arXiv:2602.01212v1 [cs.LG] 1 Feb 2026

SimpleGPT：通过一种简单的归一化策略改进 GPT

Marco Chen [1] * **Xianbiao Qi** [2] *† **Yelin He** [2] **Jiaquan Ye** [2] **Rong Xiao** [2]

摘要

在这项工作中，我们从二阶几何的视角重新审视 Transformer 优化，并建立了架构设计、激活尺度、海森矩阵与最大可容忍学习率之间的直接联系。我们提出了一种简单的归一化策略，称为 SimpleNorm，它从设计上稳定中间激活尺度。随后，通过分析网络激活相对于损失函数的海森矩阵，我们从理论上表明，SimpleNorm 显著降低了海森矩阵的谱范数，从而允许使用更大的稳定学习率。我们在参数规模为 10亿、14亿、70亿和80亿的 GPT模型上进行了大量实验，以验证理论发现。在实证方面，我们基于 SimpleNorm 构建的网络 SimpleGPT 能够容忍 3×-10× 高于标准惯例的学习率，始终表现出出色的优化稳定性，并取得远优于成熟基线的性能。具体而言，在训练 70亿规模的模型6万步时，SimpleGPT 的训练损失比采用 QKNorm 的 LLaMA2 低 0.08，从 2.290 降至 2.208。我们的源代码将在此发布：h
t
tps://github.com/Ocram7/SimpleGPT。

1. 引言

基于 Transformer 的大语言模型（LLM）（Rad-ford 等人，2018；2019；Brown 等人，2020；Touvron等人，2023a；b；Dubey 等人，2024年；Chowdhery 等人，2023；Liu 等人，2024年；Team，2023）在广泛的任务中取得了最先进的性能。随着这些模型在深度和宽度上不断扩展，优化稳定性日益成为制约性能和可扩展性的因素。现代 Transformer 中的许多架构组件 —— 例如残差连接（ He 等人，2016）、归一化层（Ioffe 和 Szegedy，2015；Ba 等人，2016；Zhang 和 Sennrich，2019；Large 等人，2024年），以及非线性激活（Shazeer，2020）——主要是为了稳定训练而引入的，在某些情况下也用于提升表达能力。因此，从原则性视角理解优化对于设计可扩展的 Transformer 架构至关重要。

经典优化理论（Nesterov，1983年；1998年；Nocedal & Wright，1999；Boyd & Vandenberghe，2004年）精确揭示了优化稳定性与二阶几何之间的联系。对于二阶可微目标函数 $\ell(\boldsymbol{x})$，其局部曲率由海森矩阵 $\boldsymbol{H}_{\boldsymbol{xx}} = \nabla^2\ell(\boldsymbol{x})$刻画。若 ℓ 是 β-光滑的，则$\|\nabla\ell(\boldsymbol{x}) - \nabla\ell(\boldsymbol{y})\|_2 \le \beta\|\boldsymbol{x}-\boldsymbol{y}\|_2$, $\forall \boldsymbol{x}$， $\boldsymbol{y}$. 标准结果表明，只有当最大可容许学习率 η 满足以下条件时，梯度下降才是稳定的

$$\eta \le \frac{2}{\beta} = \frac{2}{\sup_{\boldsymbol{x}} \|\boldsymbol{H}_{\boldsymbol{xx}}\|_2},$$

这确立了海森谱范数是决定可接受学习率和收敛行为的基本量。

相比之下，近期大量关于 Transformer 优化的文献侧重于架构启发式方法，却没有明确分析这些方法与经典优化理论之间的关系。归一化位置（Vaswani 等人，2017年；Wang 等人，2019年；Henry 等人，2020年；Qi 等人，2023a；2025c；a）、残差缩放（He 等人，2016年；Bachlechner 等人，2021年；Xie 等人，2025年）或改进的非线性函数（Hendrycks，2016年；Shazeer，2020年）等技术，通常仅凭经验结果加以论证，而它们对激活尺度、海森几何以及由此决定的最优学习率的影响仍未得到明确说明。因此，尽管网络设计与经典稳定性条件之间的理论联系对于训练超深层、大规模模型具有重要意义，我们对这种联系仍缺乏充分理解。

在本研究中，我们以基于海森矩阵的优化理论为核心视角，在考察激活尺度作用的同时，通过分析 Transformer 架构弥合了这一差距。我们提出了一种简单的归一化策略，将其命名为*SimpleNorm*，该策略通过紧随线性映射之后进行归一化来稳定中间激活尺度 。基于这一结构特性，我们分析了损失函数相对于网络激活值的海森矩阵，并表明 SimpleNorm 能够显著降低海森矩阵的谱范数，从而允许使用大幅提高的

*贡献相同 [1]清华大学 [2]云天励飞。通讯作者：Xianbiao Qi < qixianbiao@gmail.com>。

预印本。*2026*年*2*月*3*日。

学习率。通过将 Transformer 设计建立在经典优化原则（Nesterov，1983年；1998年；2013年）之上，我们的框架为现有的稳定化技术提供了统一解释，并为构建可扩展且稳定的模型提供了有原则的指导。

我们的贡献可总结如下：

- 我们从二阶几何的视角重新审视Transformer优化，并在架构设计、激活尺度、海森矩阵与最大可容忍学习率之间建立直接联系。

- 我们提出SimpleGPT，一种基于SimpleNorm的新GPT架构，并从理论上证明该设计显著降低了 $\|\boldsymbol{H}_{\boldsymbol{xx}}\|_2$，从而得到更小的梯度的利普希茨常数，并支持大幅更高的稳定学习率。

- 我们通过实验证明，这些理论优势在nanoGPT、Llama2和Llama3架构上均带来一致的实证收益，模型规模从10亿到80亿参数不等。具体而言，在以6万步训练70亿规模模型时，我们的方法取得的训练损失比采用QKNorm的Llama2低0.08，将损失从2.290降至2.208。

2. 相关工作

归一化方法。长期以来，归一化一直是稳定优化过程和提升深度网络收敛性的核心工具。批归一化（BN）（Ioffe 和 Szegedy，2019年）利用小批量统计量对激活值进行归一化，在卷积架构中取得了广泛成功，但其表现可能受批次大小和分布式同步的影响。层归一化（LN）（Ba 等人，2016）及其变体通过对每个样本的特征计算统计量来消除对批次的依赖，现已成为 Transformer 中的标准方法。实例归一化（IN）（Ulyanov 等人，2016）、组归一化（GN）（Wu 和 He，2018）、均方根归一化（RMSNorm）（Zhang 和 Sennrich，2019年）以及 nGPT（Loshchilov 等人，2025）等相关方法，则进一步针对特定的架构约束或效率需求调整归一化方式。

Transformer 中的归一化位置。除了归一化算子的选择之外，其位置在 Transformer 模块中的安排，对优化稳定性起着关键作用。原始 Transformer 架构采用后归一化（PostNorm），即在残差相加后进行归一化（Vaswani 等人，2017年）。随后，大规模实践转向预归一化（PreNorm），即在注意力和多层感知机子层之前进行归一化，以提高深层网络的可训练性（Wang 等人，2019年）。

近期研究进一步系统化了归一化位置，并探索了其他插入点。深度归一化 Transformer（DNT）（Qi 等人，2025年 a）对多种策略进行了分类，包括 InputNorm、PreNorm、MidNorm、PostNorm 和 QKNorm，并通过雅可比矩阵与梯度稳定性分析阐释其动机。DNT 最终结合了 InputNorm、PreNorm、MidNorm 和 QKNorm，同时因 PostNorm 可能导致训练不稳定而避免采用。在这些归一化位置中，QK 归一化（QKNorm）（Henry 等人，2020）专门针对注意力机制，稳定查询–键交互的几何结构，并缓解 softmax 饱和。

通过将归一化视为同时涵盖算子与位置的设计空间，这些研究强调，稳定性和条件性可以针对特定的架构子组件进行优化，而不只是针对块输出。随着模型深度增加，归一化也会与残差路径和初始化相互作用。例如，DeepNorm（Wang 等人，2022）会调整残差缩放和初始化，以限制参数更新并控制随深度增加而产生的动态增长，从而补充归一化位置策略。

无归一化 Transformer。受归一化的成本与复杂性，以及简化训练动态的需求驱动，近期研究开始质疑 Transformer 是否有必要采用显式归一化。无归一化的 *Transformer* 表明，以简单的逐点非线性函数——动态双曲正切（DyT）（Zhu 等人，2025）替代归一化层，能够在各项任务上达到与归一化基线相当的表现。这表明，恰当的有界非线性函数可以提供通常归因于 LN/RMSNorm 的大部分稳定性。在此基础上，更强的无归一化 *Transformer* (Chen 等人，2025) 更广泛地研究了逐点函数的设计，并报告称，通过搜索得到的函数族（例如 Derf），无归一化性能有所提升，且在多个领域的表现优于 LN/RMSNorm/DyT。尽管这些方法被描述为无归一化方法，但其本质上是通过有界变换控制激活值的范数，因此可以视为一种隐式归一化。

