\beta_2)\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)^2}\cdot\left(1+\beta_1^{2n}(n+1)+2\beta_1^n\sqrt{n+1}\right)\\
&= \frac{1}{k}\cdot\frac{d^2(1-\beta_1)^2\eta_0^2\left(1+\beta_1^{2n}(n+1)+2\beta_1^n\sqrt{n+1}\right)}{(1-\beta_1^n)^2(1-\beta_2)\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)^2} := \frac{C_z}{k},
\end{aligned}\tag{7.2}$$

其中倒数第二个不等式成立，因为对于 $k \ge n+1$, $\sqrt{k-n}\le\sqrt{n}\sqrt{k+1}$

将（7.2）代入（7.1），我们得到

$$\sum_{k=T_{\min}}^{T}\frac{\eta_0}{\sqrt{k}}\left\{\mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_k)\|_2^2}{nd\sqrt{5D_0nd}},\frac{\|\nabla f(x_k)\|_2}{2nd^2\sqrt{5D_1n}}\right\}\right]-\delta(\beta_2)\sqrt{D_0}-\frac{C}{\sqrt{k}}\right\}$$
$$\le \mathbb{E}f(z_{T_{\min}})-f^*+\sum_{k=T_{\min}}^{T}\frac{L}{2}\cdot\frac{C_z}{k}.\tag{7.3}$$

利用 $\sum_{k=T_{\min}}^{T}\frac{1}{\sqrt{k}}\ge 2(\sqrt{T}-\sqrt{T_{\min}})$ 和 $\sum_{k=T_{\min}}^{T}\frac{1}{k}\le\log\frac{T+1}{T_{\min}}$，我们有

$$\begin{aligned}
\min_{k\in[T_{\min},T]}\mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_k)\|_2^2}{nd\sqrt{5D_0nd}},\frac{\|\nabla f(x_k)\|_2}{2nd^2\sqrt{5D_1n}}\right\}\right] &\le \frac{\delta(\beta_2)\sqrt{D_0}\sum_{k=T_{\min}}^{T}\frac{\eta_0}{\sqrt{k}}+\left(C\eta_0+\frac{LC_z}{2}\right)\log\frac{T+1}{T_{\min}}+\mathbb{E}f(z_{T_{\min}})-f^*}{\sum_{k=T_{\min}}^{T}\frac{\eta_0}{\sqrt{k}}}\\
&\le \left(C\eta_0+\frac{LC_z}{2}\right)\frac{\log T}{\sqrt{T}}+\delta(\beta_2)\sqrt{D_0}.
\end{aligned}$$

对于足够大的 T。

经过适当的缩放后，我们得到如下关系。

$$\begin{aligned}
\min_{k\in[1,T]}\mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_k)\|_2^2}{\sqrt{D_0}},\frac{\|\nabla f(x_k)\|_2}{2\sqrt{dD_1}}\right\}\right] &\le n^{1.5}d^{1.5}\sqrt{5}\left(C\eta_0+\frac{LC_z}{2}\right)\frac{\log T}{\sqrt{T}}+n^{1.5}d^{1.5}\sqrt{5}\delta(\beta_2)\sqrt{D_0},\\
&= \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)+\mathcal{O}(\delta(\beta_2)\sqrt{D_0}).
\end{aligned}\tag{7.4}$$

至此完成了定理3.1的证明。 □

定理3.5中的发散结果证明

现在我们给出发散结果的证明。我们主要考虑一个 $=1$，这样能使结果更简洁。不过，同样的证明过程也适用于任意有限正一个 >0。我们将 $x_{k,\ i}$ 记为 x 在第 k 次外循环和第 i 次内循环时的值。我们考虑循环更新顺序，其中 $f_i(x)$ 在第 k 次外循环中按 $f_0(x)$、$f_1(x)$、$\cdots$、$f_{n-1}(x)$ 的顺序采样。我们只给出循环更新顺序下 Adam 的证明，这有助于揭示关键见解。随机采样的证明遵循相同的过程，并得出类似结论；为简洁起见，此处省略。首先，我们证明以下断言：

断言： 对于任意 $n\ge 3$，存在图5 (a)所示的一个橙色区域，使得只要 $x_{k,0}=1$，Adam 在橙色区域内取任意 (β_1,β_2) 组合都会得到 $x_{k+1,0}>1$。

Now 下面证明该断言。对于函数 (3.1)，Adam 的更新规则如下所示 .

$$x_{k,1} = \left(x_{k,0}+\delta_{k,0}\right), \quad \delta_{k,0} = -\frac{\eta_0}{\sqrt{k}}\left(\frac{n(1-\beta_1)+\beta_1 m_{k-1,n-1}}{\sqrt{(1-\beta_2)n^2+\beta_2 v_{k-1,n-1}}}\right), \tag{8.1}$$

$$x_{k,i+1} = \left(x_{k,i}+\delta_{k,i}\right), \quad i=1,\cdots,n-1; \tag{8.2}$$

其中 $\delta_{k,i} = -\frac{\eta_0}{\sqrt{k}}\left(\frac{(1-\beta_1)\sum_{j=0}^{i-1}(-1)\beta_1^j+1-\beta_1\beta_1^i+\beta_1^{i+1}m_{k-1,n-1}}{\sqrt{(1-\beta_2+\beta_2 v_{k,i-1}}}\right)$ 0

$nm,)v,$ 。

我们将总移动量 $\sum^{n-1}{}_{i=0}\delta_{k,i}$ 分解为以下三个项。

$$\begin{aligned}\sum_{i=0}^{n-1}\delta_{k,i} &= \frac{\eta_0}{\sqrt{k}}\underbrace{\left(-\frac{\beta_1 m_{k-1,n-1}}{\sqrt{(1-\beta_2)n^2+\beta_2 v_{k-1,n-1}}}-\frac{\beta_1^2 m_{k-1,n-1}}{\sqrt{(1-\beta_2)+\beta_2 v_{k,0}}}-\cdots-\frac{\beta_1^n m_{k-1,n-1}}{\sqrt{(1-\beta_2)+\beta_2 v_{k,n-2}}}\right)}_{\text{(I)}}\\ &+\frac{\eta_0}{\sqrt{k}}\underbrace{\left(\frac{1-\beta_1}{\sqrt{(1-\beta_2)+\beta_2 v_{k,0}}}+\frac{(1-\beta_1)+\beta_1(1-\beta_1)}{\sqrt{(1-\beta_2)+\beta_2 v_{k,1}}}+\cdots+\frac{(1-\beta_1)\sum_{j=0}^{n-2}\beta_1^j}{\sqrt{(1-\beta_2)+\beta_2 v_{k,n-2}}}\right)}_{\text{(II)}}\\ &+\frac{\eta_0}{\sqrt{k}}\underbrace{\left(-\frac{n(1-\beta_1)}{\sqrt{(1-\beta_2)n^2+\beta_2 v_{k-1,n-1}}}-\frac{n(1-\beta_1)\beta_1}{\sqrt{(1-\beta_2)+\beta_2 v_{k,0}}}-\cdots-\frac{n(1-\beta_1)\beta_1^{n-1}}{\sqrt{(1-\beta_2)+\beta_2 v_{k,n-2}}}\right)}_{\text{(III)}}.\end{aligned}$$

我们将证明，对于某些 β_1 和 β_2：(I)、(II) >0 和 (III) <0。此外，(I) 和 (II) 的作用超过 (III)，从而导致发散。首先，我们证明当 β_1 较小时，$m_{k-1,n-1}<0$，这意味着 (I)>0。

$$-m_{k-1,n-1} = (1-\beta_1)\sum_{j=0}^{n-2}\beta_1^j-(1-\beta_1)\beta_1^{n-1}n-\beta_1^n m_{k-2,n-1} = \left[(1-\beta_1^{n-1})-(1-\beta_1)\beta_1^{n-1}n\right]\sum_{j=0}^{k}(\beta_1^n)^j.$$

当 β_1 较小时，我们有 $(1-\beta_1^{n-1})>(1-\beta_1)\beta_1^{n-1}n$，这意味着 $-m_{k-1,n-1}>0$。对于这些 β_1的取值，我们有 (I) >0。现在我们推导 (II) 的下界。

$$\begin{aligned}\text{(II)} &\geq \frac{1-\beta_1}{\sqrt{1+\beta_2 n^2}}+\frac{(1-\beta_1)+\beta_1(1-\beta_1)}{\sqrt{1+\beta_2^2 n^2}}+\cdots+\frac{(1-\beta_1)\sum_{j=0}^{n-2}\beta_1^j}{\sqrt{1+\beta_2^{n-1}n^2}}\\ &= \frac{1-\beta_1}{\sqrt{1+\beta_2 n^2}}+\frac{1-\beta_1^2}{\sqrt{1+\beta_2^2 n^2}}+\cdots+\frac{1-\beta_1^{n-1}}{\sqrt{1+\beta_2^{n-1}n^2}}.\end{aligned}$$

该不等式源于以下事实：$v_{k,0}\leq n^2$。由于当 β_2 较小时，$\beta_2^j n^2$ 也较小，且j 接近n，因此存在某个足够小的 β_2，使得对于至少一个$j<n$，$\beta_2^j n^2\leq 0$。1。对于这些足够小的 β_2，我们保留$j\geq\log^{\beta_2}(0.1/n^2)$ 这一项，并舍去其余项。根据基本微积分，我们有

$$\text{(II)} \geq \left(n-1-\min\left\{n-1,\log_{\beta_2}\left(\frac{1}{10n^2}\right)\right\}\right)\frac{1-\beta_1^{\min\left\{n-1,\log_{\beta_2}\left(\frac{1}{10n^2}\right)\right\}}}{\sqrt{1+\max\left\{0.1,\beta_2^{n-1}n^2\right\}}}$$

现在我们推导 |(III)| 的一个上界。

$$
\begin{aligned}
|(\mathrm{III})| &\leq \left|\frac{n(1-\beta_1)}{\sqrt{(1-\beta_2)n^2+\beta_2 v_{k-1,n-1}}}\right| + \left|\frac{n(1-\beta_1)\beta_1}{\sqrt{(1-\beta_2)+\beta_2 v_{k,0}}}\right| + \cdots + \left|\frac{n(1-\beta_1)\beta_1^{n-1}}{\sqrt{(1-\beta_2)+\beta_2 v_{k,n-2}}}\right| \\
&\leq \frac{1-\beta_1}{\sqrt{1-\beta_2}}\left(1+n(\sum_{j=1}^{n-1}\beta_1^j)\right) = \frac{1-\beta_1}{\sqrt{1-\beta_2}} + \frac{\beta_1(1-\beta_1^{n-1})}{\sqrt{1-\beta_2}}n \\
&\leq \frac{1-\beta_1}{\sqrt{1-\beta_2}} + \frac{\beta_1}{\sqrt{1-\beta_2}}n.
\end{aligned}
$$

此外，我们将使用足够小的步长 $\eta_0 \leq 2\sqrt{(1-\beta_2)\beta_2^n}$，以确保迭代点保持在线性区域内，因此上述关系对轨迹中的所有迭代点都成立。总之，当满足以下条件时就会发生发散：

g

$$
(\mathbf{C1}): \left(n-1-\min\left\{n-1, \log_{\beta_2}(\frac{1}{10n^2})\right\}\right) \frac{1-\beta_1^{\min\left\{n-1,\log_{\beta_2}(\frac{1}{10n^2})\right\}}}{\sqrt{1+\max\left\{0.1, \beta_2^{n-1}n^2\right\}}} \geq \frac{1-\beta_1}{\sqrt{1-\beta_2}} + \frac{\beta_1}{\sqrt{1-\beta_2}}n; \tag{8.3}
$$

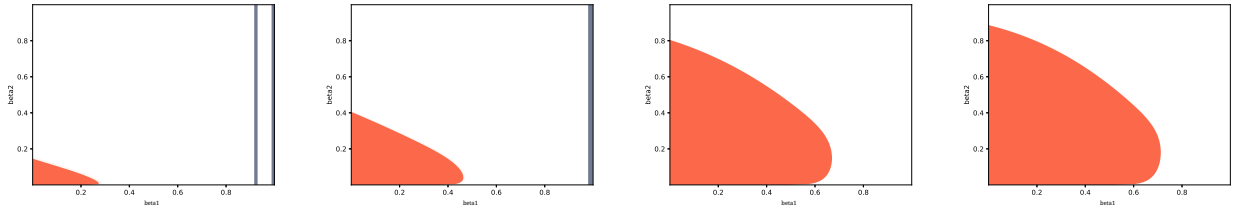
$$
(\mathbf{C2}): (1-\beta_1^{n-1}) > (1-\beta_1)\beta_1^{n-1}n; \tag{8.4}
$$

$$
(\mathbf{C3}): \eta_0 \leq 2\sqrt{(1-\beta_2)\beta_2^n}. \tag{8.5}
$$

至此，Adam 的迭代点和函数值发散的证明就完成了。按照类似的步骤，将初始化改为 $x < 0$，也可以证明梯度发散。最后，上述证明过程适用于任意有限正数 $a > 0$，而不仅仅是 $a = 1$。当 $a = 1/(n-1)^2$,时，我们的反例（3.1）满足 $D_1 = 2n^2, D_0 = 0$，因此对于任意函数类，发散都会发生$\mathcal{F}_{nL}$

$,^{D_0,D_1}(\mathbb{R}^d)$ 其中 $D_1 \geq 2n^2$ 且 $D_0 \geq 0$。至此，定理 3.5 的证明完成。□

借助 NumPy，我们将(**C1**) 和(**C2**) 成立的区域可视化。结果如图8所示。我们用橙色表示(**C1**) 成立的区域，白色表示其余区域。对于(**C2**)，我们用灰色竖线标出$(1-\beta_1^{n-1}) = (1-\beta_1)\beta_1^{n-1}n$所在的直线。请注意，这个方程有两个解：一个解为 $\beta_1 = 1$，另一个解位于 $0 < \beta_1 < 1$，这就是图中有两条竖线的原因。(**C2**) 在左侧那条灰色竖线的左边成立。



(a) $n = 5$　(b) $n = 10$　(c) $n = 50$　(d) $n = 100$

图 8：该图展示了 (8.3) 和 (8.4) 中 (C1) 与 (C2) 同时成立的区域。橙色表示 (C1) 成立的区域，白色表示其余区域。灰色垂直线表示 (C2) 的边界。请注意，(C2) 中的方程有两个解：一个解为 $\beta_1 = 1$，另一个解位于 $0 < \beta_1 < 1$，这就是图中有两条竖线的原因。(C2) 在左侧那条灰色垂直线的左边成立。该区域通过在 NumPy 中求解条件 (8.3)、(8.4) 和 (8.5) 绘制而成。

这两个区域的交集将是 Adam 发散的区域，实际上就是图 8 中的橙色区域。发散区域的大小随 n 增大。

与 $\gamma_1(n)$ 在定理 3.1中的关系。根据定理 3.1， $\gamma_1(n)$ 至少处于 $1-\mathcal{O}(n^{-5})$ 的量级。结合图 8，不难看出 $\gamma_1(n)$ 始终大于橙色区域的上界，因此并不存在矛盾。

9 局限性与未来方向

• **对临界边界进行更细致的刻画**。 综合我们的发散与收敛理论，可以得出结论：至少存在一条临界边界 (β_1^*, β_2^*) 用于划分双收敛相变。然而，我们尚未完全确定边界的确切数量和形状。图 1 (c, d) 中的实验结果表明，边界可能只有一条，其形状可能类似于图 1 (b) 中的蓝色区域。我们在此仅指出这条边界的存在，并将更精确的刻画留待未来研究。

• **Adam 与随机梯度下降的比较**。 在这项工作中，我们关注收敛这一基础问题。一个值得探讨的问题是验证 Adam 相对于随机梯度下降的优势。确定 Adam 何时以及为何收敛得更快，本身就是一个独立的研究课题，我们将其留待未来研究。

• **广义 Lipschitz 条件**。 近期一系列理论研究放宽了假设 2.1 中的标准 Lipschitz 条件，将其推广为某些广义 Lipschitz 条件（例如， [Zhang 等人，2019年，Li 等人，2023年，Wang 等人， 2024]）。如第 2.1 节前文所述，广义 Lipschitz 条件主要有助于改进 Adam 的 定量 分析，例如得出更精确的收敛速率。与此同时，本工作的核心关注点是一个更基础的问题：刻画 Adam 的 定性 行为，例如从发散到收敛的相变。为此，标准 Lipschitz 条件已足够。不过，将 Lipschitz 条件推广是一个在理论上颇具吸引力的未来研究方向。

• **更紧的上界**。 最后，我们注意到，在我们的收敛上界中， $\mathcal{O}\left(\log T\sqrt{T}\right)$ 项的乘法常数阶为 $\mathcal{O}(d^4n^{4.5})$，其中d是问题维度d，而n是小批次的数量。同样， β_2的阈值阶为n^5和$n^{5.5}$。我们并未声称这些幂次是紧的，通过更精细的分析，有可能降低这些依赖关系。我们将其留作一个有趣的未来研究方向。

• **和的 β_2最优组合**。 我们的结果指出，存在一个安全区域(β_1, β_2)，但我们尚未确定在安全区域内，和 β_2 的哪种组合能够带来最优性能。近期研究 [Zhao 等人，2024年，Orvieto 和 Gower， 2025] 指出，（当二者都足够大时）通常能够带来最优性能，并对此给出了初步解释。近期研究进一步探讨了这一发现 [Fernández-Hernández 等人，2026，Cattaneo 和 Shigida，2026]。确定最优超参数是一个独立的课题，需要投入更多精力。我们将其留作一个重要的未来研究方向。

10 条结论

在这项工作中，我们探讨了 Adam 的收敛性。当 β_2 较大时，我们证明 Adam 在任意 $\beta_1 < \sqrt{\beta_2}$ 下都能收敛。当 β_2 较小时，我们进一步表明，对于广泛范围的 β_1，Adam 可能会发散至无穷大。相变的临界边界 (β_1^*, β_2^*) 取决于具体问题，尤其取决于批量大小。我们的结果为 Adam 优化器提供了严谨的理论依据，也为如何调整 β_1 和 β_2 提供了实用建议。

致谢

Yushun Zhang 感谢 Naichen Shi、Bohan Wang 以及 NeurIPS 2022 的匿名评审，他们对论文会议版本的讨论和贡献十分宝贵。Yushun Zhang 感谢 Anthony Man-cho So 教授就有界方差条件进行了宝贵的讨论。Yushun Zhang 感谢 Lexing Ying 教授就 SignSGD 的发散与批量大小之间的关系进行了讨论。

参考资料

R. Agarwal、D. Schuurmans 和 M. Norouzi。《离线强化学习的乐观视角》。载于国际机器学习会议，页码 104–114。PMLR，2020。

A. Alacaoglu, Y. Malitsky, P. Mertikopoulos, and V. Cevher. Adam 类算法的新遗憾分析. 载于国际机器学习会议，第 202–210 页。PMLR，2020。Z. Allen-Zhu. Katyusha：首个对随机梯度方法的直接加速算法. 机器学习研究期刊，18(221):1–51，2018。J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang 等. 通义千问技术报告. arXiv 预印本 arXiv:2309.16609，2023。A. Barakat and P. Bianchi. Adam 算法在非凸随机优化中的收敛性与动力学行为. SIAM 优化杂志，31(1):244–274，2021。J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. SignSGD：面向非凸问题的压缩优化. 载于国际机器学习会议，第 560–569 页。PMLR，2018。D. Bertsekas and J. N. Tsitsiklis. 神经动态规划. Athena Scientific，1996。D. P. Bertsekas. 非线性规划. Journal of the Operational Research Society，48(3):334–334，1997。D. P. Bertsekas and J. N. Tsitsiklis. 含误差梯度法中的梯度收敛. SIAM 优化杂志，10(3):627–642，2000。L. Bottou, F. E. Curtis, and J. Nocedal. 大规模机器学习的优化方法. SIAM review，60(2):223–311，2018。T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell 等. 语言模型是少样本学习器. arXiv 预印本 arXiv:2005.14165，2020。S. Bubeck. 凸优化：算法与复杂性. Foundations and trends in Machine Learning，8(3-4):231–357，2015。M. D. Cattaneo and B. Shigida. 小批量噪声对 Adam 隐式偏置的影响. arXiv 预印本 arXiv:2602.01642，2026。C. Chen, L. Shen, F. Zou, and 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。J. Chen, D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu. 缩小深度神经网络训练中自适应梯度方法的泛化差距. arXiv 预印本 arXiv:1806.06763，2018a。X. Chen, S. Liu, R. Sun, and M. Hong. 关于一类非凸优化 Adam 类算法的收敛性. arXiv 预印本 arXiv:1808.02941，2018b。

J. M. Cohen, A. Damian, A. Talwalkar, J. Z. Kolter, and J. D. Lee. 通过中心流理解深度学习中的优化. arXiv 预印本 arXiv:2410.24206，2024。A. B. Da Silva and M. Gazeau. 用于建模一阶自适应算法的一般微分方程组. 机器学习研究杂志，21(1):5072–5113，2020。R. Das, N. Agarwal, S. Sanghavi, and I. S. Dhillon. 迈向量化 Adam 的预条件效应. arXiv 预印本 arXiv:2402.07114，2024。S. De, A. Mukherjee, and E. Ullah. RMSProp 和 Adam 在非凸优化中的收敛保证及与 Nesterov 加速的实证比较. arXiv 预印本 arXiv:1807.06766，2018。A. Défossez, L. Bottou, F. Bach, and N. Usunier. Adam 和 AdaGrad 的简明收敛证明. 机器学习研究汇刊，2022。ISSN 2835-8856。URLhttps://openreview.net/forum?id=ZPQhzTSWA7。L. Deng. 面向机器学习研究的手写数字图像 MNIST 数据库. IEEE 信号处理杂志，29(6):141–142，2012。S. Dereich and A. Jentzen. Adam 优化器的收敛速率. arXiv 预印本 arXiv:2407.21078，2024。S. Dereich, T. Do, A. Jentzen, and P. von Wurstemberger. Adam 对称性定理：随机 Adam 优化器收敛性的刻画. arXiv 预印本 arXiv:2511.06675，2025a。S. Dereich, A. Jentzen, and S. Kassing. Adam 算法的常微分方程近似：一般情形与过参数化情形. arXiv 预印本 arXiv:2511.04622，2025b。S. Dereich, A. Jentzen, and A. Riekert. Adam 优化器的高阶尖锐收敛速率. arXiv 预印本 arXiv:2504.19426，2025c。K. Ding, N. Xiao, and K.-C. Toh. 深度学习中采用解耦权重衰减的 Adam 系列方法. arXiv 预印本 arXiv:2310.08858，2023。A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and 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。URLhttps://openreview.net/forum?id=YicbFdNTTy。T. Dozat. 将 Nesterov 动量融入 Adam. 2016。R. Durrett.《概率论：理论与例子》，第 49 卷。剑桥大学出版社，2019。A. Fernández-Hernández, C. Pérez-Corral, J. I. Mestre, M. F. Dolz, and E. S. Quintana-Ortí. 为什么 Adam 的效果更好 $\beta_1 = \beta_2$：缺失的梯度尺度不变性原理. arXiv 预印本 arXiv:2601.21739，2026。S. Gadat and I. Gavra. 非凸景观中随机自适应算法的渐近研究. arXiv 预印本 arXiv:2012.05640，2020。S. Ghadimi and G. Lan. 非凸随机规划的一阶与零阶随机方法. SIAM 优化杂志，23(4):2341–2368，2013。S. Ghadimi, G. Lan, and H. Zhang. 非凸随机复合优化的小批量随机逼近方法.《数学规划》，155(1):267–305，2016。Z. Guo, Y. Xu, W. Yin, R. Jin, and T. Yang. Adam 系列算法的一种新颖收敛分析. arXiv 预印本 arXiv:2112.03459，2021。

K. He、X. Zhang、S. Ren 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，第 770–778 页，2016。G. Hinton、N. Srivastava 和 K. Swersky。机器学习神经网络第 6a 讲：小批量梯度下降概述。Cited on，14(8):2，2012。Y. Hong 和 J. Lin。无界梯度和仿射方差噪声下 Adam 的高概率收敛。arXiv 预印本 arXiv:2311.02000，2023。Y. Hong 和 J. Lin。放宽假设下 Adam 在随机优化中的收敛性。arXiv 预印本 arXiv:2402.03982，2024。F. Huang、J. Li 和 H. Huang。Super-Adam：更快、更通用的自适应梯度框架。arXiv 预印本 arXiv:2106.08208，2021。ICLR。ICLR 2018 日程概览。https://iclr.cc/Conferences/2018/ScheduleOverview，2018 年 4 月。该页面将《On the Convergence of Adam and Beyond》列为最佳论文奖获奖作品。ICLR。公布 ICLR 2015 时间检验奖得主。https://blog.iclr.cc/2025/04/14/announcing-the-test-of-time-award-winners-from-iclr-2015/，2025 年 4 月。访问日期：2025年5月27日。H. Iiduka。无需利普希茨光滑性、使用接近 1 的超参数对 Adam 进行理论分析。arXiv 预印本 arXiv:2206.13290，2022。P. Isola、J.-Y. Zhu、T. Zhou 和 A. A. Efros。利用条件对抗网络实现图像到图像转换。载于 IEEE 计算机视觉与模式识别会议论文集，第 1125–1134 页，2017。K. Jiang、D. Malik 和 Y. Li。自适应优化如何影响局部神经网络几何？《神经信息处理系统进展》，36:8305–8384，2023a。W. Jiang、D. Yu、S. Yang、W. Yang 和 L. Zhang。动量更新下基于符号的方法的改进分析。arXiv 预印本 arXiv:2507.12091，2025。Y. Jiang、J. Liu、D. Xu 和 D. P. Mandic。UAdam：用于非凸随机优化的统一 Adam 类算法框架。arXiv 预印本 arXiv:2305.05675，2023b。D. P. Kingma 和 J. Ba。Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980，2014。A. Krizhevsky、G. Hinton 等。通过微小图像学习多层特征。2009。F. Kunstner、A. Milligan、R. Yadav、M. Schmidt 和 A. Bietti。重尾类别不平衡，以及 Adam 为何在语言模型上优于梯度下降。《神经信息处理系统进展》，37:30106–30148，2024。H. Li、A. Rakhlin 和 A. Jadbabaie。放宽假设下 Adam 的收敛性。《神经信息处理系统进展》，36:52166–52196，2023。H. Li、Y. Dong 和 Z. Lin。以$o(\sqrt{d}/t^{1/4})$ 衡量的 RMSProp 及其动量扩展的收敛速率。以l_1 范数衡量。《机器学习研究期刊》，26(131):1–25，2025a。网址http://jmlr.org/papers/v26/24-0523.html。H. Li、Y. Dong 和 Z. Lin。以$o(\sqrt{d}/k^{1/4})$ 衡量的 AdamW 收敛速率。以l_1 范数衡量。CoRR，2025b。X. Li、H. Wen 和 K. Lyu。Adam 会降低一种独特的尖锐度：极小化器流形附近的理论洞见。arXiv 预印本 arXiv:2511.02773，2025c。Y. Liang、M. He、J. Liu 和 D. Xu。非凸目标函数下 Adam 的收敛性：放宽超参数与非遍历情形。《机器学习》，114(3):75，2025。

A. Liu、B. Feng、B. Xue、B. Wang、B. Wu、C. Lu、C. Zhao、C. Deng、C. Zhang、C. Ruan 等。Deepseek-v3 技术报告。arXiv 预印本 arXiv:2412.19437，2024。L. Liu、H. Jiang、P. He、W. Chen、X. Liu、J. Gao 和 J. Han。关于自适应学习率方差及其拓展。载于国际学习表征会议，2020a。网址 https://openreview.net/forum?id=rkgz2aEKDr。Y. Liu、Y. Gao 和 W. Yin。带动量的随机梯度下降的改进分析。arXiv 预印本 arXiv:2007.07989，2020b。L. Luo、Y. Xiong 和 Y. Liu。具有学习率动态界的自适应梯度方法。载于国际学习表征会议，2019a。网址 https://openreview.net/forum?id=Bkg3g2R9FX。L. Luo、Y. Xiong、Y. Liu 和 X. Sun。具有学习率动态界的自适应梯度方法。arXiv 预印本 arXiv:1902.09843，2019b。Z.-Q. Luo。自适应学习率下线性前馈网络 LMS 算法的收敛性。《神经计算》，3(2):226–245，1991。S. Malladi、K. Lyu、A. Panigrahi 和 S. Arora。自适应梯度算法的随机微分方程与缩放规则。《神经信息处理系统进展》，35:7697–7711，2022。M. Marek、S. Lotfi、A. Somasundaram、A. G. Wilson 和 M. Goldblum。语言模型的小批量训练：原始随机梯度下降何时有效，以及梯度累积为何浪费。arXiv 预印本 arXiv:2507.07101，2025。A. Nemirovski、A. Juditsky、G. Lan 和 A. Shapiro。用于随机规划的稳健随机逼近方法。《SIAM 优化杂志》，19(4):1574–1609，2009。OpenAI。ChatGPT（11 月 30 日版本）[大语言模型], 2022。网址https://chat.openai.com/。访问日期：2025-05-15。A. Orvieto 和 R. Gower。探寻 Adam 的秘方。arXiv 预印本 arXiv:2505.21829，2025。H. Peng、S. Qin、Y. Yu、F. Jiang、H. Wang 和 Z. Lin。从类符号下降视角出发的 Adam 简明收敛证明。arXiv 预印本 arXiv:2507.05966，2025。B. T. Polyak 和 A. B. Juditsky。通过平均加速随机逼近。《SIAM 控制与优化杂志》，30(4):838–855，1992。T. Porian、M. Wortsman、J. Jitsev、L. Schmidt 和 Y. Carmon。解决语言模型计算最优缩放中的差异。《神经信息处理系统进展》，37:100535–100570，2024。A. Radford、L. Metz 和 S. Chintala。利用深度卷积生成对抗网络进行无监督表征学习。arXiv 预印本 arXiv:1511.06434，2015。S. J. Reddi、S. Kale 和 S. Kumar。Adam 及其拓展的收敛性。载于国际学习表征会议，2018。网址 https://openreview.net/forum?id=ryQu7f-RZ。M. Schmidt 和 N. L. Roux。强增长条件下随机梯度下降的快速收敛。arXiv 预印本 arXiv:1308.6370，2013。M. Schmidt、N. Le Roux 和 F. Bach。利用随机平均梯度最小化有限和。《数学规划》，162(1):83–112，2017。

G. Scholar。 https://scholar.google.com/scholar?oi=bibs&hl=en&cites=16194105527543080940,
1
0
5
6
1
6
4
2
7
2
5
7
0
8
9
2
4
0
0
6,
8
7
7
6
2
1
5
5
3
0
6
7
2
3
3
8
5
3
6,
1
7
6
0
0
7
1
5
4
0
5
5
5
1
8
7
7
5
0
7
9
7,
8
4
9
8
5
1
4
2
0
9
8
8
1
2
2
2
8
2
4,
3
9
3
1
5
6
0
505490345047,16226219632797774343,16111079307796299296,13264970321150126215,11869859376987161805,11430597276676453789,5669591665425275538,19445508016579376222,2752978001160552567,6554547627726037201,1485036471182801446,3567161356552128558,1757318821126852192,9521511829532092552,5215814671580410887,
谷歌学术，2025年。N. Shi、D. Li、M. Hong 和 R. Sun。RMSProp 在超参数设置适当时能够收敛。载于《国际学习表征会议》，2020。M. V. Solodov。步长下界为正的增量梯度算法。《计算优化与应用》，11(1):23–35，1998。T. Sre´ckovi´c、J. Geiping 和 A. Orvieto。批量大小是问题所在吗？重新审视语言建模中的 Adam-SGD 差距。arXiv 预印本 arXiv:2506.12543，2025年。T. Sun、Q. Wang、D. Li 和 B. Wang。动量在较弱假设下的能保证 SignSGD 收敛。载于《国际机器学习会议》，第33077–33099页。PMLR，2023年。H. Touvron、L. Martin、K. Stone、P. Albert、A. Almahairi、Y. Babaei、N. Bashlykov、S. Batra、P. Bhargava、S. Bhosale 等。Llama 2：开放基础模型与微调聊天模型。arXiv 预印本 arXiv:2307.09288，2023年。A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin。注意力机制就是你所需要的一切。载于《神经信息处理系统进展》，第5998–6008页，2017年。S. Vaswani、F. Bach 和 M. Schmidt。过参数化模型的快速与更快速随机梯度下降收敛及一种加速感知机。载于《第22届国际人工智能与统计会议》，第1195–1204页。PMLR，2019年。B. Wang、J. Fu、H. Zhang、N. Zheng 和 W. Chen。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。B. Wang、Y. Zhang、H. Zhang、Q. Meng、R. Sun、Z.-M. Ma、T.-Y. Liu、Z.-Q. Luo 和 W. Chen。非均匀光滑性下 Adam 自适应性的可证明性。载于《第30届 ACM SIGKDD 知识发现与数据挖掘会议论文集》，第2960–2969页，2024年。N. Xiao、X. Hu、X. Liu 和 K.-C. Toh。具有收敛保证的 Adam 系列非光滑优化方法。《机器学习研究期刊》，25(48):1–53，2024年。S. Xie 和 Z. Li。AdamW 的隐式偏置：-范数约束优化。载于《国际机器学习会议》，第54488–54510页。PMLR，2024年。S. Xie、M. A. Mohamadi 和 Z. Li。Adam 通过逐坐标自适应性利用损失景观的-几何性质。arXiv 预印本 arXiv:2410.08198，2024年。Y. Yan、T. Yang、Z. Li、Q. Lin 和 Y. Yang。深度学习随机动量方法的统一分析。arXiv 预印本 arXiv:1808.10396，2018。H. Yu、R. Jin 和 S. Yang。关于通信高效动量随机梯度下降在分布式非凸优化中的线性加速分析。载于《国际机器学习会议》，第7184–7193页。PMLR，2019年。M. Zaheer、S. Reddi、D. Sachan、S. Kale 和 S. Kumar。非凸优化的自适应方法。载于 S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett（编），《神经信息处理系统进展》，第31卷。Curran Associates, Inc.，2018。网址 https://proceedings.neurips.cc/paper/2018/file/90365351ccc7437a1309dc64e4db32a3-Paper.pdf。

H. Zhang、D. Morwani、N. Vyas、J. Wu、D. Zou、U. Ghai、D. Foster 和 S. Kakade。预训练中的临界批次大小如何变化？arXiv 预印本 arXiv:2410.21676，2024a。J. Zhang、T. He、S. Sra 和 A. Jadbabaie。为什么梯度裁剪能加速训练：自适应性的理论依据。arXiv 预印本 arXiv:1905.11881，2019年。Q. Zhang、Y. Zhou 和 S. Zou。仿射噪声方差下广义光滑非凸优化中 RMSProp 和 Adam 的收敛保证。arXiv 预印本 arXiv:2404.01436，2024b。Y. Zhang、C. Chen 和 Z.-Q. Luo。Adam 会收敛吗？如果会，何时收敛？载于《ICLR 博客专栏》，2022a。网址 https://iclr-blog-track.github.io/2022/03/25/does-adam/。https://iclr-blog-track.github.io/2022/03/25/does-adam/。Y. Zhang、C. Chen、N. Shi、R. Sun 和 Z.-Q. Luo。Adam无需对更新规则做任何修改即可收敛。《神经信息处理系统进展》，35:28386–28399，2022b。Y. Zhang、C. Chen、T. Ding、Z. Li、R. Sun 和 Z. Luo。为什么 Transformer 需要 Adam：海森矩阵视角。《神经信息处理系统进展》，37:131786–131823，2024c。R. Zhao、D. Morwani、D. Brandfonbrener、N. Vyas 和 S. Kakade。解构语言模型优质优化器的构成要素。arXiv 预印本 arXiv:2407.07972，2024年。D. Zhou、J. Chen、Y. Cao、Y. Tang、Z. Yang 和 Q. Gu。非凸优化自适应梯度方法的收敛性。arXiv 预印本 arXiv:1808.05671，2018a。Z. Zhou、Q. Zhang、G. Lu、H. Wang、W. Zhang 和 Y. Yu。AdaShift：自适应学习率方法的去相关与收敛。arXiv 预印本 arXiv:1810.00143，2018b。F. Zou、L. Shen、Z. Jie、W. Zhang 和 W. Liu。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》，第11127–11135页，2019年。