我们工作的定位。虽然我们的方法可以被视为对 Transformer 中归一化位置的研究，但其关键区别在于明确将架构设计与二阶优化几何联系起来。我们不再以启发式或经验性的方式论证归一化的必要性，而是分析紧随线性映射之后进行局部归一化，如何稳定激活尺度，进而约束海森矩阵的谱范数，并使优化景观更加平滑。这一视角不仅为最大可容忍学习率提供了原则性的刻画，也为大型 Transformer 模型的优化稳定性给出了统一的理论解释。

3. 预备知识

我们考虑如下无约束凸优化问题：$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$，其中 $f\colon \mathbb{R}^d \to \mathbb{R}$ 是可微的。

3.1. 凸优化与平滑优化

Lipschitz 梯度平滑性。 如果 f 二阶可微，则其在点 $\boldsymbol{x}$ 处的二阶泰勒展开为

$$f(\boldsymbol{y}) \approx f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^\top \nabla^2 f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}).$$

二阶项刻画了局部曲率。

定义 3.1（β-平滑性）。 如果函数 f 是β-平滑的，则

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2 \le \beta \|\boldsymbol{y}-\boldsymbol{x}\|_2, \forall \boldsymbol{y}, \boldsymbol{x}.$$

对于凸且可微的函数， β-光滑性等价于以下二次上界：

$$f(\boldsymbol{y}) \le f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2, \forall \boldsymbol{y}, \boldsymbol{x}.$$

这个不等式在优化方法的步长（学习率）选择中起着核心作用。

梯度下降与学习率。 考虑标准的梯度下降迭代

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k),$$

其中 $\eta > 0$ 是学习率。我们希望了解 η 的选择如何取决于光滑常数 β，以及这一选择如何影响收敛。

下降条件。 将 $\boldsymbol{y} = \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)$ 和 $\boldsymbol{x} = \boldsymbol{x}_k$ 代入二次上界，可得

$$f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{x}_k) - \eta \|\nabla f(\boldsymbol{x}_k)\|_2^2 + \frac{\beta\eta^2}{2}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

整理各项可得

$$f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{x}_k) - \left(\eta - \frac{\beta\eta^2}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

因此，目标函数单调下降的一个充分条件是

$$0 < \eta \le \frac{2}{\beta}. \tag{1}$$

该界刻画了凸 β-光滑函数上梯度下降的 稳定区域 。 $\frac{2}{\beta}$ 通常称为最大可容忍学习率。

对于凸的 β-光滑问题，在所有能够确保下降的固定学习率中，典型的标准选择是$\eta = \frac{1}{\beta}$。这种选择兼顾了优化进展与稳定性，并带来最紧的最坏情况保证。

3.2. 线性投影的梯度和海森矩阵

给定线性投影 $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ 和损失函数 ℓ， ℓ相对于 $\boldsymbol{y}$ 的梯度和海森矩阵为，

$$\boldsymbol{g}_{\boldsymbol{y}} := \frac{\partial \ell}{\partial \boldsymbol{y}}^\top, \qquad \boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}} := \frac{\partial^2 \ell}{\partial \boldsymbol{y}\partial \boldsymbol{y}^\top}.$$

$\boldsymbol{y}$关于 $\boldsymbol{x}$的雅可比矩阵为 $\boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{W}$.因此，根据链式法则，我们有

$$\boldsymbol{g}_{\boldsymbol{x}} = \frac{\partial \ell}{\partial \boldsymbol{x}}^\top = \boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}\top}\boldsymbol{g}_{\boldsymbol{y}} = \boldsymbol{W}^\top \boldsymbol{g}_{\boldsymbol{y}},$$

而关于 $\boldsymbol{x}$ 的海森矩阵为

$$\boldsymbol{H}_{\boldsymbol{x}\boldsymbol{x}} = \frac{\partial^2 \ell}{\partial \boldsymbol{x}\partial \boldsymbol{x}^\top} = \boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}\top}\boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}}\boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}} = \boldsymbol{W}^\top \boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}}\boldsymbol{W}. \tag{2}$$

4. 方法

4.1. SimpleNorm：统一的归一化策略

SimpleNorm 的定义。 我们将*SimpleNorm*定义为在线性映射之后立即 放置一个归一化算子。给定输入向量 $\boldsymbol{x} \in \mathbb{R}^m$ 和线性变换 $\boldsymbol{W} \in \mathbb{R}^{d\times m}$，我们将 SimpleNorm 抽象为一种原始算子

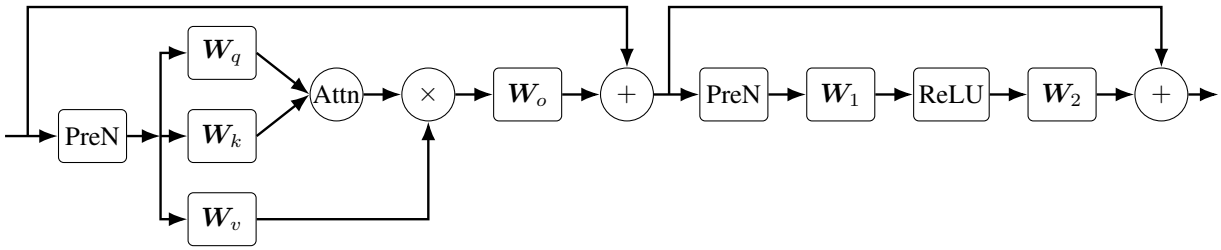
$$\boldsymbol{\Psi}(\boldsymbol{x}) = \mathrm{Norm}(\boldsymbol{W}\boldsymbol{x}), \tag{3}$$

其中，Norm(·) 是一种归一化算子，例如 LayerNorm 或均方根归一化。SimpleNorm 的设计动机是一种简单而有效的放置 策略，而非代数复杂性。与通常作用于残差块、隐藏状态或参数重参数化层面的现有归一化技术不同，SimpleNorm 会在线性映射之后就地立即 执行归一化，将“线性映射后立即进行归一化”视为一个统一算子。

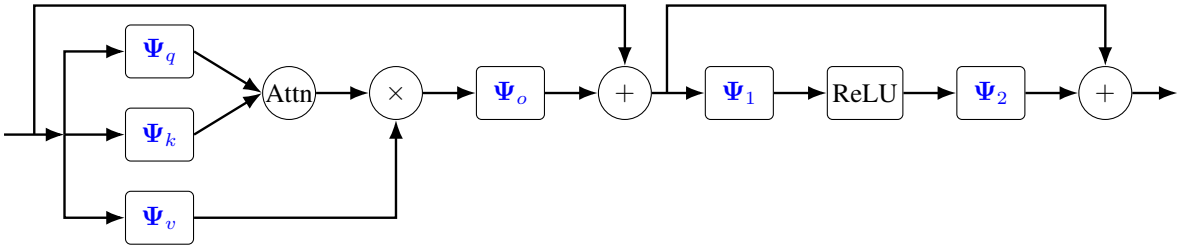
SimpleGPT 的定义。如 图1所示，*SimpleGPT* 将 SimpleNorm 用作基础构建模块。凡是出现线性层的地方，都会系统地插入 SimpleNorm，包括多层感知机投影、注意力投影（Q、K、V）、输出投影，以及门控或记忆相关模块。以图1 为例，归一化被插入到 $\boldsymbol{W}_q$、 $\boldsymbol{W}_k$、 $\boldsymbol{W}_v$、 $\boldsymbol{W}_o$、 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$投影之后。在采用 SwiGLU（Shazeer, 2020）而非多层感知机的架构中，SimpleGPT 会在 $\boldsymbol{W}_q$、 $\boldsymbol{W}_k$、 $\boldsymbol{W}_v$、 $\boldsymbol{W}_o$、 $\boldsymbol{W}_1$、 $\boldsymbol{W}_2$和 $\boldsymbol{W}_3$之后插入归一化。

用均方根归一化实例化 SimpleNorm 在这项工作中，我们将$\mathrm{Norm}(\cdot)$ 用均方根归一化（Zhang 和 Sennrich, 2019年）实例化。因此，SimpleNorm 现在定义如下：

$$\boldsymbol{\Psi}(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{\gamma}) = \boldsymbol{\gamma} \odot \sqrt{d}\,\frac{\boldsymbol{W}\boldsymbol{x}}{\|\boldsymbol{W}\boldsymbol{x}\|_2}, \tag{4}$$



(a) GPT。GPT 采用预归一化架构。应用线性投影 W_q、W_k、W_v、W_o、W_1 和 W_2。



(b) SimpleGPT。SimpleGPT 将所有线性层替换为 **SimpleNorm** 算子，记为 Ψ。在 SimpleGPT 中，我们不使用预归一化 .

图*1*。SimpleGPT 与 GPT 的比较。该图比较了标准 GPT 模块与提出的 SimpleGPT 模块，重点展示了 SimpleNorm 引入的结构简化。

其中 $\boldsymbol{x} \in \mathbb{R}^m$、$\boldsymbol{W} \in \mathbb{R}^{d\times m}$ 和 $\boldsymbol{W}$，$\boldsymbol{\gamma}$是可学习参数，而 $\odot$ 表示逐元素相乘。

为便于后续分析，我们定义中间变量

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}, \qquad s = \|\boldsymbol{z}\|_2, \qquad \boldsymbol{u} = \frac{\boldsymbol{z}}{s},$$
$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top, \quad \boldsymbol{D} = \mathrm{Diag}(\boldsymbol{\gamma}),$$

使得 $\boldsymbol{\Psi}(\boldsymbol{x};\ \boldsymbol{W},\ \boldsymbol{\gamma}) = \sqrt{d}\,\boldsymbol{D}\boldsymbol{u}$。

SimpleNorm 的核心性质 以下各节将证明 SimpleNorm 的两种重要机制：SimpleNorm 可直接将激活值的尺度稳定在 $\sqrt{d}$ 量级，并约束损失函数关于激活值的海森矩阵的谱范数，从而平滑损失景观并支持更大的学习率。此外，我们还提出一种假说，解释为何 SimpleNorm 除了具有预期的优化稳定性之外，在实际表现中也十分出色。

4.2. 机制 I：稳定的激活尺度

根据构造，SimpleNorm 在每次线性映射后立即进行归一化，从而稳定中间激活值的尺度。回顾一下，

$$\boldsymbol{\Psi}(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\gamma}) = \sqrt{d}\,\boldsymbol{D}\boldsymbol{u} \qquad \text{with } \|\boldsymbol{u}\|_2 = 1.$$

Then

$$\|\boldsymbol{\Psi}(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\gamma})\|_2 = \sqrt{d}\,\|\boldsymbol{D}\boldsymbol{u}\|_2.$$

特别地，令 $\gamma_{\min} = \min_i|\gamma_i|$ and $\gamma_{\max} = \max_i|\gamma_i|$，我们可以得到以下界

$$\gamma_{\min}\sqrt{d} \ \le\ \|\boldsymbol{\Psi}(\boldsymbol{x})\|_2 \ \le\ \gamma_{\max}\sqrt{d}.$$

因此，除学习到的逐维缩放 $\boldsymbol{\gamma}$外，每个投影都会被重新缩放，使其范数处于 $\sqrt{d}$ 量级。因此，SimpleNorm可防止中间表示范数随网络深度增加或权重增长而漂移，从而消除激活爆炸的一个常见来源。机制 II 表明，除了稳定激活值外，SimpleNorm 还通过梯度利普希茨常数来稳定曲率。

4.3. 机制 II：更平滑的损失景观

平滑性与海森矩阵。 目标函数的平滑性直接制约优化稳定性：曲率越大，安全学习率越小。给定一个二阶可微的 β-平滑目标函数 $\ell(x)$，我们用激活海森矩阵 $\boldsymbol{H}_{\boldsymbol{x}\boldsymbol{x}} = \nabla^2_{\boldsymbol{x}\boldsymbol{x}}\ell$量化局部曲率。具体而言，海森矩阵谱范数的上确界为局部曲率提供上界，并决定梯度稳定性：

$$\beta \ =\ \sup_{\boldsymbol{x}} \|\boldsymbol{H}_{\boldsymbol{x}\boldsymbol{x}}(\boldsymbol{x})\|_2.$$

为证明 SimpleNorm 能带来更平滑的损失景观，我们证明了两个结果：(i) SimpleNorm 海森矩阵可分解为 $\boldsymbol{H}_{\boldsymbol{x}\boldsymbol{x}} = \boldsymbol{L} + \boldsymbol{C}$，且在高维情况下 $\|\boldsymbol{C}\|_2 \ll \|\boldsymbol{L}\|_2$；(ii) 与曲率随 $\|\boldsymbol{W}\|_2^2$ 缩放的线性投影相比，SimpleNorm 的曲率相对于 $\|\boldsymbol{W}\|_2$ 具有尺度不变性。综合来看，这意味着 $\|\boldsymbol{H}^{\mathrm{sn}}_{\boldsymbol{x}\boldsymbol{x}}\|_2 \ll \|\boldsymbol{H}^{\mathrm{lin}}_{\boldsymbol{x}\boldsymbol{x}}\|_2$，因为 $\|\boldsymbol{W}\|_2$ 通常会在训练过程中增大。

SimpleNorm 导数。 首先，我们计算 ℓ 相对于 $\boldsymbol{x}$ 的一阶梯度和二阶海森矩阵。

Given $\boldsymbol{y} = \sqrt{d}\,\boldsymbol{D}\boldsymbol{u}$, let

$$\ell = \ell(\boldsymbol{y}), \qquad \boldsymbol{g}_{\boldsymbol{y}} := \nabla_{\boldsymbol{y}}\ell, \qquad \boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}} := \nabla^2_{\boldsymbol{y}\boldsymbol{y}}\ell.$$

一阶导数。归一化的雅可比矩阵满足 $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{z}} = \frac{1}{s}\boldsymbol{P}$，由此得到 $\overline{\boldsymbol{y}}$ 相对于 $\boldsymbol{x}$ 的雅可比矩阵：

$$\boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}} := \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \frac{\sqrt{d}}{s}\boldsymbol{D}\,\boldsymbol{P}\,\boldsymbol{W}.$$

应用链式法则，梯度为

$$\nabla_{\boldsymbol{x}}\ell = \boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}\top}\boldsymbol{g}_{\boldsymbol{y}} = \frac{\sqrt{d}}{s}\boldsymbol{W}^{\top}\boldsymbol{P}\,\boldsymbol{D}\,\boldsymbol{g}_{\boldsymbol{y}}. \tag{5}$$

二阶导数。对梯度求导可得标准分解

$$\boldsymbol{H}_{\boldsymbol{xx}} = \nabla_{\boldsymbol{x}}^2\ell = \underbrace{\boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}\top}\boldsymbol{H}_{\boldsymbol{yy}}\,\boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}}}_{\text{Gauss–Newton term}} + \underbrace{\boldsymbol{C}}_{\text{curvature term}}, \tag{6}$$

其中第一项是高斯–牛顿分量

$$\boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}\top}\boldsymbol{H}_{\boldsymbol{yy}}\boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}} = \frac{d}{s^2}\boldsymbol{W}^{\top}\boldsymbol{P}\,\boldsymbol{D}\,\boldsymbol{H}_{\boldsymbol{yy}}\,\boldsymbol{D}\,\boldsymbol{P}\,\boldsymbol{W}, \tag{7}$$

第二项则来自归一化的曲率，

$$\boldsymbol{C} = -\frac{\sqrt{d}}{s^2}\boldsymbol{W}^{\top}\Big(\boldsymbol{P}\boldsymbol{D}\boldsymbol{g}_{\boldsymbol{y}}\boldsymbol{u}^{\top} + \boldsymbol{u}^{\top}\boldsymbol{D}\boldsymbol{g}_{\boldsymbol{y}}\boldsymbol{P} + \boldsymbol{u}\boldsymbol{g}_{\boldsymbol{y}}^{\top}\boldsymbol{D}\boldsymbol{P}\Big)\boldsymbol{W}. \tag{8}$$

详见附录 A，其中详细推导了 $\nabla_{\boldsymbol{x}}\ell$和 $\nabla_{\boldsymbol{x}}^2\ell$（针对 SimpleNorm）。为完整起见，$\nabla_{\boldsymbol{\gamma}}\ell$、$\nabla_{\boldsymbol{\gamma}}^2\ell$、$\nabla_{\boldsymbol{W}}\ell$和$\nabla^2_{\mathrm{vec}(\boldsymbol{W})}\,\ell$与 $\boldsymbol{y} = \boldsymbol{\gamma}\odot\sqrt{d}\,\frac{\boldsymbol{W}\boldsymbol{x}}{\|\boldsymbol{W}\boldsymbol{x}\|_2}$和 $\nabla_{\boldsymbol{W}}\ell$和 $\nabla^2_{\mathrm{vec}(\boldsymbol{W})}\ell$与 $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$的推导也见附录 C和附录 B

高维下高斯–牛顿项占主导。接下来，我们证明，在标准的高维且非病态条件下，高斯–牛顿项主导由归一化引起的曲率。

定理 4.1(SimpleNorm 的高斯–牛顿主导性)。令 $\boldsymbol{H}_{\boldsymbol{xx}} = \nabla_{\boldsymbol{x}}^2\ell$ 表示可二次微分的目标函数 $\ell(\boldsymbol{y})$经 *SimpleNorm* 诱导的激活海森矩阵。那么，海森矩阵可分解为

$$\boldsymbol{H}_{\boldsymbol{xx}} = \boldsymbol{L} + \boldsymbol{C}, \qquad \boldsymbol{L} = (\boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}})^{\top}\boldsymbol{H}_{\boldsymbol{yy}}\boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}},$$