附录

附录目录

更多讨论 **44**A.1 更多相关工作 .44A.2 关于有界方差条件 (2.3) 的更多讨论 .45

B 更多符号和有用引理 **46****C 引理 6.1 的证明** **49****D 定理 3.3 的证明** **51**D.1 引理 D.1 的证明 .52D.2 引理 D.2 的证明 .52D.3 引理 D.3 的证明 .53D.4 引理 D.4 的证明 .54D.5 引理 D.5 的证明 .56D.6 引理 D.6 的证明 .57D.7 证明主体 .60

E 关于偏差校正项与非零 ϵ **的讨论** **64**

F 实验设置 **66**

A 更多讨论

A.1 更多相关工作

关于Adam变体。自Reddi 等人 [2018] 指出Adam的发散问题以来，一条活跃的研究路线一直试图设计可被证明收敛的新Adam变体。例如，Zou等人 [2019], Gadat和Gavra [2020], Chen等人 [2018b， 2021] 将常数超参数替换为依赖于迭代点的超参数，例如 β_{1k} 或 β_{2k}。AMSGrad [Reddi 等人， 2018] 和AdaFom [Chen 等人，2018b]将$\{\mathrm{v}_k\}$ 修改为非递减序列。Iiduka [2022] 通过放宽Lipschitz梯度条件进一步分析了AMSGrad的收敛性。然而，他们的分析需要对有界梯度和有界域都施加额外条件。AdaBound [Luo 等人，2019a] 对$\{\mathrm{v}_k\}$ 施加下界和上界，以防止有效步长消失或爆炸。Zhou等人 [2018b] 也采用了对v_k 的新估计来校正偏差。也有一些尝试将Adam与Nesterov 动量 [Dozat， 2016] 以及预热技术 [Liu 等人，2020a]相结合。Padam [Chen等人，2018a] 还引入了部分自适应参数以改善泛化性能。也有一些工作对Adam变体提供了理论分析。例如，Zhou等人 [2018a] 在有界梯度条件下研究了AdaGrad和AMSGrad的收敛性。Gadat和Gavra [2020] 从景观的角度研究了一类自适应梯度方法的渐近行为。他们的分析适用于 $\beta_1 = 0$ 和 β_2 沿迭代点递增的一类Adam变体（也可以理解为 β_2递增的RMSProp）。Alacaoglu等人 [2020]研究了AMSGrad以及另外两种Adam变体。他们的分析同时需要有界梯度和有界域假设，并且不涵盖原始Adam。Iiduka [2022] 通过放宽Lipschitz梯度条件分析了AMSGrad的收敛性。然而，他们的分析需要对有界梯度和有界域都施加额外条件。

另一类理论研究对 Adam 的轨迹施加了额外条件，这会隐式地修改 Adam 的更新规则。De 等人 [2018] 分析了 RMSProp 和 Adam，但他们假设所有随机梯度的符号始终保持不变。这要求每次迭代都执行额外的投影步骤，因此得到的算法已不再是原始 Adam。Barakat 和 Bianchi [2021] 为 Adam 给出了两种证明。第一种证明考虑 (β_1, β_2) 接近 (1, 1) 的情况，并要求满足有界梯度和迭代点有界的假设。正如论文中所述，我们需要“在紧集上进行投影步骤……以确保估计值的有界性”。这一步投影改变了 Adam 的更新规则，更重要的是，它先验地排除了发散的可能性，因此未能充分捕捉原始 Adam 的行为。第二种证明同样要求有界梯度假设，并在特定规则下，用依赖于迭代点的超参数替换常数超参数 β_{1k} 或 β_{2k}。这也改变了 Adam 的更新规则。Huang 等人 [2021] 和 Guo 等人 [2021] 提出了新颖而简洁的框架，用于分析具有较大 β_1的 Adam 系列。然而，他们要求 Adam 的有效步长被限制在特定区间内，即 $\frac{1}{\sqrt{\mathrm{v}_{k}+\epsilon}} \in [C_l, C_u]$。Zaheer 等人 [2018]也对 RMSProp 施加了这一有界性条件。该有界性条件将 Adam 改变为AdaBound [Luo 等人，2019a], ，因此他们无法解释原始 Adam 上的发散-收敛相变。

总之，这些工作研究了 Adam 的修改版本，或隐式地对 Adam 施加额外操作（例如投影），从而改变了 Adam 的更新规则。此外，所有这些工作（包括针对新变体的工作）都要求满足有界梯度假设，这从一开始就排除了可能的发散。相比之下，我们不需要这类假设，并证明改变 (β_1, β_2) 时，存在发散-收敛相变。

在特定设定下理解 Adam。已有一系列工作致力于在特定设定下理解 Adam，例如二次函数（如 [Da Silva和Gazeau，2020，Das等人，2024年，Zhang 等人，2024c]）、强凸函数 [Dereich等人，2025年a], 以及确定性且局部强凸函数 [Dereich等人，2025年c]。Jiang等人 [2023a] 研究了大批量两层线性网络上沿 Adam 的轨迹新定义的条件数的演化。也有理论工作通过常微分方程或随机微分方程方法，在连续时间逼近下分析 Adam 的动力学（例如， [Da Silva和Gazeau，2020，Malladi等人，2022，Dereich和Jentzen，2024年，Kunstner等人，2024年，Cohen等人，2024年，Dereich

等人，2025年b，Li 等人，2025年c]）。与这些工作不同，我们不采用连续逼近，而是在一般非凸函数上分析原始的（因而离散的）Adam。另有出色的工作在确定性（全批次）设定下理解 AdamW 相对于 Adam 的优势 [Xie 和 Li， 2024], 这构成了与本文正交的研究课题。

A.2 关于有界方差条件的更多讨论（2.3）

在此，我们进一步讨论为何经典的“有界方差”条件（2.3）过于受限，以及为何将其放宽为更一般的形式（例如假设 2.2）是有意义的。考虑一个凸二次最小化问题 [Bottou 等人， 2018]:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x)=\frac{1}{2n}\|Ax-b\|_2^2=\frac{1}{2n}\sum_{i=0}^{n-1}(a_i^\top x-b_i)^2:=\frac{1}{n}\sum_{i=0}^{n-1}f_i(x), \tag{A.1}$$

其中 $A \in \mathbb{R}^{d\times d}$ 且 $a^\top{}_i$ 表示第 i 行，该行取自 A ，并且我们假设 A 的秩至少为 2。对于（A.1），我们将证明有界方差条件（2.3）不成立，但假设 2.2 成立。

有界方差条件 （2.3）**在** （A.1）**上不成立**。令 $r(x) := Ax - b \in \mathbb{R}^d$ ，其分量为 $r_i(x) = a^\top{}_i x - b_i$。对于（2.4）中的分解，我们有

$$\nabla f_i(x)=(a_i^\top x-b_i)a_i=r_i(x)a_i, \qquad \nabla f(x)=\frac{1}{n}\sum_{i=0}^{n-1}\nabla f_i(x)=\frac{1}{n}A^\top r(x).$$

代入（2.3）可得

$$\frac{1}{n}\sum_{i=0}^{n-1}\|\nabla f_i(x)-\nabla f(x)\|_2^2=\frac{1}{n}\sum_{i=0}^{n-1}\|r_i(x)a_i\|^2-\frac{1}{n^2}\|A^\top r(x)\|^2. \tag{A.2}$$

假设秩$(A) \geq 2$，那么存在两个索引 $p \neq q$ ，使得 a^p 和 a^q 不共线。取 $u \in \mathbb{R}^d$使得

$$a_p^\top u=1, \qquad a_q^\top u=-1, \tag{A.3}$$

这是可行的，因为 a^p,a^q 线性独立。考虑射线 $x(t) = x_0 + tu$ ，其中任取固定的 x_0。于是

$$r_p(x(t))=r_p(x_0)+t, \qquad r_q(x(t))=r_q(x_0)-t,$$

并且对于每个 $i \in \{p,q\}$，$r_i(x(t)) = r_i(x_0) + t\, a^\top{}_i u$ 至多随 t 线性增长。

利用 (A.2)，并只保留两项 p,q ，得到下界

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n-1}\|\nabla f_i(x(t))-\nabla f(x(t))\|_2^2 &\geq \frac{1}{n}\left(\|r_p(x(t))a_p\|^2+\|r_q(x(t))a_q\|^2-\frac{1}{n}\|A^\top r(x(t))\|^2\right)\\ &=\frac{1}{n}\left((r_p(x_0)+t)^2\|a_p\|^2+(r_q(x_0)-t)^2\|a_q\|^2-\frac{1}{n}\|A^\top r(x(t))\|^2\right).\end{aligned} \tag{A.4}$$

（A.4）中的前两项随 $(\|a_p\|^2+\|a_q\|^2)t^2$ 增长，随着 $|t| \to \infty$。同时，$\|A^\top r(x(t))\|^2 = \|A^\top(Ax(t)-b)\|^2$ 也是关于 t 的二次函数，但 t^2 前的系数为 $\|A^\top Au\|^2$。由于 $a^\top p\, u = 1$ 和 $a^\top q\, u = -1$，因此 Au 至少有两个符号相反的非零坐标，特别是 $\|Au\|^2 \geq 2$。

因此，对于 $x(t)$ ，(A.2) 中的方差表达式是 t 的一个非常数二次多项式，其首项系数严格为正，因此随着 $|t| \to \infty$ 的增大，该表达式无上界。因此，不存在有限常数 D_0 ，使得 (2.3) 对所有 $x \in \mathbb{R}^d$ 都成立。

总之，对于秩$(A) \geq 2$的二次示例（2.4），（2.3）的左侧可以按$\|x\|^2$的量级增长，因此有界方差条件在一般情况下不成立。

假设2.2 在有限常数下成立。接下来我们证明，假设2.2 对（2.4）仍然可以成立。令$W := \mathrm{diag}(\|a_0\|^2, \ldots, \|a_{n-1}\|^2)$。由于 $\nabla f_i(x) = r_i(x) a_i$,

$$\sum_{i=0}^{n-1} \|\nabla f_i(x)\|^2 = \sum_{i=0}^{n-1} r_i(x)^2 \|a_i\|^2 = r(x)^\top W r(x). \tag{A.5}$$

此外,

$$\|\nabla f(x)\|^2 = \frac{1}{n^2}\|A^\top r(x)\|^2 = \frac{1}{n^2} r(x)^\top (AA^\top)\, r(x). \tag{A.6}$$

将 r(x) 分解为 $= r_\parallel(x) + r_\perp$，其中 $r_\parallel(x) \in$ A 的值域，且 $r_\perp \in \mathrm{null}(A^\top)$。由于 $r(x) = Ax - b$，且 $Ax \in$ A 的值域，正交分量与 x 无关:

$$r_\perp = -\Pi_{\mathrm{null}(A^\top)} b, \qquad A^\top r_\perp = 0.$$

利用 $(u+v)^\top W (u+v) \le 2u^\top W u + 2v^\top W v$ 以及（A.5）和（A.6），我们得到

$$\begin{aligned} \sum_{i=0}^{n-1} \|\nabla f_i(x)\|^2 = r^\top W r &\le 2 r_\parallel^\top W r_\parallel + 2 r_\perp^\top W r_\perp \\ &\le 2c\, r_\parallel^\top (AA^\top) r_\parallel + 2 r_\perp^\top W r_\perp = 2cn^2 \|\nabla f(x)\|^2 + 2 r_\perp^\top W r_\perp, \end{aligned} \tag{A.7}$$

其中

$$c := \sup_{u \in \mathrm{range}(A),\, u \neq 0} \frac{u^\top W u}{u^\top (AA^\top) u} < \infty. \tag{A.8}$$

c 有限，这是因为 $AA^\top$ 在A 的值域（A）上正定，因此（A.8）中的广义瑞利商有界。

因此，假设2.2 成立，且

$$D_1 = 2cn^2, \qquad D_0 = 2\, r_\perp^\top W r_\perp = 2 \left(\Pi_{\mathrm{null}(A^\top)} b\right)^\top W \left(\Pi_{\mathrm{null}(A^\top)} b\right), \tag{A.9}$$

两个有限常数，仅取决于 (A,b)。我们还提及一个特殊情形：若 $b \in$ 属于(A)的值域，则$r_\perp = 0$，且 (A.9) 给出 $D_0 = 0$。

总之，对于二次问题 (2.4)，假设 2.2 可以在 (D_1, D_0) 有限时成立，而有界方差条件 (2.3) 不成立。

B 更多符号与有用引理

更多符号。对于随机打乱下的 Adam（算法 2），我们将 $x_{k,i}, m_{k,i}, v_{k,i} \in \mathbb{R}^d$ 记作第 k 个外循环和第 i 个内循环中 x、m、v 的值。此外，我们将 $x_{l,k,i}, m_{l,k,i}, v_{l,k,i} \in \mathbb{R}$ 记作 $x_{k,i}, m_{k,i}, v_{k,i}$ 的第 l 个分量。此外，我们用 $\tau_{k,i}$ 表示第 k 轮次中随机选取的第 i 个批次。基于此，我们将 $\partial_l f_{\tau_{k,i}}(x)$ 记作 $\frac{\partial}{\partial x_l} f_{\tau_{k,i}}(x)$。对于每个轮次 $k \ge 1$，我们用 $\mathbb{E}_k[\cdot]$ 简记给定截至（并包括）轮次起始迭代点 $x_{k,0}$ 的历史、但不包括在轮次 k 中采样的排列 $\{\tau_{k,0}, \ldots, \tau_{k,n-1}\}$ 时的条件期望。换言之，该期望以初始状态（$m_{1,-1}, v_{1,-1}$ 和 $x_{1,0}$）以及所有过去的排列 $\{\tau_{s,0}, \ldots, \tau_{s,n-1}\}$（$s = 1, \ldots, k-1$）为条件。类似地，$\mathbb{E}_{k-1}[\cdot]$表示给定截至（并包括）$x_{k-1,0}$ 的全部历史、但不包括轮次 k 中的排列时的条件期望。我们对 α 的记法作如下约定：对于 $\partial_l f_\alpha(x)$，定义 $\alpha := \arg\max_i |\partial_l f_i(x)|$；对于 $\partial_\alpha f(x)$，定义 $\alpha := \arg\max_l |\partial_l f(x)|$。$m_{\alpha,k,i}$ 和 $v_{\alpha,k,i}$ 也采用类似记法。

我们进一步定义以下常数，它们将在分析采用有放回抽样（算法 1）和随机打乱（算法 2）时的 Adam 中反复使用。

$$\Delta_k := \frac{\eta_0}{\sqrt{k}} \frac{L\sqrt{d}}{\sqrt{1-\beta_2}} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}, \qquad Q_k := \frac{32(n+1)\Delta_1}{\sqrt{k}} \left(\left\lceil \frac{\log(1/2)}{\log \beta_2} \right\rceil + n \right). \tag{B.1}$$

定理 3.1 和 3.3 的一些有用引理。 现在我们证明以下引理 B.2、引理 B.3和引理 B.4，这些引理稍后将被反复使用。

引理 B.1. 对于任意 $\beta \in (0,1)$，我们有

$$(1-\beta)\sum_{j=1}^{\infty}\beta^{j-1}=1,\quad (1-\beta)\sum_{j=1}^{\infty}j\beta^{j-1}=\frac{1}{1-\beta},\quad (1-\beta)\sum_{j=1}^{\infty}j^2\beta^{j-1}=\frac{1+\beta}{(1-\beta)^2}.$$

证明。 证明仅涉及基本计算，我们在此省略证明。 □

引理 B.2. 考虑算法 1，对于任意 $f \in \mathcal{F}_{nL,D_0,D_1}(\mathbb{R}^d)$，$l \in [d], i \in [n]$, 若 $0 \le \beta_1 < \sqrt{\beta_2} < 1$，则我们有

$$|\partial_l f_i(x_{k+1}) - \partial_l f_i(x_k)| \le \frac{\eta_0}{\sqrt{k}}\frac{L\sqrt{d}}{\sqrt{1-\beta_2}}\frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} := \Delta_k. \tag{B.2}$$

同样，对于算法2，在相同条件下，我们有如下结果：

$$|\partial_l f_i(x_{k,i+1}) - \partial_l f_i(x_{k,i})| \le \Delta_k. \tag{B.3}$$

证明。 此处,我们给出（B.2）的证明。（B.3）的证明按照相同的过程完成 e.

$$\begin{aligned}
|x_{l,k+1} - x_{l,k}| &= \eta_k \frac{|m_{l,k}|}{\sqrt{v_{l,k}}} \\
&\le \eta_k(1-\beta_1)\sum_{j=0}^{k}\beta_1^j \frac{|\partial_l f_{\tau_{k-j}}(x_{k-j})|}{\sqrt{v_{l,k}}} \qquad (B.4)\\
&\overset{(i)}{\le} \eta_k(1-\beta_1)\sum_{j=0}^{k}\beta_1^j \frac{|\partial_l f_{\tau_{k-j}}(x_{k-j})|}{|\partial_l f_{\tau_{k-j}}(x_{k-j})|\sqrt{(1-\beta_2)\beta_2^j}} \\
&\overset{(ii)}{=} \eta_k\frac{(1-\beta_1)}{\sqrt{1-\beta_2}}\sum_{j=0}^{k}\left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^j \le \eta_k\frac{(1-\beta_1)}{\sqrt{1-\beta_2}}\sum_{j=0}^{\infty}\left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^j \\
&= \eta_k\frac{(1-\beta_1)}{\sqrt{1-\beta_2}}\frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}. \qquad (B.5)
\end{aligned}$$

注意到，(i) 成立是因为 $v_{l,k} \ge (1-\beta_2)\beta_2^j \partial_l f^{\tau_{k-j}}(x^{k-j})^2$。当 $|\partial_l f^{\tau_{k-j}}(x^{k-j})| > 0$ 时，(ii) 显然成立。当$|\partial_l f^{\tau_{k-j}}(x^{k-j})| = 0$ 时，在非零初始化 $v \succ 0$ 或 $\epsilon > 0$ 下，(B.4) 中的被加数等于 0，这使得 Adam 定义良好。因此，(ii) 中的不等式显然成立。无论哪种情况，我们都得到最终结果 (B.5)，且该最终结果独立于初始化 v 或 ϵ。我们还要指出，这是分析中唯一用到条件 $v \succ 0$ 的地方。

通过应用假设2.1中的 Lipschitz 条件，我们完成了引理 B.2 的证明。 □

引理 B.3。 考虑算法 1，对于任意 $f \in \mathcal{F}_{nL,D_0,D_1}(\mathbb{R}^d)$，$0 \le \beta_1 < \sqrt{\beta_2} < 1, l \in [d], i \in [n]$ *and* $j \in [k-2]$，当 $|\partial f_i(x_k)| \ge {}^{8\sqrt{2}j\sqrt{}}{}^{\Delta_1}{}_{k}$ 时，我们有

$$(\partial_l f_i(x_{k-j}))^2 \ge \frac{(\partial_l f_i(x_k))^2}{2}. \tag{B.6}$$

类似地，对于算法 2，当 $|\partial f_i(x_{k,0})| \ge {}^{8\sqrt{2}jn\sqrt{}}{}^{\Delta_1}{}_{k}$ 时，对于任意 $i' \in [n]$，，我们有

$$(\partial_l f_i(x_{k-j,i'}))^2 \ge \frac{(\partial_l f_i(x_{k,0}))^2}{2}. \tag{B.7}$$

证明。 这里给出（B.6）的证明。（B.7）的证明按照相同的步骤完成。

$$
\begin{aligned}
(\partial_l f_i(x_{k-j}))^2 &= (\partial_l f_i(x_{k-j}) - \partial_l f_i(x_k) + \partial_l f_i(x_k))^2 \\
&\geq (\partial_l f_i(x_k))^2 - 2|\partial_l f_i(x_{k-j}) - \partial_l f_i(x_k)||\partial_l f_i(x_k)| \\
&= (\partial_l f_i(x_k))^2 \left(1 - \frac{2|\partial_l f_i(x_{k-j}) - \partial_l f_i(x_k)|}{|\partial_l f_i(x_k)|}\right).
\end{aligned}
$$

根据引理B.2，我们有

$$
\begin{aligned}
|\partial_l f_i(x_{k-j}) - \partial_l f_i(x_k)| &\overset{\text{Lemma B.2}}{\leq} \sum_{i=1}^{j} \Delta_{k-i}. \\
&\overset{\text{(i)}}{\leq} \frac{2j\Delta_1}{\sqrt{k-1}} \leq \frac{2j\sqrt{2}\Delta_1}{\sqrt{k}},
\end{aligned} \tag{B.8}
$$

其中，(i) 是由 $\sum_{i=1}^{j} \frac{1}{\sqrt{k-i}} \leq \int_{k-1}^{k-j-1} \frac{dt}{\sqrt{t}} = \frac{2j}{\sqrt{k-j-1}+\sqrt{k-1}} \leq \frac{2j}{\sqrt{2k-1}}$ 得出的。因此，我们有

$$
(\partial_l f_i(x_{k-j}))^2 \geq (\partial_l f_i(x_k))^2 \left(1 - \frac{4j\sqrt{2}\Delta_1}{|\partial_l f_i(x_k)|\sqrt{k}}\right).
$$

因此，当 $|\partial_l f_i(x_k)| \geq \frac{8\sqrt{2}j\Delta_1}{\sqrt{k}}$ 时，我们有 $(\partial_l f_i(x^{k-j}))^2 \geq \frac{(\partial_l f_i(x_k))^2}{2}$，由此完成证明。 □

引理 B.4. 考虑算法 *1*，对于任意 $f \in \mathcal{F}_{nL,D_0,D_1}(\mathbb{R}^d)$、$0 \leq \beta_1 < \sqrt{\beta_2} < 1$、$l \in [d], i \in [n]$ 和 $j \in [k-1]$，我们有以下两个结果：

$$
|\partial_l f_{\tau_{k-j}}(x_{k-j})| \leq \sum_{t=1}^{j} \Delta_{k-t} + \sqrt{D_1}\sqrt{nd}\left(|\partial_\alpha f(x_k)| + \sqrt{\frac{D_0}{D_1 d}}\right), \tag{B.9}
$$

其中 $|\partial_\alpha f(x_k)| := \max_{l\in[d]} |\partial_l f(x_k)|$。

同样，对于算法*2*，在相同条件下，我们有以下结果。

$$
|\partial_l f_{\tau_{k,i}}(x_{k,i})| \leq i\Delta_k + \sqrt{D_1}\sqrt{nd}\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right), \tag{B.10}
$$

其中 $|\partial_\alpha f(x_{k,0})| := \text{最大值}_{l\in[d]} |\partial_l f(x_{k,0})|$。

证明。这里给出（B.9）的证明。（B.10）的证明采用相同的过程 e.

$$
\begin{aligned}
|\partial_l f_{\tau_{k-j}}(x_{k-j})| &\overset{\text{Lemma B.2}}{\leq} \sum_{t=1}^{j}\Delta_{k-t} + |\partial_l f_{\tau_{k-j}}(x_k)| \\
&\leq \sum_{t=1}^{j}\Delta_{k-t} + \sum_{l'=1}^{d}\sum_{i=0}^{n-1}|\partial_{l'} f_i(x_k)| \\
&\overset{\text{Cauchy–Schwarz inequality}}{\leq} \sum_{t=1}^{j}\Delta_{k-t} + \sqrt{n}\sum_{l'=1}^{d}\sqrt{\sum_{i=0}^{n-1}|\partial_{l'} f_i(x_k)|^2} \\
&\overset{\text{Cauchy–Schwarz inequality}}{\leq} \sum_{t=1}^{j}\Delta_{k-t} + \sqrt{n}\sqrt{d}\sqrt{\sum_{l'=1}^{d}\sum_{i=0}^{n-1}|\partial_{l'} f_i(x_k)|^2} \\
&\overset{\text{Assumption 2.2}}{\leq} \sum_{t=1}^{j}\Delta_{k-t} + \sqrt{n}\sqrt{d}\sqrt{D_1\|\nabla f(x_k)\|_2^2 + D_0} \\
&\leq \sum_{t=1}^{j}\Delta_{k-t} + \sqrt{D_1}\sqrt{n}d\sqrt{|\partial_\alpha f(x_k)|^2 + \frac{D_0}{D_1 d}} \qquad \text{(B.11)} \\
&\leq \sum_{t=1}^{j}\Delta_{k-t} + \sqrt{D_1}\sqrt{n}d\left(|\partial_\alpha f(x_k)| + \sqrt{\frac{D_0}{D_1 d}}\right). \qquad \text{(B.12)}
\end{aligned}
$$

证明完毕。顺便指出，如果 $D_1 = 0$ （即假设具有有界二阶矩），利用 $\sum_{l'=1}^{d}\sum^{n-1}{}_{i=0}\,|\partial_{l'} f_i(x_k)|^2 \leq D_0$，，则 (B.11) 还可以进一步简化，最终结果为$|\partial_l f_{\tau_{k-j}}(x^{k-j})| \leq \sum^{j}{}_{t=1}\Delta_{k-t} + \sqrt{}ndD_0$。这是 (B.12) 的一种特殊情形，而且由于消除了对 $|\partial_\alpha f(x_k)|$ 的依赖，后续分析会显著简化。简化后的分析中误差项会显著减少，可以沿用当前证明中的相同步骤完成。在我们的分析中，我们不限制 D_1 必须为 0。

□

引理 B.5。考虑算法 2，对于任意满足 $f \in \mathcal{F}nL,D_0,D_1(\mathbb{R}^d)$、$0 \leq \beta_1 < \sqrt{\beta_2} < 1$、$l \in [d], i \in [n]$，且 $k \geq 4$ 足够大以使 $\frac{1-\beta_1^n}{1-\beta_1}\beta(k-1+1 \leq \sqrt{}1)nk$ 的情形，我们有

$$
\sum_{i=0}^{n-1}|m_{l,k,i}| \leq C_{m,1}\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{C_{m,2}}{\sqrt{k}}, \qquad \text{(B.13)}
$$

$$
\sum_{i=0}^{n-1}|m_{l,k-1,i}| \leq \tilde{C}_{m,1}\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{\tilde{C}_{m,2}}{\sqrt{k}}, \qquad \text{(B.14)}
$$

其中

$$
C_{m,1} = 2n, \quad C_{m,2} = \frac{n(n+1)\Delta_1}{2(1-\beta_1)\sqrt{n}} + \beta_1\frac{1-\beta_1^n}{(1-\beta_1)^2}\frac{2\sqrt{2}\Delta_1}{\sqrt{n}} + \sum_{i=1}^{n-1}\|\nabla f_i(x_{1,0})\|_1,
$$

$$
\tilde{C}_{m,1} = n, \quad \tilde{C}_{m,2} = \frac{n}{1-\beta_1}\frac{2\sqrt{2}\Delta_1}{\sqrt{n}} + \frac{\sqrt{2}n(n-1)\Delta_1}{\sqrt{n}} + \sum_{i=1}^{n-1}\|\nabla f_i(x_{1,0})\|_1.
$$

证明。引理B.5 可通过重复应用引理B.4得证。证明很简单，为简洁起见，此处略去证明过程。

□

C 引理6.1 的证明

证明。我们现在证明引理6.1。我们讨论以下两种情况。

情况 1：当 $\partial_l f(x_k) \geq 0$： 当 $\partial_l f(x_k) \leq 0$，我们有

$$\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}} \overset{(6.13)}{\leq} \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}} \left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right)^{\frac{1}{2}}.$$

当 $\partial_l f(x_k) > 0$ 时，我们有

$$\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}} \leq \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}} \frac{1}{\sqrt{\beta_2}}.$$

总之，我们有

$$\begin{aligned}
\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}} &\leq \max\left\{\frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}}\left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right)^{\frac{1}{2}}, \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}}\frac{1}{\sqrt{\beta_2}}\right\} \\
&\leq \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}} + \max\left\{\frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}}\left[\left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right)^{\frac{1}{2}}-1\right], \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}}\left(\frac{1}{\sqrt{\beta_2}}-1\right)\right\} \\
&\leq \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}} + \frac{|\partial_l f(x_k)|}{\sqrt{v_{l,k-1}}} \underbrace{\left(\left[1-\left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right)^{\frac{1}{2}}\right]+\frac{1}{\sqrt{\beta_2}}-1\right)}_{\delta_{5,1}(\beta_2)} \\
&\overset{(6.13)}{\leq} \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}} + \frac{|\partial_l f(x_k)|}{\sqrt{v_{l,k}}} \underbrace{\left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right)^{-\frac{1}{2}}}_{\delta_{5,2}(\beta_2)} \delta_{5,1}(\beta_2) \\
&\overset{(6.2)(6.3)}{\leq} \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}} + \underbrace{\frac{2\sqrt{2n}}{\sqrt{\beta_2^n-(1-\beta_2)16n}}}_{\delta_{5,3}(\beta_2)} \delta_{5,1}(\beta_2)\delta_{5,2}(\beta_2).
\end{aligned}$$

记作 $\delta_5(\beta_2) = \delta_{5,1}(\beta_2)\delta_{5,2}(\beta_2)\delta_{5,3}(\beta_2)$。

$$\begin{aligned}
\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}} - \frac{\partial_l f(x_{k-1})}{\sqrt{v_{l,k-1}}} &\leq \frac{\partial_l f(x_k)-\partial_l f(x_{k-1})}{\sqrt{v_{l,k-1}}} + \delta_5(\beta_2) \\
&\overset{\text{Lemma B.2}}{\leq} \frac{\Delta_{k-1}}{\sqrt{v_{l,k-1}}} + \delta_5(\beta_2).
\end{aligned}$$

情况 2：当 $\partial_l f(x_k) \leq 0$： 当 $\partial_l f(x_k) \geq 0$ 时，我们有

$$\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}} \overset{(6.13)}{\geq} \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}} \left(1-(1-\beta_2)\left(\frac{8n}{\beta_2^n-(1-\beta_2)16n}+\frac{1}{\beta_2}\right)\right)^{\frac{1}{2}}.$$

当 $\partial_l f(x_k) < 0$ 时，我们有

$$\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}} \geq \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}} \frac{1}{\sqrt{\beta_2}}.$$

按照与情况 1 相同的策略，我们可以证明

$$\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}} \geq \frac{\partial_l f(x_k)}{\sqrt{v_{l,k-1}}} - \delta_5(\beta_2),$$

这进一步意味着

$$\begin{aligned}
\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}} - \frac{\partial_l f(x_{k-1})}{\sqrt{v_{l,k-1}}} &\geq \frac{\partial_l f(x_k)-\partial_l f(x_{k-1})}{\sqrt{v_{l,k-1}}} - \delta_5(\beta_2) \\
&\overset{\text{Lemma B.2}}{\geq} -\frac{\Delta_{k-1}}{\sqrt{v_{l,k-1}}} - \delta_5(\beta_2).
\end{aligned}$$

将**情况 1** 和**情况 2**结合起来，我们得到

$$\left|\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}}-\frac{\partial_l f(x_{k-1})}{\sqrt{v_{l,k-1}}}\right|\leq\frac{\Delta_{k-1}}{\sqrt{v_{l,k-1}}}+\delta_5(\beta_2).$$

根据上述不等式，我们得到

$$\begin{aligned}\left|\frac{\partial_l f(x_k)}{\sqrt{v_{l,k}}}-\frac{\partial_l f(x_{k-n})}{\sqrt{v_{l,k-n}}}\right| &\leq \left(\frac{\Delta_{k-1}}{\sqrt{v_{l,k-1}}}+\frac{\Delta_{k-2}}{\sqrt{v_{l,k-2}}}+\cdots+\frac{\Delta_{k-n}}{\sqrt{v_{l,k-n}}}\right)+n\delta_5(\beta_2)\\ &\leq \left(\frac{\Delta_{k-1}}{\sqrt{\beta_2^n v_{l,k-n}}}+\frac{\Delta_{k-2}}{\sqrt{\beta_2^n v_{l,k-n}}}+\cdots+\frac{\Delta_{k-n}}{\sqrt{\beta_2^n v_{l,k-n}}}\right)+n\delta_5(\beta_2)\\ &\leq \frac{n}{\sqrt{\beta_2^n}}\frac{\Delta_{k-n}}{\sqrt{v_{l,k-n}}}+n\delta_5(\beta_2).\end{aligned}$$

引理 6.1的证明至此完成。

□

D 定理3.3的证明

在深入证明定理 3.3之前，我们先给出证明中将用到的一些额外技术引理。回顾一下，根据算法 2，我们定义了 $v_{k,-1}=v_{k-1,n-1}$。

引理 *D.1*。考虑算法 *2*。假设 $k\geq\lceil \log(1/2)/\log\beta_2\rceil+1$ 且 $\partial_l f_\alpha(x_{k,0}):=\max_i|\partial_l f_i(x_{k,0})|\geq\sqrt[8]{2}n\Delta_1 k\log(1/2)/\log\beta_2$。那么对于任意 $f\in\mathcal{F}nL,D_0,D_1(\mathbb{R}^d)$, $0\leq\beta_1<\sqrt{\beta_2}<1$，我们有

$$\frac{\left(\partial_l f_\alpha\left(x_{k,0}\right)\right)^2}{v_{l,k,-1}}\leq\frac{4n}{\beta_2^n}.\tag{D.1}$$

引理 L 的证明 引理 D.1 见附录D.1。现在我们给出$\sqrt{}$的集中不等式 ——.