其中，$\boldsymbol{C}$ 是归一化诱导的曲率项。

假设 $\|\boldsymbol{x}\|_2 = \sqrt{d}$、$\overline{\boldsymbol{D}} = \boldsymbol{I}$、$\boldsymbol{W} \in \mathbb{R}^{d\times d}$ 具有较高的有效秩 $\|\boldsymbol{W}\|_F^2/\|\boldsymbol{W}\|_2^2 \geq c\,d$，且输入和损失导数与 $\boldsymbol{W}$ 的对齐不存在病态情况。定义

$$\kappa := \left\|\frac{\sqrt{d}\,\boldsymbol{W}}{\|\boldsymbol{W}\boldsymbol{x}\|_2}\right\|_2.$$

于是，以高概率有 $\kappa = \Theta(1)$，且存在一个常数 $\tau = \Theta(1)$使得

$$\|\boldsymbol{L}\|_2 = \tau\,\kappa^2\,\|\boldsymbol{H}_{\boldsymbol{yy}}\|_2, \qquad \|\boldsymbol{C}\|_2 \leq \frac{3\kappa^2}{\sqrt{d}}\,\|\boldsymbol{g}_{\boldsymbol{y}}\|_2.$$

特别地，如果 $\|\boldsymbol{g}_{\boldsymbol{y}}\|_2 = O(\|\boldsymbol{H}_{\boldsymbol{yy}}\|_2)$，那么

$$\|\boldsymbol{C}\|_2 \ll \|\boldsymbol{L}\|_2$$

因此，以高概率，高斯–牛顿项主导海森矩阵。

A complete proof is given in Appendix D.

SimpleNorm 海森矩阵具有权重尺度不变性。最后，我们将 SimpleNorm 海森矩阵的大小与普通线性投影的大小进行比较。我们证明，线性曲率会随权重矩阵的谱范数$\|\boldsymbol{W}\|_2$的增大而呈二次增长，而 SimpleNorm 消除了这种依赖。

定理 4.2（线性曲率随 $\|\boldsymbol{W}\|_2^2$ 缩放，而 SimpleNorm 不会）。设 $\ell = \ell(\boldsymbol{y})$二阶可微，且满足 $\boldsymbol{H}_{yy} = \nabla_{\boldsymbol{yy}}^2\ell$和 $\boldsymbol{g}_{\boldsymbol{y}} = \nabla_{\boldsymbol{y}}\ell$。
考虑其海森矩阵为

$$\boldsymbol{y}_1 = \boldsymbol{W}_1\boldsymbol{x}, \qquad \boldsymbol{H}_{xx}^{\mathrm{lin}} = \boldsymbol{W}_1^{\top}\boldsymbol{H}_{yy}\boldsymbol{W}_1,$$

以及其海森矩阵为

$$\boldsymbol{y}_2 = \boldsymbol{D}\frac{\sqrt{d}\,\boldsymbol{W}_2\boldsymbol{x}}{\|\boldsymbol{W}_2\boldsymbol{x}\|_2}, \qquad \boldsymbol{H}_{xx}^{\mathrm{sn}} = \boldsymbol{L} + \boldsymbol{C}.$$

假设 $\boldsymbol{H}_{y_1y_1} = \boldsymbol{H}_{y_2y_2} := \boldsymbol{H}_{yy}$，$\boldsymbol{W}_1 = \boldsymbol{W}_2 := \boldsymbol{W}$，且满足定理 4.1 的条件，使得 $\|\boldsymbol{L}\|_2 \gg \|\boldsymbol{C}\|_2$。那么，以高概率有：

$$\|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{sn}}\|_2 = \Theta\big(\kappa^2\,\|\boldsymbol{H}_{\boldsymbol{yy}}\|_2\big), \qquad \kappa^2 = \frac{d}{\|\widetilde{\boldsymbol{W}}\boldsymbol{x}\|_2^2} = \Theta(1),$$

where $\widetilde{\boldsymbol{W}} = \boldsymbol{W}/\|\boldsymbol{W}\|_2$.

此外，如果 $\boldsymbol{W}$ 的值域与 $\boldsymbol{H}_{\boldsymbol{yy}}$ 的主特征子空间并非对抗性对齐，那么存在一个常数 $c_{\mathrm{lin}} = \Theta(1)$使得

$$\|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{lin}}\|_2 = \|\boldsymbol{W}^{\top}\boldsymbol{H}_{\boldsymbol{yy}}\boldsymbol{W}\|_2 \;\geq\; c_{\mathrm{lin}}\,\|\boldsymbol{W}\|_2^2\,\|\boldsymbol{H}_{\boldsymbol{yy}}\|_2.$$

因此，随着训练过程中 $\|\boldsymbol{W}\|_2$ 增大，

$$\|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{lin}}\|_2 \;\gg\; \|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{sn}}\|_2 \qquad \textit{(with high probability)}.$$

直观而言，SimpleNorm 通过归一化激活值，消除了曲率对权重尺度的依赖；而随着 $\|\boldsymbol{W}\|_2$ 增大，线性投影会放大曲率。我们在附录 E中给出定理 4.2的证明。

SimpleNorm 支持更大的学习率。对于二阶可微的 β-平滑目标函数，梯度下降的最大稳定学习率与 β成反比，其中 β是梯度的 Lipschitz常数，等价于海森矩阵谱范数的上确界：$\eta \leq \frac{2}{\beta}$其中 $\beta = \sup_{\boldsymbol{x}}\|\boldsymbol{H}_{\boldsymbol{xx}}(\boldsymbol{x})\|_2$

–

定理 4.1 和 定理 4.2 证明，在标准的高维且非病态条件下，

SimpleNorm 的海森矩阵不受权重矩阵谱范数的影响，而线性投影的海森矩阵则随权重范数的平方变化。

因此，由于权重谱范数通常会在训练过程中不断增大，基于 SimpleNorm 的 SimpleGPT 架构拥有更平滑的损失景观，与基于直接线性投影的方法相比，能够容忍显著更大的学习率。

4.4. 解释：超越优化稳定性

我们已经确立了 SimpleNorm 的两项核心性质：(i) 它能稳定 $\Theta(\sqrt{d})$处的激活尺度；(ii) 它通过约束激活海森矩阵的谱范数来平滑损失景观，从而支持更大且更稳定的学习率。尽管这些性质解释了 SimpleNorm 优异的优化稳定性，却无法充分解释其在第 5 节中展现出的强劲实证表现。

我们假设，SimpleNorm 在更全局的表征层面还带来了额外优势。通过在每次线性投影后立即进行归一化，SimpleNorm 确保每一层都能产生真正的非线性变换，即使在周围网络原本表现得近乎线性的情况下也是如此。这实际上增加了非线性交互深度，并在不增加参数数量的前提下增强了表达能力。

在这一视角下，SimpleNorm 通过双重作用提升性能：局部而言，它通过降低曲率变化改善优化几何；全局而言，它通过普遍存在的归一化诱导非线性增强表达能力。我们认为，这种组合解释了为何 SimpleNorm 能带来持续的实证收益，且超出仅凭学习率稳定性所能预期的收益。

4.5. 使用 torch.compile 加速训练

归一化层受内存限制且执行频繁，因此可能成为瓶颈。通过融合归约操作与逐点操作，并利用 torch.compile，SimpleNorm 增加的归一化开销大多得以摊销，使训练时间相比带 QKNorm 的 GPT 仅增加约 3%。

5. 实验

实验设置。 我们在三个 Transformer 主干模型上评估 SimpleNorm：nanoGPT、Llama2 和 Llama3。SimpleNorm 应用于所有 Transformer 块，但不包括嵌入层和输出层。所有模型均使用 Kingma& Ba，2014；Loshchilov & Hutter，2019年提出的 AdamW 优化器进行训练，并采用余弦学习率调度和 bfloat16 精度。针对每种方法调整学习率。由于 SimpleNorm 允许使用明显更大的稳定学习率，因此我们相应地调整权重衰减。

有关架构、超参数和训练的更多细节见 附录 F和 附录 G。

5.1. 最大可容忍学习率

我们通过比较不同归一化方案下的优化稳定性，同时保持其他训练设置不变，来评估最大可容忍学习率。如图 2所示，PreNorm 在学习率为 2×10^{-3}时已出现收敛问题。相比之下，PreNorm+QKNorm 在 2×10^{-3} 和 2×10^{-2}时仍保持稳定，但学习率升高至2×10^{-1}时则变得不稳定。SimpleNorm 在2×10^{-3} 和 2×10^{-2}下均能稳定收敛，并且在 2×10^{-1}时明显比 PreNorm+QKNorm 更稳定。总体而言，这些结果表明，SimpleNorm 能够始终耐受更高的学习率，说明其优化鲁棒性更强。