引理 *D.2*。考虑算法 *2*。假设 $k\geq\lceil \log(1/2)/\log\beta_2\rceil+1$ 且 $\partial_l f_\alpha(x_{k,0}):=\max_i|\partial_l f_i(x_{k,0})|\geq\sqrt[8]{2}n\Delta_1 k\log(1/2)/\log\beta_2$。那么对于任意 $f\in\mathcal{F}nL,D_0,D_1(\mathbb{R}^d)$, $0\leq\beta_1<\sqrt{\beta_2}<1$，以及任意 $i\in[n]$，，我们有

$$\left(1-(1-\beta_2)n\left(\frac{8n}{\beta_2^n}+\frac{1}{2}\right)\right)\frac{1}{\sqrt{v_{l,k,-1}}}\leq\frac{1}{\sqrt{v_{l,k,i}}}\leq\frac{1}{\sqrt{\beta_2^n}}\frac{1}{\sqrt{v_{l,k,-1}}},\tag{D.2}$$

引理 D.2 的证明见附录 D.2。

引理 *D.3*。考虑算法 *2*。假设 $k\geq\lceil \log(1/2)/\log\beta_2\rceil+1$ 且 $\max_i|\partial_l f_i(x_{k,0})|\geq\sqrt[16]{n}\Delta_1 k\log(1/2)/\log\beta_2$。那么对于任意 $f\in\mathcal{F}nL,D_0,D_1(\mathbb{R}^d)$, $0\leq\beta_1<\sqrt{\beta_2}<1$，以及任意 $i=-1,0,\cdots,n-2$，我们有

$$\sqrt{\beta_2^n}\frac{1}{\sqrt{v_{l,k,-1}}}\leq\frac{1}{\sqrt{v_{l,k-1,i}}}\leq\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{-\frac{1}{2}}\frac{1}{\sqrt{v_{l,k,-1}}}\tag{D.3}$$

引理D.3 的证明见附录D.3。

引理 *D.4*。考虑算法 *2*。假设 $k\geq\lceil \log(1/2)/\log\beta_2\rceil+1$。假设 $\partial_l f_\alpha(x_{k,0}):=\max_i|\partial_l f_i(x_{k,0})|\geq\sqrt[16]{n}\Delta_1 k\log(1/2)/\log\beta_2$, $\partial_l f_\alpha(x_{k-1,0}):=\max_i|\partial_l f_i(x_{k-1,0})|\geq\sqrt[16]{n}\Delta_1 k-1\log(1/2)/\log\beta_2$, ..., $\partial_l f_\alpha(x_{k-j,0}):=\max_i|\partial_l f_i(x_{k-j,0})|\geq\sqrt[16]{n}\Delta_1\sqrt{k-j}\log(1/2)/\log\beta_2$，那么对于任意 $f\in\mathcal{F}nL,D_0,D_1(\mathbb{R}^d)$, $0\leq\beta_1<\sqrt{\beta_2}<1$，我们有

$$\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}-\frac{\partial_l f(x_{k-j,0})}{\sqrt{v_{l,k-j,-1}}}\right|\leq\frac{j}{\sqrt{\beta_2^{nj}}}\frac{n\Delta_{(k-j)}}{\sqrt{v_{l,k-j,-1}}}+j\tilde{\delta}_1(\beta_2),\tag{D.4}$$

其中 $\tilde{\delta}_1(\beta_2) = 4n\left(\frac{1}{\sqrt{\beta_2^n}} - 1 + \left[1 - \left(1 - (1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{\frac{1}{2}}\right]\right)\left(1 - (1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{-\frac{1}{2}}$ 是一个常数，当 β_2 趋近于 *1* 时趋近于 *0*。

引理 D.4 的证明可见附录 D.4。定理 3.3 证明的主体随后在附录 D.7中给出。

D.1 引理 D.1D.1

我们先给出$v_{l,k,-1}$的一个下界。

$$\begin{aligned} v_{l,k,-1} &= (1-\beta_2)\left(\left(\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1})\right)^2 + \cdots + \beta_2^{n(k-1)}\left(\partial_l f_{\tau_{1,0}}(x_{1,0})\right)^2\right) \\ &\geq (1-\beta_2)\left(\sum_{j=1}^{k-1}\beta_2^{nj}\sum_{i=0}^{n-1}\left(\partial_l f_{\tau_{k-j,i}}(x_{k-j,i})\right)^2\right) \\ &\geq (1-\beta_2)\left(\sum_{j=1}^{k-1}\beta_2^{nj}\sum_{i=0}^{n-1}\left(\partial_l f_{\tau_{k-j,i}}(x_{k-j,i})\right)^2 \cdot \mathbb{I}(\tau_{k-j,i}=\alpha)\right) \end{aligned}$$

其中 $\alpha = \arg\max_i |\partial_l f_i(x_{k,0})|$。由于 $k \geq \lceil \log(1/2)\, n\log\beta_2 \rceil + 1$，我们进一步得到：

$$\begin{aligned} v_{l,k,-1} &\geq (1-\beta_2)\left(\sum_{j=1}^{\lceil \frac{\log(1/2)}{n\log\beta_2}\rceil}\beta_2^{nj}\sum_{i=0}^{n-1}\left(\partial_l f_{\tau_{k-j,i}}(x_{k-j,i})\right)^2 \cdot \mathbb{I}(\tau_{k-j,i}=\alpha)\right) \\ &\overset{\text{(i)}}{\geq} \frac{(1-\beta_2)}{2}\left(\sum_{j=1}^{\lceil \frac{\log(1/2)}{n\log\beta_2}\rceil}\beta_2^{nj}\sum_{i=0}^{n-1}\left(\partial_l f_{\tau_{k-j,i}}(x_{k,0})\right)^2 \cdot \mathbb{I}(\tau_{k-j,i}=\alpha)\right) \\ &\geq \frac{1-\beta_2}{2}\cdot\beta_2^n\cdot\frac{1-\beta_2^{n\lceil \frac{\log(1/2)}{n\log\beta_2}\rceil}}{1-\beta_2^n}\left(\partial_l f_\alpha(x_{k,0})\right)^2 \\ &\overset{\text{(ii)}}{\geq} \frac{\beta_2^n}{4}\cdot\frac{1-\beta_2}{1-\beta_2^n}\left(\partial_l f_\alpha(x_{k,0})\right)^2 \\ &\overset{\text{(iii)}}{\geq} \frac{\beta_2^n}{4n}\left(\partial_l f_\alpha(x_{k,0})\right)^2 \end{aligned}$$

其中 (i) 是通过应用引理 B.3 得到的，其条件 $\partial_l f_\alpha(x_{k,0}) := \max_i |\partial_l f_i(x_{k,0})| \geq 8\sqrt{2}n\sqrt{\Delta_1 k\left(\log(1/2\beta_2)\right)n\log}$ 得到满足。(ii) 源于 $1-\beta^{n\lceil \log(\frac{1}{2}\frac{2}{\beta_2}\rceil} \geq \frac{1}{2}$ /) nlog；(iii) 源于 $\frac{1-\beta_2}{1-\beta_2^n} \geq 1$ n 。整理不等式后，我们完成了引理 D.1 的证明。

我们进一步说明引理 D.1与有放回抽样版本的集中性结果，即引理 4.1之间的一点区别。在随机打乱的情况下，索引 α 在一个轮次中会且仅会出现一次，而有放回抽样则无法保证这一点。因此，引理 D.1 的结论是确定性的，而引理 4.1 的结论则不是。

□

引理 D.2 的证明

上界为 (D.2) 由于 $v_{l、k、i} \geq \beta_2^n v_{l、k、-1}$，这一结论显而易见。现在我们证明（D.2).

$$\begin{aligned}\frac{1}{\sqrt{v_{l,k,i}}} &\geq \frac{1}{\sqrt{v_{l,k,-1}}}\frac{1}{\sqrt{1+\frac{|v_{l,k,i}-v_{l,k,-1}|}{v_{l,k,-1}}}}\\ &\overset{\text{(i)}}{\geq} \frac{1}{\sqrt{v_{l,k,-1}}}\left(1-\frac{|v_{l,k,i}-v_{l,k,-1}|}{2v_{l,k,-1}}\right),\end{aligned}$$

其中 (i)：$\frac{1}{\sqrt{1+x}} \geq 1 - x/2$。现在我们给出 $|v_{l,k,i} - v_{l,k,-1}|$ 的一个上界。

$$\begin{aligned}|v_{l,k,i}-v_{l,k,-1}| &= \left|(1-\beta_2)\left(\left(\partial_l f_{\tau_{k,i}}(x_{k,i})\right)^2+\beta_2\left(\partial_l f_{\tau_{k,i-1}}(x_{k,i-1})\right)^2+\cdots+\beta_2^i\left(\partial_l f_{\tau_{k,0}}(x_{k,0})\right)^2\right)-(1-\beta_2^{i+1})v_{l,k,-1}\right|\\ &= \left|(1-\beta_2)\left(\left(\left(\partial_l f_{\tau_{k,i}}(x_{k,i})\right)^2-v_{l,k,-1}\right)+\beta_2\left(\left(\partial_l f_{\tau_{k,i-1}}(x_{k,i-1})\right)^2-v_{l,k,-1}\right)+\cdots+\beta_2^i\left(\left(\partial_l f_{\tau_{k,0}}(x_{k,0})\right)^2-v_{l,k,-1}\right)\right)\right|\\ &\leq (1-\beta_2)\left(\left|\left(\partial_l f_{\tau_{k,i}}(x_{k,i})\right)^2-v_{l,k,-1}\right|+\beta_2\left|\left(\partial_l f_{\tau_{k,i-1}}(x_{k,i-1})\right)^2-v_{l,k,-1}\right|+\cdots+\beta_2^i\left|\left(\partial_l f_{\tau_{k,0}}(x_{k,0})\right)^2-v_{l,k,-1}\right|\right)\end{aligned}$$

回顾

$$\begin{aligned}\left|\partial_l f_{\tau_{k,i}}(x_{k,i})\right| &\leq \left|\partial_l f_{\tau_{k,i}}(x_{k,i})-\partial_l f_{\tau_{k,i}}(x_{k,0})\right|+\left|\partial_l f_{\tau_{k,i}}(x_{k,0})\right|\\ &\overset{\text{Lemma B.2}}{\leq} n\Delta_k+\left|\partial_l f_{\tau_{k,i}}(x_{k,0})\right|,\end{aligned}$$

因此我们有

$$\begin{aligned}\left|\partial_l f_{\tau_{k,i}}(x_{k,i})\right|^2 &\leq (n\Delta_k)^2+2n\Delta_k\left|\partial_l f_{\tau_{k,i}}(x_{k,0})\right|+\left|\partial_l f_{\tau_{k,i}}(x_{k,0})\right|^2\\ &\overset{\text{(i)}}{\leq} 4\max_i\left|\partial_l f_i(x_{k,0})\right|^2,\end{aligned}$$

其中，(i) 是由 $\max_i |\partial_l f_i(x_{k,0})| \geq \frac{8\sqrt{2}n\sqrt{\Delta_1 k\log(1/2)}}{n\log\beta_2} > n\Delta_k$ 得出的。因此，我们有

$$\begin{aligned}|v_{l,k,i}-v_{l,k,-1}| &\leq (1-\beta_2)\left(\left|\left(\partial_l f_{\tau_{k,i}}(x_{k,i})\right)^2-v_{l,k,-1}\right|+\beta_2\left|\left(\partial_l f_{\tau_{k,i-1}}(x_{k,i-1})\right)^2-v_{l,k,-1}\right|+\cdots+\beta_2^i\left|\left(\partial_l f_{\tau_{k,1}}(x_{k,1})\right)^2-v_{l,k,-1}\right|\right)\\ &\leq (1-\beta_2)\left(\left|\left(\partial_l f_{\tau_{k,i}}(x_{k,i})\right)^2+v_{l,k,-1}\right|+\beta_2\left|\left(\partial_l f_{\tau_{k,i-1}}(x_{k,i-1})\right)^2+v_{l,k,-1}\right|+\cdots+\beta_2^i\left|\left(\partial_l f_{\tau_{k,1}}(x_{k,1})\right)^2+v_{l,k,-1}\right|\right)\\ &\leq (1-\beta_2)n\left(4\max_i\left|\partial_l f_i(x_{k,0})\right|^2+v_{l,k,-1}\right).\end{aligned}$$

最后，我们有

$$\begin{aligned}\frac{1}{\sqrt{v_{l,k,i}}} &\geq \frac{1}{\sqrt{v_{l,k,-1}}}\left(1-\frac{|v_{l,k,i}-v_{l,k,-1}|}{2v_{l,k,-1}}\right)\\ &\geq \frac{1}{\sqrt{v_{l,k,-1}}}\left(1-\frac{(1-\beta_2)n\left(4\max_i\left|\partial_l f_i(x_{k,0})\right|^2+v_{l,k,-1}\right)}{2v_{l,k,-1}}\right)\\ &\overset{\text{Lemma D.1}}{\geq} \frac{1}{\sqrt{v_{l,k,-1}}}\left(1-(1-\beta_2)n\left(\frac{8n}{\beta_2^n}+\frac{1}{2}\right)\right).\end{aligned}$$

这就完成了引理D.2的证明。 □

D.3 引理D.3的证明

(D.3) 中的下界很容易得出，因为 $v_{l,k-1,i} \leq v_{l,k-1,}/\beta_2^n$。关于 (D.2) 中的上界，我们考虑

$$
\begin{aligned}
v_{l,k-1,i} &= v_{l,k,-1} \cdot \frac{v_{l,k-1,i}}{v_{l,k,-1}} \\
&= v_{l,k,-1} \cdot \frac{v_{l,k,-1} + v_{l,k-1,i} - v_{l,k,-1}}{v_{l,k,-1}} \\
&\geq v_{l,k,-1} \cdot \left(1 - \frac{|v_{l,k-1,i} - v_{l,k,-1}|}{v_{l,k,-1}}\right).
\end{aligned}
$$

证明的其余部分与引理 D.2 的证明步骤相同。首先，可以证明：当 $\max_i |\partial_l f_i(x_{k,0})| \geq \sqrt[16]{n}\sqrt[\Delta_1]{k \log(1/2)\, n \log \beta_2}$ 时，我们有

$$
\left(\partial_l f_{\tau_{k-1,j}}(x_{k-1,j})\right)^2 \leq 4 \max_i \left(\partial_l f_i(x_{k,0})\right)^2. \tag{D.5}
$$

因此，对于$i = -1, 0, \cdots, n-2$，我们得到以下结果

$$
\begin{aligned}
\frac{|v_{l,k,-1} - v_{l,k-1,i}|}{v_{l,k,-1}} &= \left| \frac{(1-\beta_2)}{v_{l,k,-1}} \left[\left(\left(\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1})\right)^2 - v_{l,k-1,i} \right) + \cdots + \beta_2^{n-i-2} \left(\left(\partial_l f_{\tau_{k-1,i+1}}(x_{k-1,i+1})\right)^2 - v_{l,k-1,i} \right) \right] \right| \\
&\leq \frac{(1-\beta_2)}{v_{l,k,-1}} \left[\left(\left(\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1})\right)^2 + v_{l,k-1,i} \right) + \cdots + \beta_2^{n-i-2} \left(\left(\partial_l f_{\tau_{k-1,i+1}}(x_{k-1,i+1})\right)^2 + v_{l,k-1,i} \right) \right] \\
&\overset{(D.5)}{\leq} \frac{(1-\beta_2)}{v_{l,k,-1}} (1 + \cdots + \beta_2^{n-i-2}) \left(4 \max_i |\partial_l f_i(x_{k,0})|^2 + v_{l,k-1,i} \right) \\
&\leq (1-\beta_2) n \left(\frac{16n}{\beta_2^n} + \frac{1}{\beta_2^n} \right),
\end{aligned}
$$

其中，最后一个不等式由$v_{l,k,-1} \geq \beta_2^{n-1-i} v_{l,k-1,i}$对于$i = -1, 0, \cdots, n-2$以及引理D.1 得出（注意，当 $\max_i |\partial_l f_i(x_{k,0})| \geq \sqrt[16]{n}\sqrt[\Delta_1]{k \log(1/2)\, n \log \beta_2}$时，引理D.1 的条件成立）。综上，我们有

$$
\begin{aligned}
v_{l,k-1,i} &\geq v_{l,k,-1} \left(1 - \frac{|v_{l,k-1,i} - v_{l,k,-1}|}{v_{l,k,-1}}\right) \\
&\geq v_{l,k,-1} \left(1 - (1-\beta_2) n \left(\frac{16n+1}{\beta_2^n}\right)\right).
\end{aligned}
$$

整理不等式即可得出引理D.3的证明。 □

D.4 引理D.4的证明

我们先证明$j = 1$时的结果。接下来讨论以下两种情况。

情况 1：当 $\partial_l f(x_{k,0}) \geq \partial_l \sqrt{v_{l,k,-1}}\, f(x_{k-1,0}) \sqrt{v_{l,k-1,-1}}$：当 $\partial_l f(x_{k,0}) \leq 0$时，我们有

$$
\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}} \overset{\text{Lemma D.3}}{\leq} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}} \left(1 - (1-\beta_2) n \left(\frac{16n+1}{\beta_2^n}\right)\right)^{\frac{1}{2}}.
$$

当 $\partial_l f(x_{k,0}) > 0$ 时，我们有

$$
\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}} \overset{\text{Lemma D.3}}{\leq} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}} \frac{1}{\sqrt{\beta_2^n}}.
$$

总之，我们得到

$$
\begin{aligned}
&\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}} \\
\leq\ & \max\left\{\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}}\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{\frac{1}{2}}, \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}}\frac{1}{\sqrt{\beta_2^n}}\right\} \\
\leq\ & \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}} + \max\left\{\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}}\left[\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{\frac{1}{2}}-1\right], \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}}\left(\frac{1}{\sqrt{\beta_2^n}}-1\right)\right\} \\
\leq\ & \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}} + \frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{l,k-1,-1}}}\left(\frac{1}{\sqrt{\beta_2^n}}-1+\left[1-\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{\frac{1}{2}}\right]\right) \\
\overset{\text{Lemma D.3}}{\leq}\ & \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}} + \frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{l,k,-1}}}\left(\frac{1}{\sqrt{\beta_2^n}}-1+\left[1-\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{\frac{1}{2}}\right]\right)\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{-\frac{1}{2}} \\
\overset{\text{Lemma D.1}}{\leq}\ & \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}} + \sqrt{\frac{4n}{\beta_2^n}}\left(\frac{1}{\sqrt{\beta_2^n}}-1+\left[1-\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{\frac{1}{2}}\right]\right)\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{-\frac{1}{2}} \\
:=\ & \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}} + \tilde{\delta}_1(\beta_2),
\end{aligned}
$$

其中 $\tilde{\delta}_1(\beta_2) := \sqrt{\frac{4n}{\beta_2^n}}\left(\frac{1}{\sqrt{\beta_2^n}}-1+\left[1-\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{\frac{1}{2}}\right]\right)\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{-\frac{1}{2}}$ 是一个常数，当 β_2 趋近于 1 时趋近于 0。因此，我们有

$$
\begin{aligned}
\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}} - \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{l,k-1,-1}}} \leq\ & \frac{\partial_l f(x_{k,0}) - \partial_l f(x_{k-1,0})}{\sqrt{v_{l,k-1,-1}}} + \tilde{\delta}_1(\beta_2) \\
\overset{\text{Lemma B.2}}{\leq}\ & \frac{n\Delta_{(k-1)}}{\sqrt{v_{l,k-1,-1}}} + \tilde{\delta}_1(\beta_2).
\end{aligned}
$$

情况 2：当 ~~$\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}} \leq (\partial_l)\sqrt{v_{l,k}}f(x_{k-1,0})\sqrt{v_{l,k-1,-1}}$~~：当 $\partial_l f(x_{k,0}) \geq 0$，我们有

$$
\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}} \overset{(D.3)}{\geq} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}}\left(1-(1-\beta_2)n\left(\frac{16n+1}{\beta_2^n}\right)\right)^{\frac{1}{2}}.
$$

当 $\partial_l f(x_{k,0}) < 0$，我们有

$$
\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}} \geq \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}}\frac{1}{\sqrt{\beta_2^n}}.
$$