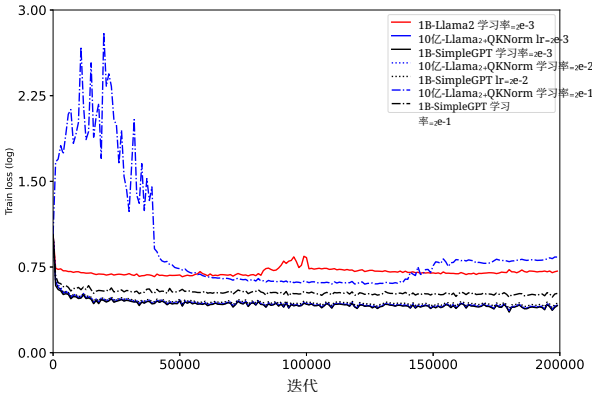
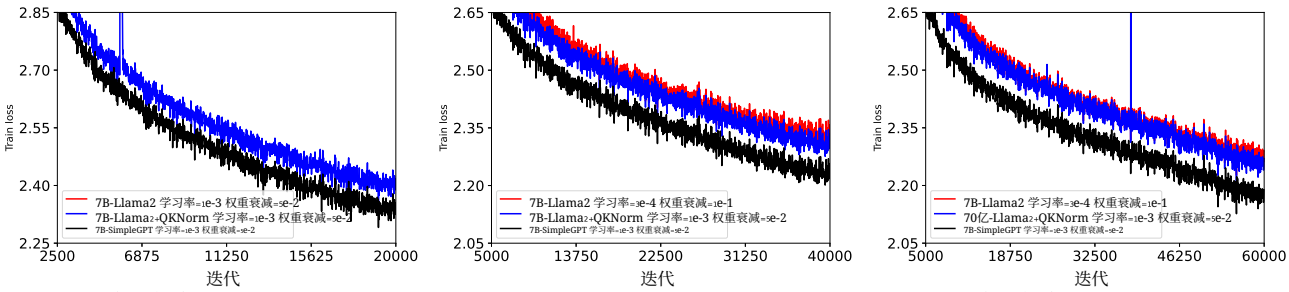


图 2. Llama2-B、采用 QKNorm 的 Llama2-1B 以及 SimpleGPT-1B 的最大可接受学习率。

5.2. 基于 Llama2 的 SimpleGPT 10亿

在本小节中，我们评估 SimpleGPT 1B，并将其与标准的 Llama2 1B 以及采用 QKNorm 的 Llama2 1B 进行比较。我们按照（Zhang 等人，2024年）的方法训练模型 200K 步，全局批次大小为 256，序列长度为 512，训练词元总数约为 26B。我们在 C4数据集上训练所有模型，并遵循各自对应基线的相同训练方案。结果如图 4所示。Tensorboard 中的损失曲线经过了 80% 的平滑处理。对于所有实验，我们报告最后一步的训练损失。

在 图 4中，SimpleGPT 1B 相较于采用 QKNorm 的 Llama2 1B 取得了显著提升。具体而言，训练损失从 2.478 降至 2.446，绝对提升为 0.032。因此，即使在较小的 1B 规模下，SimpleGPT 也能带来可衡量的收益。



7B-Llama2 学习率=1e-3 权重衰减=5e-2：最后 = 4.115
7B-Llama2+QKNorm 学习率=1e-3 权重衰减=5e-2：最后 = 2.393
7B-SimpleGPT 学习率=1e-3 权重衰减=5e-2：最后 = 2.331

(a) SimpleGPT 7B，训练 20K 步。

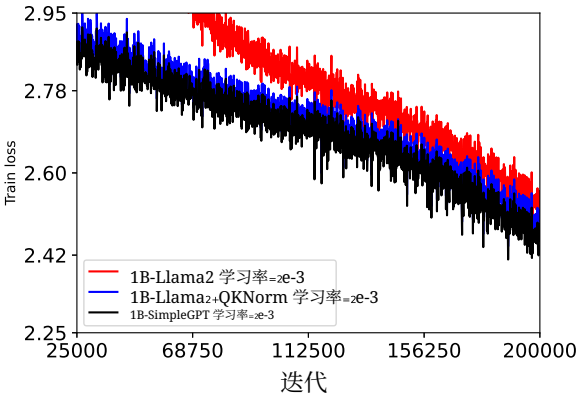
7B-Llama2 lr=3e-4 wd=1e-1：最后 = 2.345
7B-Llama2+QKNorm lr=1e-3 wd=5e-2：最后 = 2.320
7B-SimpleGPT lr=1e-3 wd=5e-2：最后 = 2.243

(b) 训练步数为4万步的 SimpleGPT 7B。

7B-Llama2 学习率=3e-4 权重衰减=1e-1：最后 = 2.311
7B-Llama2+QKNorm 学习率=1e-3 权重衰减=5e-2：最后 = 2.290
7B-SimpleGPT 学习率=1e-3 权重衰减=5e-2：最后 = 2.208

(c) 训练步数为6万步的 SimpleGPT 7B。

图 *3.* 训练损失曲线：Llama2 7B、采用 QKNorm 的 Llama2 7B 和 SimpleGPT 7B 在训练步数为2万、4万和6万步时的训练损失曲线 s.



1B-Llama2 学习率=2e-3：最后 = 2.520
1B-Llama2+QKNorm 学习率=2e-3：最后 = 2.478
1B-SimpleGPT 学习率=2e-3：最后 = 2.446

图 *4.* Llama2 1B、采用 QKNorm 的 Llama2 1B 和 SimpleGPT 1B 在训练步数为20万步时的训练损失曲线。

5.3. 基于 Llama2 的 SimpleGPT 7B

我们在 图 3中将 SimpleGPT 7B 与标准 Llama2 7B 以及采用 QKNorm 的 Llama2 7B 进行比较。我们分别训练模型 2万、4万和6万步，对应约 80亿、160亿和240亿个词元。所有模型均在 C4数据集 上训练，并遵循与各自基线相同的训练方案。SimpleGPT 7B 使用 0.001 的学习率，比用于 Llama2（Touvron等人，2023年 b）7B模型的学习率更高。

我们得出以下观察结果。首先，SimpleGPT 相较于 Llama2+QKNorm 的性能提升在整个训练过程中始终显著：在训练 2 万步时提升达到 0.062，4 万步时增至 0.077，而在 6 万步时仍保持在相近水平（0.082）。其次，与采用 QKNorm 的 Llama2 相比，SimpleGPT 的训练动态更加稳定。第三，随着训练推进、观察到更多词元，相对提升并未减弱，这表明其优势属于 SimpleGPT 的稳定优势，而非训练初期的短暂效果。最后，我们观察到模型规模呈现出明显的缩放趋势。使用 260亿个词元训练的 1B模型提升幅度较小，约为 0.03；而使用 240亿个词元训练的 7B模型则取得了大得多的提升，达到 0.08。

5.4. 基于 Llama3 的 SimpleGPT 8B

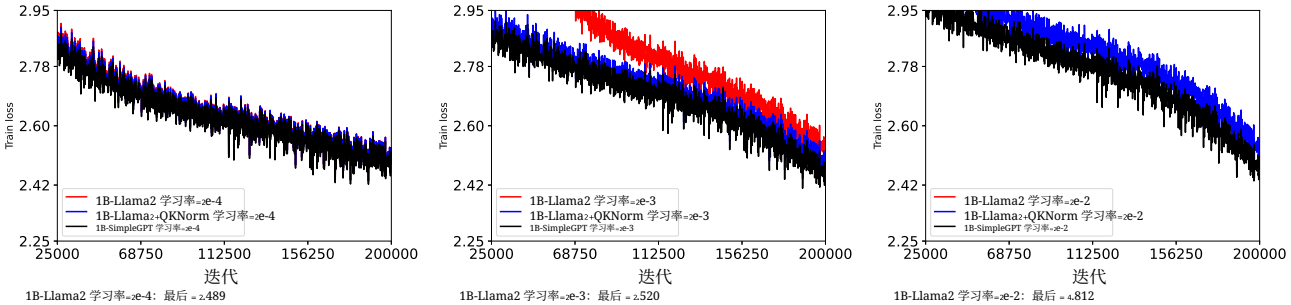
在 8B 规模下，我们的实验基于 Llama3 8B架构。我们在 C4数据集上训练 SimpleGPT 8B 和 Llama3 8B，全局批次大小为 192，序列长度为 2048。我们训练了 20K 步，对应约 8B 训练词元。由于计算资源限制，我们没有训练更多步。SimpleGPT 8B 使用了比 Llama3 8B 3×更高的学习率，正如图6所示，其训练损失显著更低。此外，性能提升幅度与 7B模型的观察结果一致，这表明随着模型规模增大，我们的方法展现出良好的缩放行为。

5.5. 基于 nanoGPT 的 SimpleGPT 1.4B

最后，我们在 nanoGPT 代码库上评估 SimpleGPT 1.4B。所有模型均训练 100K 步，对应约 50B 词元。SimpleGPT 1.4B 的训练学习率比基线高 3×。

我们在图7中报告验证损失。请注意，由于验证损失每 1,000步记录一次，图7 中的曲线与之前的图相比有所不同。

我们观察到，采用 QKNorm 的 GPT-2 性能几乎与原始 GPT-2 相同，这表明在这种设置下，单独使用 QKNorm 带来的收益有限。与 LLaMA2 1B 上的结果一致，基于 nanoGPT 的 SimpleGPT 1.4B 带来了约 0.043 的提升。这些发现表明，SimpleNorm 带来的收益在不同架构中都很稳定。



1B-Llama2 学习率=2e-4：最后 = 2.489
1B-Llama2+QKNorm 学习率=2e-4：最后 = 2.488
1B-SimpleGPT 学习率=2e-4：最后 = 2.477

(a) 学习率为=2e-4 的 SimpleGPT 1B。

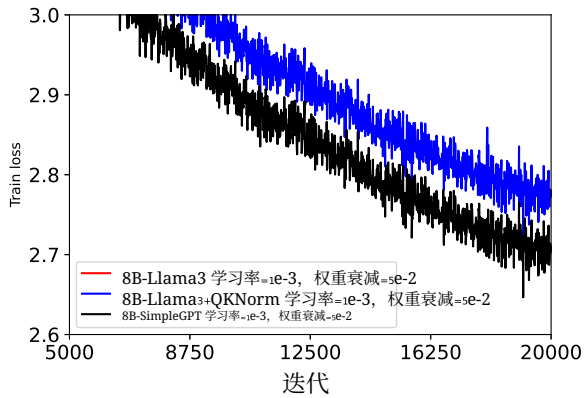
1B-Llama2 学习率=2e-3：最后 = 2.520
1B-Llama2+QKNorm 学习率=2e-3：最后 = 2.478
1B-SimpleGPT 学习率=2e-3：最后 = 2.446

(b) 学习率为=2e-3 的 SimpleGPT 1B。

1B-Llama2 学习率=2e-2：最后 = 4.812
1B-Llama2+QKNorm 学习率=2e-2：最后 = 2.517
1B-SimpleGPT 学习率=2e-2：最后 = 2.461

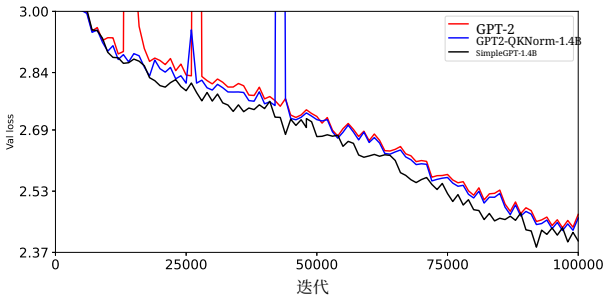
(c) 学习率为=2e-2 的 SimpleGPT 1B。

图*5*。 在三种不同学习率下，对 Llama2 1B、使用 QKNorm 的 Llama2 1B 和 SimpleGPT 1B 进行整体比较。Adam-mini 使用 2×10^{-4} 学习率。在 SimpleGPT 中，我们启用 $10\times$ 学习率，并获得了更好的性能。



8B-Llama3 学习率=1e-3，权重衰减=5e-2：最后 = 4.640
8B-Llama3+QKNorm 学习率=1e-3，权重衰减=5e-2：最后 = 2.752
8B-SimpleGPT 学习率=1e-3，权重衰减=5e-2：最后 = 2.678

图*6*。 Llama3 8B、使用 QKNorm 的 Llama3 8B 和 SimpleGPT 8B 的训练损失曲线。



GPT2-1.4B：损失 = 2.434，
GPT2-QKNorm-1.4B：损失 = 2.427，
SimpleGPT-1.4B：损失 = 2.384

图*7*. GPT2 1.4B、采用 QKNorm 的 GPT2 1.4B 和 SimpleGPT 1.4B 在 10 万训练步数下的验证损失曲线。

5.6. 消融研究

不同的学习率。如图 5所示，我们使用三种不同的学习率在 1B 模型上进行了实验。当学习率为 2×10^{-4}时，采用 QKNorm 的 LLaMA2 1B 仅略优于原始 LLaMA2 1B。当学习率提高到 2×10^{-3}时，QKNorm 带来的提升更为明显，我们认为这是由于与 PreNorm 相比，QKNorm 带来了更平滑的优化景观。重要的是，在所有学习率下，SimpleGPT 相较于采用 QKNorm 的 LLaMA2 1B 均取得了稳定的提升。为确保公平比较，报告的结果采用了表现最佳的 LLaMA2 1B 与 QKNorm 配置。

5.7. 关于训练时间的讨论

我们将 SimpleGPT 8B 模型的训练速度与采用 QKNorm 的 Llama3 8B 模型进行了比较。平均而言，Llama3 8B 模型每个训练步骤耗时 1553 毫秒，而我们的 SimpleGPT 8B 模型每步耗时 1603 毫秒。这相当于约 3% 的合理减速，通过更巧妙的内核设计，或将归一化算子替换为更适合融合的逐点函数（例如 Derf (Chen 等人，2025）。

6. 结论

在这项工作中，我们从二阶视角重新审视 Transformer 优化，并建立了架构设计、海森矩阵与优化稳定性之间的直接联系。通过引入 SimpleNorm 并分析其所诱导的海森结构，我们表明，降低激活相对于损失的海森范数，能够支持显著更大的可用学习率。由此得到的模型 **SimpleGPT** 能够可靠地支持比强基线高 3×-10× 倍的学习率，同时保持优化稳定。在针对 nanoGPT、Llama2 和 Llama3 风格模型开展的大量实验中，涵盖从 10亿到 80亿的参数规模，我们的方法始终比带 QKNorm 的 GPT 实现显著更强的性能。重要的是，这些提升仅带来了极少的额外计算开销。

影响声明

本文旨在推动机器学习领域的发展。我们的工作可能会带来许多社会影响，但我们认为没有哪项影响需要在此特别指出。

参考文献

Ba, J. L., Kiros, J. R., andHinton, G. E. 层归一化。*arXiv*预印本 *arXiv:1607.06450*，2016年。

Bachlechner, T., Majumder, B. P., Mao, H., Cottrell, G., and McAuley, J. ReZero就是你所需要的一切：大深度下的快速收敛。载于人工智能中的不确定性，第1352–1361页。PMLR，2021年。

Boyd, S.与Vandenberghe, L. 凸优化。剑桥大学出版社，2004年。

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shya m, P., Sastry, G., Askell, A. 等。语言模型是少样本学习器。神经信息处理系统进展，33: 1877–1901，2020。

Chen, M., Lu, T., Zhu, J., Sun, M. 和 Liu, Z. 更强的无归一化Transformer。*arXiv*预印本 *arXiv:2512.10938*，2025。

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S. 等。PaLM：通过 Pathways 扩展语言建模。机器学习研究期刊，24(240):1–113，2023。

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang,A., Fan, A., 等。《Llama 3 模型群》。*arXiv* 预印本 *arXiv:2407.21783*，2024年。

Gage, P. 一种用于数据压缩的新算法。*The CUsersJournal*，12(2):23–38，1994年。Golub, G. H. 和 Van Loan, C. F. 矩阵计算。JHU 出版社，2013年。He, K.、Zhang, X.、Ren, S. 和 Sun, J. 用于图像识别的深度残差学习。载于*IEEE* 计算机视觉与模式识别会议论文集，第 770–778 页，2016年。Hendrycks, D. 高斯误差线性单元（gelu）。*arXiv* 预印本 *arXiv:1606.08415*，2016年。Henry, A.、Dachapally, P. R.、Pawar, S. S. 和 Chen, Y. 面向 Transformer 的查询-键归一化。载于 计算语言学协会：*EMNLP 2020* 论文集，第 4246–4253 页，2020年。

Horn, R. A. 和 Johnson, C. R. 矩阵分析。剑桥大学出版社，2012年。

Ioffe, S. 和 Szegedy, C. 批归一化：通过减少内部协变量偏移来加速深度网络训练。载于国际机器学习会议，第 448–456 页。PMLR，2015年。

Karpathy, A. nanoGPT。https://github.com/karpathy/nanoGPT，2022年。

Kingma, D. P. 和 Ba, J. Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014。

Large, T., Liu, Y., Huh, M., Bahng, H., Isola, P. 和 Bernstein, J. 模块化范数中的可扩展优化。*arXiv* 预印本 *arXiv:2405.14813*，2024年。

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., 等。DeepSeek-V3 技术报告。*arXiv* 预印本 *arXiv:2412.19437*，2024年。

Loshchilov, I. 和 Hutter, F. 修正 Adam 中的权重衰减正则化。在国际学习表征会议，2019年。

Loshchilov, I.、Hsieh, C.-P.、Sun, S. 和 Ginsburg, B. nGPT：在超球面上进行表征学习的归一化 Transformer。第十三届国际学习表征会议，2025年。

Magnus, J. R. 和 Neudecker, H. 矩阵微分演算及其在统计学和计量经济学中的应用。约翰·威立父子出版公司，2019年。

Nesterov, Y. 一种无约束凸最小化问题的求解方法，其收敛速率为 o (1/kˆ 2)。载于苏联科学院院刊，第269卷，第543–547页，1983年。

Nesterov, Y.《凸规划导论讲座》第一卷：基础课程。讲义，3(4):5，1998年。Nesterov, Y. 凸优化导论：基础课程，第87卷。施普林格科学与商业媒体，2013年。Nocedal, J. 和 Wright, S. J. 数值优化。Springer，1999年。Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. 等。PyTorch：一种命令式风格、高性能的深度学习库。神经信息处理系统进展，32，2019年。Petersen, K. B., Pedersen, M. S. 等。《矩阵烹饪书》。丹麦技术大学，7(15):510，2008年。