按照情况 1 中的相同策略，我们可以证明

$$
\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}} \geq \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k-1,-1}}} - \tilde{\delta}_1(\beta_2),
$$

这进一步意味着

$$
\begin{aligned}
\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}} - \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{l,k-1,-1}}} \geq\ & \frac{\partial_l f(x_{k,0}) - \partial_l f(x_{k-1,0})}{\sqrt{v_{l,k-1,-1}}} - \tilde{\delta}_1(\beta_2) \\
\overset{\text{Lemma B.2}}{\geq}\ & -\frac{n\Delta_{(k-1)}}{\sqrt{v_{l,k-1,-1}}} - \tilde{\delta}_1(\beta_2).
\end{aligned}
$$

结合情况 1 和情况 2，我们得到

$$\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}-\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{l,k-1,-1}}}\right|\le\frac{n\Delta_{(k-1)}}{\sqrt{v_{l,k-1,-1}}}+\tilde{\delta}_1(\beta_2).$$

现在我们考虑 $j>1$ 的情况。根据上述不等式，我们得到

$$\begin{aligned}\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}-\frac{\partial_l f(x_{k-j,0})}{\sqrt{v_{l,k-j,-1}}}\right| &\le \left(\frac{n\Delta_{(k-1)}}{\sqrt{v_{l,k-1,-1}}}+\frac{n\Delta_{(k-2)}}{\sqrt{v_{l,k-2,-1}}}+\cdots+\frac{n\Delta_{(k-j)}}{\sqrt{v_{l,k-j,-1}}}\right)+j\tilde{\delta}_1(\beta_2)\\ &\le \left(\frac{n\Delta_{(k-1)}}{\sqrt{\beta_2^{nj}v_{l,k-j,-1}}}+\frac{n\Delta_{(k-2)}}{\sqrt{\beta_2^{nj}v_{l,k-j,-1}}}+\cdots+\frac{n\Delta_{(k-j)}}{\sqrt{\beta_2^{nj}v_{l,k-j,-1}}}\right)+j\tilde{\delta}_1(\beta_2)\\ &\le \frac{j}{\sqrt{\beta_2^{nj}}}\frac{n\Delta_{(k-j)}}{\sqrt{v_{l,k-j,-1}}}+j\tilde{\delta}_1(\beta_2).\end{aligned}$$

引理D.4的证明到此结束。 □

D.5 引理D.5的证明

现在我们证明引理 D.5，该引理稍后会在附录 D.7 中给出。首先，对于满足 $\max_i|\partial_l f_i(x_k,0)|\ge Q_k$ 的那些 $l\in[d]$，我们给出 $\sum_{i=0}^{n-1}|\partial_l f(x_k,0)|\,m_{l,k,i}/\sqrt{v_{l,k,i}}$，$v_{l,k,i}$ 的一个下界，其中 Q_k 在 (B.1) 中定义。我们讨论以下两种情况。

情况 1： 当 $\partial_l f(x_k,0)m_{l,k,i}\ge 0$ 时，我们有

$$\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}\overset{\text{Lemma D.2}}{\ge}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}\left(1-(1-\beta_2)n(\frac{8n}{\beta_2^n}+\frac{1}{2})\right).$$

情况 2： 当 $\partial_l f(x_k,0)m_{l,k,i}<0$ 时，我们有

$$\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}\overset{\text{Lemma D.2}}{\ge}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}\frac{1}{\sqrt{\beta_2^n}}.$$

令 $i+:=\{i\in\{0,\dots,n-1\}:\partial_l f(x_k,0)m_{l,k,i}\ge 0\}$ 和 $i-:=\{i\in\{0,\dots,n-1\}:\partial_l f(x_k,0)m_{l,k,i}<0\}$。相应地，记作 $\sum_{i+}(\cdot):=\sum_{i\in i+}(\cdot)$ 和 $\sum_{i-}(\cdot):=\sum_{i\in i-}(\cdot)$。结合情况 1 和情况 2，我们得到 l 在 $\max_i|\partial_l f_i(x_k,0)|\ge Q_k$ 时的以下结果：

$$\begin{aligned}\sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} &= \sum_{i+}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}+\sum_{i-}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}\\ &\ge \sum_{i+}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}\left(1-(1-\beta_2)n(\frac{8n}{\beta_2^n}+\frac{1}{2})\right)+\sum_{i-}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}\frac{1}{\sqrt{\beta_2^n}}\\ &= \sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}+\sum_{i+}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}\left(-(1-\beta_2)n(\frac{8n}{\beta_2^n}+\frac{1}{2})\right)\\ &\quad+\sum_{i-}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}\left(\frac{1}{\sqrt{\beta_2^n}}-1\right)\\ &\ge \sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}-\tilde{\delta}_2(\beta_2)\sum_{i=0}^{n-1}\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\right||m_{l,k,i}|\\ &\overset{\text{Lemma D.1 and B.5}}{\ge} \sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}-\tilde{\delta}_2(\beta_2)2\sqrt{\frac{n}{\beta_2^n}}\left(C_{m,1}\left(|\partial_\alpha f(x_{k,0})|+\sqrt{\frac{D_0}{D_1 d}}\right)+\frac{C_{m,2}}{\sqrt{k}}\right),\end{aligned}$$

其中$\tilde{\delta}_2(\beta_2) = (1-\beta_2)n(\frac{8n}{\beta_2^n}+\frac{1}{2}) + \left(\frac{1}{\sqrt{\beta_2^n}}-1\right)$和$C_{m,1}$、$C_{m,2}$在引理B.5中定义。我们有

$$\begin{aligned}
\text{(II-2)} &= \frac{\eta_k}{1-\beta_1^n}\sum_{l=1}^{d}\sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}\mathbb{I}\left(\max_i|\partial_l f_i(x_{k,0})|\geq Q_k\right)\\
&\geq \frac{\eta_k}{1-\beta_1^n}\sum_{l=1}^{d}\sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}\mathbb{I}\left(\max_i|\partial_l f_i(x_{k,0})|\geq Q_k\right)\\
&\quad -\frac{d\eta_k}{1-\beta_1^n}\tilde{\delta}_2(\beta_2)2\sqrt{\frac{n}{\beta_2^n}}\left(C_{m,1}\left(|\partial_\alpha f(x_{k,0})|+\sqrt{\frac{D_0}{D_1 d}}\right)+\frac{C_{m,2}}{\sqrt{k}}\right)\\
&:= \frac{\eta_k}{1-\beta_1^n}\sum_{l=1}^{d}\sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}}\mathbb{I}\left(\max_i|\partial_l f_i(x_{k,0})|\geq Q_k\right)-\eta_k\tilde{\delta}_3(\beta_2)\left(|\partial_\alpha f(x_{k,0})|+\sqrt{\frac{D_0}{D_1 d}}\right)-\frac{\widetilde{C}_4}{k},
\end{aligned}$$

其中

$$\tilde{\delta}_3(\beta_2)=\frac{d}{1-\beta_1^n}\tilde{\delta}_2(\beta_2)2\sqrt{\frac{n}{\beta_2^n}}\cdot C_{m,1},\quad \widetilde{C}_4=\frac{d\eta_0}{1-\beta_1^n}\tilde{\delta}_2(\beta_2)2\sqrt{\frac{n}{\beta_2^n}}\cdot C_{m,2},$$

其中 $C_{m,1}$ 和 $C_{m,2}$ 定义见引理B.5。注意，$\tilde{\delta}_3(\beta_2)$ 是一个常数，当 β_2趋近于 1 时趋近于 0。至此，(II-2) 的证明完成。(II-3) 的证明完全遵循相同的步骤，使用引理 D.3 ，为简洁起见，我们省略其证明。

□

引理D.6 的证明

现在我们证明引理 D.6，该引理稍后将在附录 D.7 中出现。对于满足 $\max_i|\partial_l f_i(x_{k,0})|\geq Q_k$ 的 $l\in[d]$ ，我们有

$$\begin{aligned}
&\mathbb{E}_{k-1}\mathbb{E}_k\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\left(\sum_{i=0}^{n-1}(m_{l,k,i}-\beta_1^n m_{l,k-1,i})\right)\right]\\
=\ &\underbrace{\mathbb{E}_{k-1}\mathbb{E}_k\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\sum_{i=0}^{n-1}(1-\beta_1)\left(\partial_l f_{\tau_{k,i}}(x_{k,i})+\beta_1\partial_l f_{\tau_{k,i-1}}(x_{k,i-1})+\cdots+\beta_1^i\partial_l f_{\tau_{k,0}}(x_{k,0})\right)\right]}_{\text{(III-1)}}\\
&+\underbrace{\mathbb{E}_{k-1}\mathbb{E}_k\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\sum_{i=0}^{n-2}(1-\beta_1)\beta_1^{i+1}\left(\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1})+\cdots+\beta_1^{n-i-2}\partial_l f_{\tau_{k-1,i+1}}(x_{k-1,i+1})\right)\right]}_{\text{(III-2)}}.
\end{aligned}$$

我们首先给出 (III-1) 的下界。

$$\begin{aligned}
\text{(III-1)} &\overset{\text{Lemma B.2}}{\geq} \mathbb{E}_{k-1}\mathbb{E}_k\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\sum_{i=0}^{n-1}(1-\beta_1)\left(\partial_l f_{\tau_{k,i}}(x_{k,0})+\beta_1\partial_l f_{\tau_{k,i-1}}(x_{k,0})+\cdots+\beta_1^i\partial_l f_{\tau_{k,0}}(x_{k,0})\right)\right]\\
&\quad -\mathbb{E}_{k-1}\mathbb{E}_k\left[\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\right|\right]\cdot n^3\Delta_k\\
&\overset{\text{Lemma D.1}}{\geq} \mathbb{E}_{k-1}\mathbb{E}_k\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\sum_{i=0}^{n-1}(1-\beta_1)\left(\partial_l f_{\tau_{k,i}}(x_{k,0})+\beta_1\partial_l f_{\tau_{k,i-1}}(x_{k,0})+\cdots+\beta_1^i\partial_l f_{\tau_{k,0}}(x_{k,0})\right)\right]\\
&\quad -2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\Delta_k
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{=} \mathbb{E}_{k-1}\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\cdot\partial_l f(x_{k,0})\sum_{i=0}^{n-1}(1-\beta_1)(1+\beta_1+\cdots+\beta_1^i)\right]-2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\Delta_k\\
&\geq \mathbb{E}_{k-1}\left[\frac{(\partial_l f(x_{k,0}))^2}{\sqrt{v_{l,k,-1}}}\cdot(1-\beta_1)(1+\beta_1+\cdots+\beta_1^{n-1})\right]-2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\Delta_k\\
&= \mathbb{E}_{k-1}\left[\frac{(\partial_l f(x_{k,0}))^2}{\sqrt{v_{l,k,-1}}}\cdot(1-\beta_1^n)\right]-2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\Delta_k,
\end{aligned}
$$

其中，(i) 是通过对截至 $x_{k,0}$ 的全部历史取条件期望，同时排除第 k 轮中索引的新随机排列而得到的。在 $\mathbb{E}_k[\cdot]$ 下，量 $\partial_l f(x_{k,0})$ 和 $v_{l,k,-1}$ 均视为固定，因此我们有

$$
\mathbb{E}_k\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\,\partial_l f_{\tau_{k,i}}(x_{k,0})\right]=\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\,\mathbb{E}_k[\partial_l f_{\tau_{k,i}}(x_{k,0})].
$$

此外，给定此历史，对于任意固定的 $i\in[n]$，$\tau_{k,i}$ 其在 $[n]$，上均匀分布，因此$\mathbb{E}_k[\partial_l f_{\tau_{k,i}}(x_{k,0})]=\partial_l f(x_{k,0})$。

现在，我们用类似的步骤对 (III-2) 进行界定。

$$
\begin{aligned}
\text{(III-2)} &\overset{\text{Lemma B.2}}{\geq} \mathbb{E}_{k-1}\mathbb{E}_k\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}(1-\beta_1)\sum_{i=0}^{n-2}\beta_1^{i+1}\left(\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,0})+\cdots+\beta_1^{n-i-2}\partial_l f_{\tau_{k-1,i+1}}(x_{k-1,0})\right)\right]\\
&\quad -\mathbb{E}_{k-1}\mathbb{E}_k\left[\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\right|\right]\cdot n^3\Delta_{k-1}\\
&\geq \mathbb{E}_{k-1}\mathbb{E}_k\left[\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{l,k-1,-1}}}(1-\beta_1)\sum_{i=0}^{n-2}\beta_1^{i+1}\left(\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,0})+\cdots+\beta_1^{n-i-2}\partial_l f_{\tau_{k-1,i+1}}(x_{k-1,0})\right)\right]\\
&\quad -\left|\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{l,k-1,-1}}}-\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\right|\left|(1-\beta_1)\sum_{i=0}^{n-2}\beta_1^{i+1}\left(\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,0})+\cdots+\beta_1^{n-i-2}\partial_l f_{\tau_{k-1,i+1}}(x_{k-1,0})\right)\right|\\
&\quad -\mathbb{E}_{k-1}\mathbb{E}_k\left[\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\right|\right]\cdot n^3\Delta_{k-1}\\
&\overset{\text{Lemma D.4}}{\geq} \mathbb{E}_{k-1}\mathbb{E}_k\left[\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{l,k-1,-1}}}(1-\beta_1)\sum_{i=0}^{n-2}\beta_1^{i+1}\left(\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,0})+\cdots+\beta_1^{n-i-2}\partial_l f_{\tau_{k-1,i+1}}(x_{k-1,0})\right)\right]\\
&\quad -\frac{1}{\sqrt{\beta_2^n}}\cdot\frac{\Delta_{k-1}}{\sqrt{v_{l,k-1,-1}}}\mathbb{E}_{k-1}\mathbb{E}_k(1-\beta_1)\sum_{i=0}^{n-2}\beta_1^{i+1}\left(\left|\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,0})\right|+\cdots+\left|\beta_1^{n-i-2}\partial_l f_{\tau_{k-1,i+1}}(x_{k-1,0})\right|\right)\\
&\quad -\tilde{\delta}_1(\beta_2)\mathbb{E}_{k-1}\mathbb{E}_k(1-\beta_1)\sum_{i=0}^{n-2}\beta_1^{i+1}\left(\left|\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,0})\right|+\cdots+\left|\beta_1^{n-i-2}\partial_l f_{\tau_{k-1,i+1}}(x_{k-1,0})\right|\right)\\
&\quad -\mathbb{E}_{k-1}\mathbb{E}_k\left[\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\right|\right]\cdot n^3\Delta_{k-1}\\
&\overset{\text{(i)}}{\geq} \mathbb{E}_{k-1}\mathbb{E}_k\left[\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{l,k-1,-1}}}(1-\beta_1)\sum_{i=0}^{n-2}\beta_1^{i+1}\left(\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,0})+\cdots+\beta_1^{n-i-2}\partial_l f_{\tau_{k-1,i+1}}(x_{k-1,0})\right)\right]\\
&\quad -\frac{1}{\sqrt{\beta_2^n}}\cdot\Delta_{k-1}\cdot(1-\beta_1)\cdot 2\sqrt{\frac{n}{\beta_2^n}}\cdot n^2\\
&\quad -\tilde{\delta}_1(\beta_2)n^2\left(n\Delta_k+\sqrt{D_1}\sqrt{n}d\left(\mathbb{E}_{k-1}\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)\right)\\
&\quad -2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\cdot\Delta_{k-1},
\end{aligned}
$$

$$
\begin{aligned}
\frac{\eta_k}{1-\beta_1^n}\sum_{l=1}^{d}\mathbb{E}_{k-1}\mathbb{E}_k(\text{III-2})\mathbb{I}\left(\max_i|\partial_l f_i(x_{k,0})|\geq Q_k\right) \quad &\geq \quad \frac{\eta_k}{1-\beta_1^n}\sum_{l=1}^{d}\mathbb{E}_{k-1}\mathbb{E}_k(\text{III-2}) \\
&\quad -\left|\frac{\eta_k}{1-\beta_1^n}\sum_{l=1}^{d}\mathbb{E}_{k-1}\mathbb{E}_k(\text{III-2})\mathbb{I}\left(\max_i|\partial_l f_i(x_{k,0})|<Q_k\right)\right| \\
&\overset{\text{(ii)}}{\geq} \quad \left[\frac{(\partial_l f(x_{k-1,0}))^2}{\sqrt{v_{l,k-1,-1}}}(1-\beta_1)\sum_{i=0}^{n-2}\beta_1^{i+1}\left(1+\cdots+\beta_1^{n-i-2}\right)\right] \\
&\quad -d\tilde{\delta}_1(\beta_2)n^2\left(n\Delta_k+\sqrt{D_1}\sqrt{n}d\left(\mathbb{E}_{k-1}\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)\right) \\
&\quad -d\Delta_{k-1}\left(\frac{1}{\sqrt{\beta_2^n}}\cdot(1-\beta_1)\cdot 2\sqrt{\frac{n}{\beta_2^n}}\cdot n^2+2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\right) \\
&\quad -\frac{d\eta_k}{1-\beta_1^n}\frac{(1-\beta_1)^2n^2Q_k}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\geq \quad -d\tilde{\delta}_1(\beta_2)n^2\left(n\Delta_k+\sqrt{D_1}\sqrt{n}d\left(\mathbb{E}_{k-1}\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)\right) \\
&\quad -d\Delta_{k-1}\left(\frac{1}{\sqrt{\beta_2^n}}\cdot(1-\beta_1)\cdot 2\sqrt{\frac{n}{\beta_2^n}}\cdot n^2+2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\right) \\
&\quad -\frac{d\eta_0}{1-\beta_1^n}\frac{(1-\beta_1)^2n^2}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)}\frac{32(n+1)\Delta_1\left(\lceil\frac{\log(1/2)}{\log\beta_2}\rceil+n\right)}{k} \\
&:= \quad -d\tilde{\delta}_1(\beta_2)n^2\left(n\Delta_k+\sqrt{D_1}\sqrt{n}d\left(\mathbb{E}_{k-1}\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)\right) \\
&\quad -d\Delta_{k-1}\left(\frac{1}{\sqrt{\beta_2^n}}\cdot(1-\beta_1)\cdot 2\sqrt{\frac{n}{\beta_2^n}}\cdot n^2+2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\right) \\
&\quad -\frac{\widetilde{C}_8}{k},
\end{aligned}
$$