Qi, X., Wang, J., Chen, Y., Shi, Y. 和 Zhang, L. Lips-former：将 Lipschitz 连续性引入视觉 Transformer。发表于第十一届国际学习表征会议，2023a。

Qi, X., Wang, J. 和 Zhang, L. 通过雅可比矩阵和Lipschitz常数理解深度学习的优化。*arXiv* 预印本 *arXiv:2306.09338*，2023年 b。

Qi, X., Chen, M., Xiao, W., Ye, J., He, Y., Li, C.-G. 和 Lin, Z. DNT：一种可通过动量随机梯度下降进行训练的深度归一化 Transformer。*arXiv* 预印本 *arXiv:2507.17501*，2025年 a。

Qi, X., He, Y., Ye, J., Li, C.-G., Zi, B., Dai, X., Zou, Q. 和 Xiao, R. 无需使用学习率预热，驯服Transformer。第十三届国际学习表征会议，2025年 b。

Qi, X., Ye, J., He, Y., Li, C.-G., Zi, B., Dai, X., Zou, Q. 和 Xiao, R. Stable-transformer：迈向稳定的Transformer训练，2025c。网址https://openreview.net/forum?id=lkRjnNW0gb。

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. 等。通过生成式预训练提升语言理解。2018年。

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. 等。语言模型是无监督的多任务学习者。*OpenAI* 博客，1(8):9，2019年。

Shazeer, N. GLU变体改进 Transformer。*arXiv* 预印本 *arXiv:2002.05202*，2020年。

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B. 和 Roformer, Y. L. 采用旋转位置嵌入的增强型 Transformer。，2021年。*DOI: https://doi. org/10.1016/j.neucom*，2023年。

Team, Q. 通义千问技术报告。*arXiv*预印本 *arXiv:2309.16609*，2023。Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi`ere, B., Goyal, N., Hambro, E., Azhar, F. 等人。Llama：开放且高效的基础语言模型。*arXiv*预印本 *arXiv:2302.13971*，2023a。Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. 等人。Llama 2：开放基础模型与微调聊天模型。*arXiv*预印本 *arXiv:2307.09288*，2023年 b。Ulyanov, D., Vedaldi, A. 与 Lempitsky, V. 实例归一化：快速风格化所缺失的要素。*arXiv*预印本 *arXiv:1607.08022*，2016。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J.,Jones, L., Gomez, A. N., Kaiser, Ł. 与 Polosukhin, I. 你所需要的只是注意力。神经信息处理系统进展，30，2017。

Wang, H., Ma, S., Dong, L., Huang, S., Zhang, D. 和 Wei, F. Deepnet：将 Transformer 扩展至 1,000 层。*arXiv* 预印本 *arXiv:2203.00555*，2022年。

Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F. 和 Chao, L. S. 用于机器翻译的深度 Transformer 模型学习。*arXiv* 预印本 *arXiv:1906.01787*，2019年。

Wu, Y. 和 He, K. 组归一化。载于欧洲计算机视觉会议论文集（*ECCV*），第 3–19 页，2018 年。

Xie, Z., Wei, Y., Cao, H., Zhao, C., Deng, C., Li, J., Dai, D., Gao, H., Chang, J., Zhao, L. 等。mhc：流形约束超连接。*arXiv* 预印本 *arXiv:2512.24880*，2025 年。

Zhang, B. 和 Sennrich, R. 均方根层归一化。神经信息处理系统进展，第 32 卷，2019年。

Zhang, Y., Chen, C., Li, Z., Ding, T., Wu, C., Kingma, D. P., Ye, Y., Luo, Z.-Q., 和 Sun, R. Adam-mini：用更少的学习率获益更多。*arXiv* 预印本 *arXiv:2406.16793*，2024年。

Zhu, J.、Chen, X.、He, K.、LeCun, Y. 和 Liu, Z.《无归一化的 Transformer》。载于计算机视觉与模式识别会议论文集，第 14901–14911 页，2025年。

正文和附录中出现的公式推导可参见以下资料（Golub & Van Loan，2013年；Horn &Johnson，2012年；Petersen etal.，2008年；Magnus & Neudecker，2019年；Nesterov，1983年；1998年；2013年；Boyd & Vandenberghe，2004年；Qi 等人，2025年 b；2023年 b；2025年 a）。

A. SimpleNorm 的 $\nabla_{\boldsymbol{x}}\ell$ 和 $\nabla^2_{\boldsymbol{x}}\ell$ 的导数

我们考虑映射

$$\boldsymbol{y} = \boldsymbol{\gamma} \odot \sqrt{d}\,\frac{\boldsymbol{W}\boldsymbol{x}}{\|\boldsymbol{W}\boldsymbol{x}\|_2},$$

其中，$\boldsymbol{\gamma} \in \mathbb{R}^d$ 和 $\odot$ 表示逐元素乘法，$\ell = \ell(\boldsymbol{y})$ 是一个标量损失。

和之前一样，我们定义一些中间变量，

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}, \qquad s = \|\boldsymbol{z}\|_2, \qquad \boldsymbol{u} = \frac{\boldsymbol{z}}{s}, \qquad \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top, \qquad \boldsymbol{D} = \mathrm{Diag}(\boldsymbol{\gamma}),$$

其中 $\boldsymbol{P}$，$\boldsymbol{D}$ 是对称的。

将归一化并缩放的输出定义为

$$\boldsymbol{y} = \sqrt{d}\,\boldsymbol{D}\,\boldsymbol{u},$$

并令

$$\boldsymbol{g}_{\boldsymbol{y}} := \nabla_{\boldsymbol{y}}\ell, \qquad \boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}} := \nabla^2_{\boldsymbol{y}\boldsymbol{y}}\ell.$$

将归一化映射 $\boldsymbol{u}$ 关于 $\boldsymbol{z}$ 的雅可比矩阵 $\boldsymbol{A}$定义为

$$\boldsymbol{A} := \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{z}} = \frac{1}{s}\,\boldsymbol{P},$$

因此

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} = \sqrt{d}\,\boldsymbol{D}\,\boldsymbol{A} = \frac{\sqrt{d}}{s}\,\boldsymbol{D}\,\boldsymbol{P}.$$

我们将反复对这条路径应用链式法则

$$\boldsymbol{W} \to \boldsymbol{z} \to \boldsymbol{u} \to \boldsymbol{y} \to \ell, \qquad \text{and similarly} \qquad \boldsymbol{\gamma} \to \boldsymbol{D} \to \boldsymbol{y} \to \ell.$$

(1) 一阶导数。由于 $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$，我们有

$$d\boldsymbol{z} = \boldsymbol{W}\,d\boldsymbol{x}.$$

由 $\boldsymbol{u} = \boldsymbol{z}/s$和 $\boldsymbol{A} = \partial \boldsymbol{u}/\partial \boldsymbol{z}$可得

$$d\boldsymbol{u} = \boldsymbol{A}\,d\boldsymbol{z} = \frac{1}{s}\,\boldsymbol{P}\,d\boldsymbol{z}.$$

因此，

$$d\boldsymbol{y} = \sqrt{d}\,\boldsymbol{D}\,d\boldsymbol{u} = \sqrt{d}\,\boldsymbol{D}\,\boldsymbol{A}\,\boldsymbol{W}\,d\boldsymbol{x},$$

所以雅可比矩阵为

$$\boldsymbol{J}^{\boldsymbol{y}}_{\boldsymbol{x}} := \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \sqrt{d}\,\boldsymbol{D}\,\boldsymbol{A}\,\boldsymbol{W} = \frac{\sqrt{d}}{s}\,\boldsymbol{D}\,\boldsymbol{P}\,\boldsymbol{W}.$$

应用链式法则可得

$$\nabla_{\boldsymbol{x}}\ell = \boldsymbol{J}^{\boldsymbol{y}}_{\boldsymbol{x}}{}^\top \boldsymbol{g}_{\boldsymbol{y}} = \sqrt{d}\,\boldsymbol{W}^\top \boldsymbol{A}^\top \boldsymbol{D}^\top \boldsymbol{g}_{\boldsymbol{y}}.$$

利用 $\boldsymbol{A}^\top = \boldsymbol{A}$ 和 $\boldsymbol{D}^\top = \boldsymbol{D}$，我们得到

$$\nabla_{\boldsymbol{x}}\ell = \sqrt{d}\,\boldsymbol{W}^\top \boldsymbol{A}\,\boldsymbol{D}\,\boldsymbol{g}_{\boldsymbol{y}} = \frac{\sqrt{d}}{s}\,\boldsymbol{W}^\top \boldsymbol{P}\,\boldsymbol{D}\,\boldsymbol{g}_{\boldsymbol{y}}. \tag{9}$$