其中，(i) 由引理 D.1 和 B.4 得出，(ii) 则由随机排列的性质得出：对于任意 $i \in [n]$，$\mathbb{E}_{k-1}(\partial_l f_{\tau_{k-1}}(x_{k-1,0})) = \partial_l f(x_{k-1,0})$。在 (i) 中，我们可以对第 $(k-1)$ 轮使用引理 D.1，因为当 $\max_i|\partial_l f_i(x_{k,0})| \geq Q_k$ 时，$\max_i|\partial_l f_i(x_{k-1,0})|$ 也会自动满足引理 D.1 的条件。C

ombining (III-1) and (III -2) togeth呃，我们 have

$$
\begin{aligned}
\mathbb{E}_{k-1}\mathbb{E}_k(\mathrm{III}) &= \mathbb{E}_{k-1}\mathbb{E}_k\left[\frac{\eta_k}{1-\beta_1^n}\sum_{l=1}^{d}\sum_{i=0}^{n-1}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}}\left(m_{l,k,i}-\beta_1^n m_{l,k-1,i}\right)\mathbb{I}\left(\max_i|\partial_l f_i(x_{k,0})|\geq Q_k\right)\right]\\
&\geq \eta_k\mathbb{E}_{k-1}\left[\sum_{l=1}^{d}\frac{(\partial_l f(x_{k,0}))^2}{\sqrt{v_{l,k,-1}}}\mathbb{I}\left(\max_i|\partial_l f_i(x_{k,0})|\geq Q_k\right)\right]-\frac{d\eta_k}{1-\beta_1^n}2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\Delta_k\\
&\quad-\frac{d\eta_k}{1-\beta_1^n}\left(\tilde{\delta}_1(\beta_2)n^2\left(n\Delta_k+\sqrt{D_1}\sqrt{n}d\left(\mathbb{E}_{k-1}\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)\right)\right)\\
&\quad-\frac{d\eta_k}{1-\beta_1^n}\Delta_{k-1}\left(\frac{1}{\sqrt{\beta_2^n}}\cdot(1-\beta_1)\cdot 2\sqrt{\frac{n}{\beta_2^n}}\cdot n^2+2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\right)-\frac{\widetilde{C}_8}{k}\\
&\geq \eta_k\mathbb{E}_{k-1}\left[\frac{(\partial_\alpha f(x_{k,0}))^2}{\sqrt{v_{\alpha,k,-1}}}\right]-\eta_k\tilde{\delta}_6(\beta_2)\left(\mathbb{E}_{k-1}\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)-\frac{\widetilde{C}_9}{k},
\end{aligned}
$$

其中

$$
\begin{aligned}
\tilde{\delta}_6(\beta_2) &= \frac{1}{1-\beta_1^n}\tilde{\delta}_1(\beta_2)n^2\sqrt{D_1}\sqrt{n}d^2,\\
\widetilde{C}_9 &= \frac{\eta_0 d\Delta_1}{1-\beta_1^n}\left(2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3+n^3\tilde{\delta}_1(\beta_2)+\sqrt{2}\left(\frac{1}{\sqrt{\beta_2^n}}\cdot(1-\beta_1)\cdot 2\sqrt{\frac{n}{\beta_2^n}}\cdot n^2+2\sqrt{\frac{n}{\beta_2^n}}\cdot n^3\right)\right)+\widetilde{C}_8.
\end{aligned}
$$

引理的证明到此结束D.6。 □

D.7 证明主体

现在我们可以证明定理 3.3 了。我们对辅助变量 $z_k = x_{k,0} - \beta_1^n x_{k-1,0}/(1-\beta_1^n)$ 应用下降引理。

$$
f(z_{k+1}) \leq f(z_k)+\langle\nabla f(z_k), z_{k+1}-z_k\rangle+\frac{L}{2}\|z_{k+1}-z_k\|_2^2 \tag{D.6}
$$

我们将给出 $L_2\|z^{k+1}-z_k\|_2^2$ 的上界，以及 $\langle\nabla f(z_k)、z_k-z_{k+1}\rangle$ 的下界。我们从 $L_2\|z^{k+1}-z_k\|_2^2$ 开始。

$$
\begin{aligned}
\|z_{k+1}-z_k\|_2^2 &= \frac{1}{(1-\beta_1^n)^2}\left\|x_{k+1,0}-\beta_1^n x_{k,0}-x_{k,0}+\beta_1^n x_{k-1,0}\right\|_2^2\\
&\leq \frac{1}{(1-\beta_1^n)^2}\left(\left\|x_{k+1,0}-x_{k,0}\right\|_2+\left\|x_{k,0}-x_{k-1,0}\right\|_2\right)^2\\
&\leq \frac{1}{(1-\beta_1^n)^2}\left(\left\|x_{k+1,0}-x_{k,n-1}+x_{k,n-1}-x_{k,n-2}+\cdots+x_{k,1}-x_{k,0}\right\|_2\right.\\
&\quad\left.+\left\|x_{k,0}-x_{k-1,n-1}+x_{k-1,n-1}-\cdots+x_{k-1,1}-x_{k-1,0}\right\|_2\right)^2\\
&\overset{(B.5)}{\leq} \frac{1}{(1-\beta_1^n)^2}\cdot\left(\frac{(1-\beta_1)}{\sqrt{1-\beta_2}}\cdot\frac{d}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)^2\cdot\left(n\eta_k+n\eta_{k-1}\right)^2\\
&\overset{(i)}{\leq} \frac{1}{(1-\beta_1^n)^2}\cdot\left(\frac{(1-\beta_1)}{\sqrt{1-\beta_2}}\cdot\frac{d}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)^2\cdot\frac{(1+\sqrt{2})^2n^2\eta_0^2}{k}\\
&:= \frac{\widetilde{C}_1}{k},
\end{aligned}
\tag{D.7}
$$

其中，(i) 是由 $\sqrt{k-1} \leq \sqrt{k}$ 导致的，当 $k \geq 2$ 时。现在我们给出 $\langle \nabla f(z_k)$、$z_k - z_{k+1} \rangle$ 的下界。

$$\begin{aligned}
\langle \nabla f(z_k), z_k - z_{k+1} \rangle &= \left\langle \nabla f(z_k),\ \frac{x_{k,0} - \beta_1^n x_{k-1,0}}{1-\beta_1^n} - \frac{x_{k+1,0} - \beta_1^n x_{k,0}}{1-\beta_1^n} \right\rangle \\
&= \underbrace{\frac{1}{1-\beta_1^n} \left\langle \nabla f(z_k) - \nabla f(x_{k,0}),\ x_{k,0} - x_{k+1,0} - \beta_1^n (x_{k-1,0} - x_{k,0}) \right\rangle}_{\text{(I)}} \\
&\quad + \underbrace{\frac{1}{1-\beta_1^n} \left\langle \nabla f(x_{k,0}),\ x_{k,0} - x_{k+1,0} - \beta_1^n (x_{k-1,0} - x_{k,0}) \right\rangle}_{\text{(II)}}.
\end{aligned}$$

我们首先证明 (I) 的一个下界：

$$\begin{aligned}
\text{(I)} &\geq -\frac{1}{1-\beta_1^n} \|\nabla f(z_k) - \nabla f(x_{k,0})\|_2 \|x_{k,0} - x_{k+1,0} - \beta_1^n (x_{k-1,0} - x_{k,0})\|_2 \\
&\geq -\frac{L}{1-\beta_1^n} \|z_k - x_{k,0}\|_2 \|x_{k,0} - x_{k+1,0} - \beta_1^n (x_{k-1,0} - x_{k,0})\|_2 \\
&\overset{\text{(i)}}{\geq} -\frac{L}{1-\beta_1^n} \cdot \frac{\beta_1^n}{1-\beta_1^n} \cdot \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \cdot \frac{nd}{1 - \frac{\beta_1}{\sqrt{\beta_2}}} \cdot \eta_k \cdot \|x_{k,0} - x_{k+1,0} - \beta_1^n (x_{k-1,0} - x_{k,0})\|_2 \\
&\geq -\frac{L}{1-\beta_1^n} \cdot \left(\frac{\beta_1^n}{1-\beta_1^n} \cdot \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \cdot \frac{nd}{1 - \frac{\beta_1}{\sqrt{\beta_2}}} \right)^2 \cdot \left(n + \sqrt{2} n \beta_1^n \right) \cdot \left(\frac{\eta_0}{\sqrt{k}} \right)^2 \\
&:= -\frac{\widetilde{C}_2}{k},
\end{aligned} \tag{D.8}$$

其中 (i) 是由于

$$\begin{aligned}
\|z_k - x_{k,0}\|_2 &= \left\| \frac{x_{k,0} - \beta_1^n x_{k-1,0}}{1-\beta_1^n} - x_{k,0} \right\|_2 \\
&= \left\| \frac{\beta_1^n (x_{k,0} - x_{k-1,0})}{1-\beta_1^n} \right\|_2 \\
&\overset{\text{Lemma B.2}}{\leq} \frac{\beta_1^n}{1-\beta_1^n} \cdot \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \cdot \frac{\sqrt{2} nd}{1 - \frac{\beta_1}{\sqrt{\beta_2}}} \cdot \eta_k.
\end{aligned}$$

现在我们给出 (II) 的一个下界，这部分要复杂一些。

$$\begin{aligned}
\text{(II)} &= \frac{1}{1-\beta_1^n} \sum_{l=1}^{d} \partial_l f(x_{k,0}) \left[\eta_k \sum_{i=0}^{n-1} \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} - \beta_1^n \eta_{k-1} \sum_{i=0}^{n-1} \frac{m_{l,k-1,i}}{\sqrt{v_{l,k-1,i}}} \right] \\
&= \underbrace{\frac{1}{1-\beta_1^n} \sum_{l=1}^{d} \partial_l f(x_{k,0}) \left[\eta_k \sum_{i=0}^{n-1} \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} - \beta_1^n \eta_{k-1} \sum_{i=0}^{n-1} \frac{m_{l,k-1,i}}{\sqrt{v_{l,k-1,i}}} \right] \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \leq Q_k \right)}_{\text{(II-1)}} \\
&\quad + \underbrace{\frac{\eta_k}{1-\beta_1^n} \sum_{l=1}^{d} \left[\sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \right] \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \right)}_{\text{(II-2)}} \\
&\quad \underbrace{- \frac{\beta_1^n \eta_{k-1}}{1-\beta_1^n} \sum_{l=1}^{d} \left[\sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k-1,i}}{\sqrt{v_{l,k-1,i}}} \right] \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \right)}_{\text{(II-3)}}.
\end{aligned}$$

我们从 (II-1) 开始。注意，(II-1) 施加了有界梯度条件，因此可以简单地将其界定如下：

$$\begin{aligned}
\text{(II-1)} &\overset{\text{(B.5)}}{\geq} -\frac{\sum_{l=1}^{d} n \max_i |\partial_l f_i(x_{k,0})|}{1-\beta_1^n} \frac{\eta_0}{\sqrt{k}} \left(\frac{1-\beta_1}{\sqrt{1-\beta_2}} \cdot \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} + \sqrt{2} \cdot \frac{1-\beta_1}{\sqrt{1-\beta_2}} \cdot \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} \right) \\
&\geq -\frac{dQ_k}{1-\beta_1^n} \frac{2\sqrt{2}n\eta_0}{\sqrt{k}} \left(\frac{1-\beta_1}{\sqrt{1-\beta_2}} \cdot \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} \right) \\
&:= -\frac{\widetilde{C}_3}{k}. \qquad \text{(D.9)}
\end{aligned}$$

现在处理 (II-2) 和 (II-3)，它们关注梯度无界的情况。我们在以下引理中对 (II-2) 和 (II-3) 进行界定D.5。

引理 D.5。 考虑算法 2。假设 $k \geq \lceil \log(1/2)\, n \log \beta_2 \rceil + 1$。那么对于任意 $f \in \mathcal{F}_{nL,D_0,D_1}(\mathbb{R}^d)$，我们有

$$\begin{aligned}
\textit{(II-2)} &= \frac{\eta_k}{1-\beta_1^n} \sum_{l=1}^{d} \left[\sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \right] \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \right) \\
&\geq \frac{\eta_k}{1-\beta_1^n} \sum_{l=1}^{d} \sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k,i}}{\sqrt{v_{l,k,-1}}} \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \right) \\
&\quad -\eta_k \tilde{\delta}_3(\beta_2) \left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}} \right) - \frac{\widetilde{C}_4}{k}, \qquad \text{(D.10)}
\end{aligned}$$

其中

$$\tilde{\delta}_2(\beta_2) = (1-\beta_2)\, n\left(\frac{8n}{\beta_2^n} + \frac{1}{2}\right) + \left(\frac{1}{\sqrt{\beta_2^n}} - 1 \right), \quad \tilde{\delta}_3(\beta_2) = \frac{d}{1-\beta_1^n} \tilde{\delta}_2(\beta_2) 2\sqrt{\frac{n}{\beta_2^n}} \cdot C_{m,1}, \quad \widetilde{C}_4 = \frac{d\eta_0}{1-\beta_1^n} \tilde{\delta}_2(\beta_2) 2\sqrt{\frac{n}{\beta_2^n}} \cdot C_{m,2},$$

并且 $C_{m,1}$ 和 $C_{m,2}$ 是引理 B.5 中定义的常数。注意，$\tilde{\delta}_2(\beta_2)$ 是一个当 β_2趋近于 1 时趋近于 0 的常数。类似地，我们有

$$\begin{aligned}
\textit{(II-3)} &= -\frac{\beta_1^n \eta_{k-1}}{1-\beta_1^n} \sum_{l=1}^{d} \left[\sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k-1,i}}{\sqrt{v_{l,k-1,i}}} \right] \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \right) \\
&\geq -\frac{\beta_1^n \eta_{k-1}}{1-\beta_1^n} \sum_{l=1}^{d} \sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k-1,i}}{\sqrt{v_{l,k,-1}}} \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \right) \\
&\quad -\eta_{k-1} \tilde{\delta}_4(\beta_2) \left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}} \right) - \frac{\widetilde{C}_5}{k}, \qquad \text{(D.11)}
\end{aligned}$$

其中

$$\tilde{\delta}_4(\beta_2) = \frac{d\beta_1^n}{1-\beta_1^n} \tilde{\delta}_2(\beta_2) 2\sqrt{\frac{n}{\beta_2^n}} \cdot \tilde{C}_{m,1}, \quad \widetilde{C}_5 = \frac{d\beta_1^n \eta_0}{1-\beta_1^n} \tilde{\delta}_2(\beta_2) 2\sqrt{\frac{n}{\beta_2^n}} \cdot \tilde{C}_{m,2},$$

并且 $\tilde{C}_{m,1}$，以及 $\tilde{C}_{m,2}$ 是引理 B.5中定义的常数。

引理 D.5 的证明见附录 D.5。将 (II-1)、(II-2) 和 (II-3) 合并起来，我们得到 (II) 的以下下界：

$$
\begin{aligned}
\text{(II)} \quad &= \quad \text{(II-1)} + \text{(II-2)} + \text{(II-3)} \\
&\overset{\text{(D.9)}}{\geq} \quad -\frac{\widetilde{C}_3}{k} + \text{(II-2)} + \text{(II-3)} \\
&\overset{\text{Lemma D.5}}{\geq} \quad \underbrace{\frac{\eta_k}{1-\beta_1^n} \sum_{l=1}^{d} \sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k,i} - \beta_1^n m_{l,k-1,i}}{\sqrt{v_{l,k,-1}}} \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \right)}_{\text{(III)}} \\
&\qquad -\frac{1}{1-\beta_1^n} (\eta_{k-1} - \eta_k) \beta_1^n \sum_{l=1}^{d} \sum_{i=0}^{n-1} \left| \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,-1}}} \right| |m_{l,k-1,i}| \\
&\qquad - \left(\eta_k \tilde{\delta}_3(\beta_2) + \eta_{k-1} \tilde{\delta}_4(\beta_2) \right) \left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}} \right) - \frac{\widetilde{C}_3 + \widetilde{C}_4 + \widetilde{C}_5}{k} \\
&\overset{\text{(i)}}{\geq} \quad \underbrace{\frac{\eta_k}{1-\beta_1^n} \sum_{l=1}^{d} \sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k,i} - \beta_1^n m_{l,k-1,i}}{\sqrt{v_{l,k,-1}}} \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \right)}_{\text{(III)}} \\
&\qquad -\frac{d\beta_1^{\frac{n}{2}} 4n\sqrt{n}\eta_0}{(1-\beta_1^n)k} \left(C_{m,1} \left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}} \right) + \frac{C_{m,2}}{\sqrt{k}} \right) \\
&\qquad - \left(\eta_k \tilde{\delta}_3(\beta_2) + \eta_{k-1} \tilde{\delta}_4(\beta_2) \right) \left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}} \right) - \frac{\widetilde{C}_3 + \widetilde{C}_4 + \widetilde{C}_5}{k} \\
&\geq \quad \underbrace{\frac{\eta_k}{1-\beta_1^n} \sum_{l=1}^{d} \sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k,i} - \beta_1^n m_{l,k-1,i}}{\sqrt{v_{l,k,-1}}} \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \right)}_{\text{(III)}} \\
&\qquad -\eta_k \tilde{\delta}_5(\beta_2) \left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}} \right) - \frac{\widetilde{C}_6}{k} \left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}} \right) - \frac{\widetilde{C}_7}{k}, \qquad \text{(D.12)}
\end{aligned}
$$

其中

$$
\tilde{\delta}_5(\beta_2) = \tilde{\delta}_3(\beta_2) + \sqrt{2}\tilde{\delta}_4(\beta_2), \quad \widetilde{C}_6 := \frac{d\beta_1^{\frac{n}{2}} 4n\sqrt{n}\eta_0 C_{m,1}}{(1-\beta_1^n)}, \quad \widetilde{C}_7 = \widetilde{C}_3 + \widetilde{C}_4 + \widetilde{C}_5 + \frac{d\beta_1^{\frac{n}{2}} 4n\sqrt{n}\eta_0 C_{m,2}}{(1-\beta_1^n)}.
$$

(i) 由引理 D.1 和引理 B.5 以及 $\sqrt{}^{1}_{k-1} - \sqrt{}^{1}_{k} \leq {}^{2}_{k}$（其中 $k \geq 2$）得到。现在我们取条件期望 $\mathbb{E}_k[\cdot]$，然后取 $\mathbb{E}_{k-1}[\cdot]$，并在以下引理中对 (III) 进行界定。

引理 D.6. 考虑算法 2。假设 $k \geq \lceil \log(1/2)\, n \log \beta_2 \rceil + 1$。那么，对于任意 $f \in \mathcal{F}_{nL,D_0,D_1}(\mathbb{R}^d)$，$0 \leq \beta_1 < \sqrt{\beta_2} < 1$，均有

$$
\mathbb{E}_{k-1}\mathbb{E}_k(III) \quad \geq \quad \eta_k \mathbb{E}_{k-1}\left[\frac{(\partial_\alpha f(x_{k,0}))^2}{\sqrt{v_{\alpha,k,-1}}} \right] - \eta_k \tilde{\delta}_6(\beta_2) \left(\mathbb{E}_{k-1} |\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}} \right) - \frac{\widetilde{C}_9}{k}, \qquad \text{(D.13)}
$$

其中

$$
\begin{aligned}
\tilde{\delta}_6(\beta_2) &= \frac{1}{1-\beta_1^n} \tilde{\delta}_1(\beta_2) n^2 \sqrt{D_1}\sqrt{n} d^2 \\
\widetilde{C}_8 &= \frac{d\eta_0}{1-\beta_1^n} \cdot 32(n+1)\Delta_1 \left(\lceil \frac{\log(1/2)}{\log \beta_2} \rceil + n \right) \cdot \frac{(1-\beta_1)^2 n^2}{\sqrt{1-\beta_2}\left(1 - \frac{\beta_1}{\sqrt{\beta_2}}\right)} \\
\widetilde{C}_9 &= \frac{\eta_0 d \Delta_1}{1-\beta_1^n} \left(2\sqrt{\frac{n}{\beta_2^n}} \cdot n^3 + n^3 \tilde{\delta}_1(\beta_2) + \sqrt{2} \left(\frac{1}{\sqrt{\beta_2^n}} \cdot (1-\beta_1) \cdot 2\sqrt{\frac{n}{\beta_2^n}} \cdot n^2 + 2\sqrt{\frac{n}{\beta_2^n}} \cdot n^3 \right) \right) + \widetilde{C}_8.
\end{aligned}
$$

引理的证明 mma D.6 如附录所示D.6。综合 (I) 至 (III) 的所有结果，我们得到

$$\begin{aligned}\mathbb{E}_{k-1}\mathbb{E}_k\langle\nabla f(z_k), z_k - z_{k+1}\rangle &\overset{(D.8),(D.12),(D.13)}{\geq} \underbrace{-\frac{\tilde{C}_2}{k}}_{\text{from (I)}} \underbrace{-\eta_k\tilde{\delta}_5(\beta_2)\left(\mathbb{E}_{k-1}|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) - \frac{\tilde{C}_6}{k}\left(\mathbb{E}_{k-1}|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) - \frac{\tilde{C}_7}{k}}_{\text{from (II)}} \\ &\quad \underbrace{+\eta_k\mathbb{E}_{k-1}\left[\frac{(\partial_\alpha f(x_{k,0}))^2}{\sqrt{v_{\alpha,k,-1}}}\right] - \eta_k\tilde{\delta}_6(\beta_2)\left(\mathbb{E}_{k-1}|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) - \frac{\tilde{C}_9}{k}}_{\text{from (III)}} \\ &= \eta_k\mathbb{E}_{k-1}\left[\frac{(\partial_\alpha f(x_{k,0}))^2}{\sqrt{v_{\alpha,k,-1}}}\right] - \eta_k\left(\tilde{\delta}_5(\beta_2) + \tilde{\delta}_6(\beta_2)\right)\left(\mathbb{E}_{k-1}|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) \\ &\quad - \frac{\tilde{C}_6}{k}\left(\mathbb{E}_{k-1}|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) - \frac{\tilde{C}_2 + \tilde{C}_7 + \tilde{C}_9}{k}. \end{aligned} \tag{D.14}$$

对（D.14）两边取期望，并重复推导（6.44）时所用的步骤，我们得到以下不等式：当 $\tilde{\delta}_5(\beta_2) + \tilde{\delta}_6(\beta_2) \leq \frac{1}{4\sqrt{5}dD_1 n}$ 时，可通过令 $1-\beta_2 = \mathcal{O}((1-\beta_1^n)/n^{5.5})$ 达到；~~并且~~ k 足够大，使得 $\frac{\tilde{C}_6}{\eta_0\sqrt{k}} \leq \frac{1}{4\sqrt{5}dD_1 n}$，

$$\mathbb{E}\left[\frac{(\partial_\alpha f(x_{k,0}))^2}{n\sqrt{v_{\alpha,k,-1}}} - (\tilde{\delta}_5(\beta_2) + \tilde{\delta}_6(\beta_2) + \frac{\tilde{C}_6}{\eta_0\sqrt{k}})|\partial_\alpha f(x_{k,0})|\right] \geq \mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_{k,0})\|_2^2}{d\sqrt{5D_0 nd}}, \frac{\|\nabla f(x_{k,0})\|_2}{2d^2\sqrt{5D_1 n}}\right\}\right]$$

$$- (\tilde{\delta}_5(\beta_2) + \tilde{\delta}_6(\beta_2))\sqrt{\frac{D_0}{D_1 d}} - \frac{\tilde{C}_{10}}{\sqrt{k}},$$

其中 $\tilde{C}_{10} = \tilde{C}_6\sqrt{D_0 D_1 d} + \frac{32\Delta_1^2}{(1-\beta_2)^2 D_1 d^2\sqrt{5D_0 nd}} + \frac{2\sqrt{2}\Delta_1}{(1-\beta_2)D_1 nd^2\sqrt{5}} + (\tilde{\delta}_5(\beta_2) + \tilde{\delta}_6(\beta_2))\frac{\tilde{C}_6}{\eta_0}\sqrt{D_0 D_1 d} + \frac{\tilde{C}_6}{\eta_0}4\sqrt{2n\Delta_1(1-\beta_2)d\sqrt{D_1}}$。

$$\mathbb{E}\langle\nabla f(z_k), z_k - z_{k+1}\rangle \geq \eta_k\left\{\mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_{k,0})\|_2^2}{d\sqrt{5D_0 nd}}, \frac{\|\nabla f(x_{k,0})\|_2}{2d^2\sqrt{5D_1 n}}\right\}\right] - \tilde{\delta}(\beta_2)\sqrt{D_0} - \frac{\tilde{C}}{\sqrt{k}}\right\},$$

其中 $\tilde{\delta}(\beta_2) = (\tilde{\delta}_5(\beta_2) + \tilde{\delta}_6(\beta_2))\sqrt{\frac{1}{D_1 d}}$，$\tilde{C} = \tilde{C}_{10} + \frac{\tilde{C}_6\sqrt{D_0 D_1 d} + \tilde{C}_2 + \tilde{C}_7 + \tilde{C}_{\eta_0}}{9}$。

按照定理 3.1 证明中的相同步骤，我们得到

$$\min_{k\in[1,T]}\mathbb{E}\left[\min\left\{\frac{\|\nabla f(x_{k,0})\|_2^2}{\sqrt{D_0}}, \frac{\|\nabla f(x_{k,0})\|_2}{2\sqrt{dD_1}}\right\}\right] \leq \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right) + \mathcal{O}(\tilde{\delta}(\beta_2)\sqrt{D_0}).$$

最后，当 $D_1 = 0$，即有界二阶矩条件成立时，经过对常数项作一些调整，可以得到类似的结论。在 $D_1 = 0$ 下的证明严格来说比当前证明简单，因为它可归结为有界梯度情形。至此，我们完成了定理 3.3 的证明。 □

E 关于偏差校正项和非零 ϵ 的讨论

在上述分析中，我们关注的是不含偏差校正项的 Adam，并考虑 $\epsilon = 0$ (ϵ 是算法中用于数值稳定性的超参数1 和 2)。为完整起见，下面简要讨论如何将偏差校正项和非零 ϵ 纳入上述分析。基于当前的收敛证明，我们只需进行几处简单的补充。我们先讨论非零 ϵ，然后讨论偏差校正项。我们将采用算法 1 的符号，但所有论证同样适用于算法 2。

Adam 在非零 ϵ 下：在我们当前的分析中，我们考虑 $\epsilon = 0$。实践中，ϵ 通常被设置为一个很小的正数，例如 10^{-8}。证明在 $\epsilon > 0$ 下的收敛严格来说更为简单。它只需在当前证明的基础上做几处简单修改。我们如下说明。

当 $\epsilon \neq 0$ 时，新的二阶动量变为 $\sqrt{\hat{v}_{l,k}} := \sqrt{v_{l,k}} + \epsilon$。经过这一改动，新证明可通过对第6节中当前证明做以下细微改动来完成。首先，我们对当前证明做如下轻微修改：

- 首先，对于引理 B.3，我们将条件从 $|\partial_l f_i(x_k)| \geq \frac{\sqrt{1-\beta_2}}{2\Delta_1 jk}$ 到 $|\partial_l f_i(x_k)| \geq \max\left(1, \frac{8\sqrt{1-\beta_2}}{\Delta_1\sqrt{k}}\right)$。引理 B.3 的结论不受影响，保持不变。
- 其次，基于新版引理 B.3，我们将引理 4.1 的条件从 $\max_i |\partial_l f_i(x_k)| \geq R_k$ 改为 $\max_i |\partial_l f_i(x_k)| \geq \max(1, R_k)$。引理 4.1 的结论不受影响，保持不变。

- 我们定义常数 $\hat{Q}_k = \max(1, Q_k)$。

现在，我们重新陈述第6节证明中的两种情况：

- **情况 1（有界梯度）：** 当 $\max_i |\partial_l f_i(x_k)| \leq \hat{Q}_k$ 时，我们考虑两个子情况：若 $\hat{Q}_k = Q_k$，则从 $\sqrt{v_{l,k}}$ 改为 $\sqrt{\hat{v}_{l,k}}$ 不会影响证明，因为这不会改变 (B.5) 的结果。特别是，“ $\mathbb{E}(c)$”段落中的推导仍然成立。若 $\hat{Q}_k = 1$，则梯度由常数 1 所界定。这等价于在有界梯度条件下以常数 1 分析 Adam。特别地，我们有：

$$\sqrt{v_{l,k}} \leq \sqrt{\hat{v}_{l,k}} = \sqrt{v_{l,k}} + \epsilon \leq C,$$

其中 $C > 0$ 是一个与 ϵ 无关的常数。因此，每当我们需要 $\frac{1}{\sqrt{\hat{v}_{l,k}}}$ 的上界时，就使用 $\frac{1}{\sqrt{\hat{v}_{l,k}}} \leq \frac{1}{\sqrt{v_{l,k}}}$，然后按照当前证明中的相同步骤进行。每当我们需要 $\frac{1}{\sqrt{\hat{v}_{l,k}}}$ 的下界时，就使用 $\frac{1}{\sqrt{\hat{v}_{l,k}}} \geq \frac{1}{C}$。这样一来，证明比当前证明更简单，因为它将 Adam 化归为随机梯度下降。为简洁起见，我们省略证明。

- **情况 2（无界梯度）：** 情况 1 的补集，即无界偏梯度事件 $B_{cl,k}$。在这种情况下，以高概率，$\sqrt{v_{l,k}}$ 的下界是一个与 $\hat{Q}_k$ 同阶的大常数。参照引理4.1 的证明（从（5.11）到（5.19））即可得出这一点。由于 ϵ 通常选取得远小于 1，因此有 $\sqrt{v_{l,k}} \approx \hat{Q}_k > \epsilon$。因此，以高概率有以下关系：

$$\sqrt{v_{l,k}} \leq \sqrt{\hat{v}_{l,k}} = \sqrt{v_{l,k}} + \epsilon \leq 2\sqrt{v_{l,k}}.$$

因此，每当我们需要 $\frac{1}{\sqrt{\hat{v}_{l,k}}}$ 的上界时，就使用 $\frac{1}{\sqrt{\hat{v}_{l,k}}} \leq \frac{1}{\sqrt{v_{l,k}}}$，然后按照当前证明中的相同步骤进行。每当我们需要 $\frac{1}{\sqrt{\hat{v}_{l,k}}}$ 的下界时，就使用 $\frac{1}{\sqrt{\hat{v}_{l,k}}} \geq \frac{1}{2\sqrt{v_{l,k}}}$，然后对常数稍作调整，按照当前证明中的相同步骤进行。最后，尾概率可以使用当前段落“**下界** $\mathbb{E}(e)$”中的相同步骤加以控制，并且尾界随 $\beta_2 \to 1$ 指数级地趋于零。

最终的收敛结果将独立于 ϵ。将符号改为 2 算法中的 $v_{k,0}$、$\hat{v}_{k,0}$、$\nabla f_{\tau_{k,0}}(x)$，上述论证同样适用。

带偏差校正项的 Adam： 这些偏差校正项由 [Kingma 和 Ba, 2014]提出。其形式如下：

$$x_{k+1} = x_k - \eta_k \frac{m_k/(1-\beta_1^k)}{\sqrt{v_k/(1-\beta_2^k)} + \epsilon} = x_k - \eta_k \frac{\sqrt{1-\beta_2^k}}{1-\beta_1^k} \frac{m_k}{\sqrt{v_k} + \epsilon\sqrt{(1-\beta_2^k)}} := x_k - \hat{\eta}_k \frac{m_k}{\sqrt{v_k} + \hat{\epsilon}}$$

如上所示，偏差校正项可以通过将步长 η_k 改为 $\hat{\eta}_k = \frac{\sqrt{1-\beta_2^k}}{1-\beta_1^k}\eta_k = \frac{\sqrt{1-\beta_2^k}}{1-\beta_1^k}\frac{\eta_0}{\sqrt{k}}$，并将 ϵ 改为 $\hat{\epsilon} = \epsilon\sqrt{(1-\beta_2^k)}$ 来实现。我们现在说明如何将这一改动纳入我们的分析。

我们观察到，新的步长 $\hat{}\eta_k$ 在旧步长 η_k 附近有良好的界限，即 $\hat{\eta}_k \in [\sqrt{1-\beta_2}\eta_k, \frac{1}{1-\beta_1}\eta_k]$。因此，为了证明 Adam 在 $\hat{}\eta_k$ 下的收敛，我们在当前证明中加入以下步骤。

- 由于 $\hat{}\epsilon \le \epsilon$，先前关于 ϵ 的分析可以直接应用于 $\hat{}\epsilon$。
- 每当我们需要 $\hat{}\eta_k$ 的上界时，我们使用 $\hat{}\eta_k \le \frac{1}{1-\beta_1}\eta_k$。然后，我们按照原有分析进行，只需额外引入常数 $\frac{1}{1-\beta_1}$。结果表明，我们只需将 (B.1) 中的常数 Δ_k 替换为 $\frac{1}{1-\beta_1}\frac{\eta_0}{\sqrt{k}}L\sqrt{1-\beta_2}\frac{1-\beta_1}{1-\sqrt{\beta_2}}d$。其余分析保持不变。
- 每当我们需要得到 $\hat{}\eta_k$ 的下界时，就使用 $\hat{}\eta_k \ge \sqrt{1-\beta_2}\eta_k$。然后，我们在原有分析的基础上增加一个常数 $\sqrt{1-\beta_2}$。因此，我们只需修改最终结果中的常数项。其余分析保持不变。

F 实验设置

这里介绍我们的实验设置。

- **图 3 (b) 中使用的不可实现函数实验**。这里说明图 3 (b) 中使用的不可实现函数。这个例子改述自 [Shi 等人，2020，附录 A。4]。

$$f_j(x) = \begin{cases} (x-a)^2 \text{ if } j = 0 \\ -0.1\left(x - \frac{10}{9}a\right)^2 \text{ if } 1 \le j \le 9 \end{cases} \tag{F.1}$$

我们可以看到，$f(x) = \frac{1}{10}\sum_{j=0}^{9} f_j(x) = \frac{1}{10}\left(\frac{1}{10}x^2 - \frac{1}{9}a^2\right)$ 是一个凸函数，并且 $f(x)$ 的下界为 $-\frac{1}{90}a^2$。我们在图 3 中使用了 a = 10 (b)。注意，对于函数，我们有 $D_0 > 0$ （F.1）。

- 反例 (3.1) 上的实验。我们使用算法 2，按循环顺序 f0、f1、f2 等最小化函数 (3.1)。我们报告经过 50k 次迭代后的最优性差距 $x - x^*$，这等价于 50000/n 轮次。我们使用 $\epsilon = 10^{-8}$ 以保证数值稳定性。我们采用递减步长 $\eta k = 0.1/\sqrt{k}$，其中 k 是轮次的索引。除非另有说明，此设置适用于所有其他关于函数 (3.1) 的实验。
- **MNIST [Deng，2012]**。我们使用宽度为 =16 的单隐藏层神经网络。我们设置批量大小为 =1，权重衰减为 =0，步长为 =0.0001，并训练 20 轮。我们使用 $\epsilon = 10^{-8}$ 以保证数值稳定性。
- **CIFAR-10 [Krizhevsky 等人，2009]**。我们使用 ResNet-18 [He 等人，2016]作为架构。我们选择批量大小为 =16，权重衰减为 =5e-4，初始步长为=1e-3。我们采用分阶段恒定学习率调度，在第 30、60 和 90 轮时将学习率乘以 0.1。我们使用 $\epsilon = 10^{-8}$ 以保证数值稳定性。

对于 MNIST 和 CIFAR-10，较大的批量大小会呈现出与图1中类似的模式，但相变会在某个更小的 β_2。