(2) 二阶导数。对 $\nabla_{\boldsymbol{x}}\ell = \sqrt{d}\,\boldsymbol{W}^\top \boldsymbol{A}\,\boldsymbol{D}\,\boldsymbol{g_y}$ 求导可得

$$d(\nabla_{\boldsymbol{x}}\ell) = \sqrt{d}\,\boldsymbol{W}^\top\Big(\boldsymbol{A}\,\boldsymbol{D}\,d\boldsymbol{g_y} + d\boldsymbol{A}\,\boldsymbol{D}\,\boldsymbol{g_y}\Big).$$

这导出了标准分解

$$\nabla^2_{\boldsymbol{x}}\ell = \boldsymbol{J_x^y}^\top \boldsymbol{H_{yy}} \boldsymbol{J_x^y} + \boldsymbol{C},$$

其中两项的计算如下。

（A）线性（高斯–牛顿）项。

由于 $d\boldsymbol{g_y} = \boldsymbol{H_{yy}}\,d\boldsymbol{y}$，且 $d\boldsymbol{y} = \boldsymbol{J_x^y}\,d\boldsymbol{x}$，我们有

$$d\boldsymbol{g_y} = \boldsymbol{H_{yy}}\,\boldsymbol{J_x^y}\,d\boldsymbol{x}.$$

代入 $\sqrt{d}\,\boldsymbol{W}^\top\boldsymbol{A}\,\boldsymbol{D}\,d\boldsymbol{g_y}$ 得到

$$\sqrt{d}\,\boldsymbol{W}^\top\boldsymbol{A}\,\boldsymbol{D}\,d\boldsymbol{g_y} = \sqrt{d}\,\boldsymbol{W}^\top\boldsymbol{A}\,\boldsymbol{D}\,\boldsymbol{H_{yy}}\,\boldsymbol{J_x^y}\,d\boldsymbol{x} = \boldsymbol{J_x^y}^\top\boldsymbol{H_{yy}}\boldsymbol{J_x^y}\,d\boldsymbol{x}.$$

利用 $\boldsymbol{A} = \boldsymbol{P}/s$ 和 $\boldsymbol{J_x^y} = (\sqrt{d}/s)\boldsymbol{D}\boldsymbol{P}\boldsymbol{W}$，可得其显式形式

$$\boldsymbol{J_x^y}^\top\boldsymbol{H_{yy}}\boldsymbol{J_x^y} = \frac{d}{s^2}\,\boldsymbol{W}^\top\boldsymbol{P}\,\boldsymbol{D}\,\boldsymbol{H_{yy}}\,\boldsymbol{D}\,\boldsymbol{P}\,\boldsymbol{W}.$$

(B) 曲率（归一化）项。

回顾 $\boldsymbol{A} = \frac{1}{s}\boldsymbol{P}$。利用 $d\boldsymbol{u} = \boldsymbol{A}\,d\boldsymbol{z}$ 和 $ds = d\|\boldsymbol{z}\|_2 = \boldsymbol{u}^\top d\boldsymbol{z}$，可得

$$d\boldsymbol{A} = -\frac{1}{s^2}\Big((\boldsymbol{u}^\top d\boldsymbol{z})\boldsymbol{P} + \boldsymbol{P}\,d\boldsymbol{z}\,\boldsymbol{u}^\top + \boldsymbol{u}\,d\boldsymbol{z}^\top\boldsymbol{P}\Big).$$

右乘 $\boldsymbol{D}\,\boldsymbol{g_y}$，得到

$$d\boldsymbol{A}\,\boldsymbol{D}\,\boldsymbol{g_y} = -\frac{1}{s^2}\Big((\boldsymbol{u}^\top d\boldsymbol{z})\boldsymbol{P}\,\boldsymbol{D}\,\boldsymbol{g_y} + \boldsymbol{P}\,d\boldsymbol{z}\,\boldsymbol{u}^\top\boldsymbol{D}\,\boldsymbol{g_y} + \boldsymbol{u}\,d\boldsymbol{z}^\top\boldsymbol{P}\,\boldsymbol{D}\,\boldsymbol{g_y}\Big).$$

等价地，将标量收缩项移至左侧，可得 $d\boldsymbol{z}$ 中的线性映射：

$$d\boldsymbol{A}\,\boldsymbol{D}\,\boldsymbol{g_y} = -\frac{1}{s^2}\Big(\boldsymbol{P}\boldsymbol{D}\boldsymbol{g_y}\boldsymbol{u}^\top + \boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{g_y}\boldsymbol{P} + \boldsymbol{u}\boldsymbol{g_y}^\top\boldsymbol{D}\boldsymbol{P}\Big)\,d\boldsymbol{z}.$$

代入 $d\boldsymbol{z} = \boldsymbol{W}\,d\boldsymbol{x}$ 并左乘 $\sqrt{d}\,\boldsymbol{W}^\top$，得到曲率项

$$\boldsymbol{C} = -\frac{\sqrt{d}}{s^2}\,\boldsymbol{W}^\top\Big(\boldsymbol{P}\boldsymbol{D}\boldsymbol{g_y}\boldsymbol{u}^\top + \boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{g_y}\boldsymbol{P} + \boldsymbol{u}\boldsymbol{g_y}^\top\boldsymbol{D}\boldsymbol{P}\Big)\boldsymbol{W}.$$

（如有需要，可定义 $\boldsymbol{g}_1 := \boldsymbol{D}\boldsymbol{g_y}$ 以简化该表达式，但我们仍显式保留图中的符号。）

将线性项与曲率项合并后，关于 $\boldsymbol{x}$ 的海森矩阵为

$$\boldsymbol{H_{xx}} = \nabla^2_{\boldsymbol{x}}\ell = \frac{d}{s^2}\,\boldsymbol{W}^\top\boldsymbol{P}\,\boldsymbol{D}\,\boldsymbol{H_{yy}}\,\boldsymbol{D}\,\boldsymbol{P}\,\boldsymbol{W} \;-\; \frac{\sqrt{d}}{s^2}\,\boldsymbol{W}^\top\Big(\boldsymbol{P}\boldsymbol{D}\boldsymbol{g_y}\boldsymbol{u}^\top + \boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{g_y}\boldsymbol{P} + \boldsymbol{u}\boldsymbol{g_y}^\top\boldsymbol{D}\boldsymbol{P}\Big)\boldsymbol{W}, \tag{10}$$

其中 $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$，$s = \|\boldsymbol{z}\|_2$，$\boldsymbol{u} = \frac{\boldsymbol{z}}{s}$，$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top$，$\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{\gamma})$。

B. $\nabla_{\boldsymbol{W}}\ell$ 和 $\nabla^2_{\mathrm{vec}(\boldsymbol{W})}\ell$ 关于 $\boldsymbol{y}=\boldsymbol{W}\boldsymbol{x}$ 的导数

我们考虑线性映射

$$\boldsymbol{y}=\boldsymbol{W}\boldsymbol{x},$$

其中 $\boldsymbol{W}\in\mathbb{R}^{d\times m}$、$\boldsymbol{x}\in\mathbb{R}^m$ 和 $\boldsymbol{y}\in\mathbb{R}^d$。令 $\ell=\ell(\boldsymbol{y})$ 为一个取值为标量的损失函数。

我们将损失 l 关于 $\boldsymbol{y}$ 的一阶和二阶导数定义为

$$\boldsymbol{g}_{\boldsymbol{y}}\ :=\ \nabla_{\boldsymbol{y}}\ell(\boldsymbol{y})\in\mathbb{R}^d,\qquad \boldsymbol{H}_{\boldsymbol{yy}}\ :=\ \nabla^2_{\boldsymbol{y}}\ell(\boldsymbol{y})\in\mathbb{R}^{d\times d}.$$

(1) 关于 $\boldsymbol{W}$ 的一阶导数。由于 $\boldsymbol{y}$ 关于 $\boldsymbol{W}$ 呈线性关系，一阶变分满足

$$d\boldsymbol{y}=d\boldsymbol{W}\,\boldsymbol{x}.$$

应用链式法则，

$$d\ell=\boldsymbol{g}_{\boldsymbol{y}}^{\top}d\boldsymbol{y}=\boldsymbol{g}_{\boldsymbol{y}}^{\top}(d\boldsymbol{W}\,\boldsymbol{x})=\langle\boldsymbol{g}_{\boldsymbol{y}}\boldsymbol{x}^{\top},\,d\boldsymbol{W}\rangle_F.$$

在弗罗贝尼乌斯内积下匹配系数可得梯度

$$\nabla_{\boldsymbol{W}}\ell\ =\ \boldsymbol{g}_{\boldsymbol{y}}\,\boldsymbol{x}^{\top}. \tag{11}$$

(2) 关于 $\boldsymbol{W}$ 的二阶导数。为了表示二阶导数，将 $\boldsymbol{W}$ 向量化较为方便。利用矩阵-向量乘积的标准恒等式，

$$\boldsymbol{y}=(\boldsymbol{x}^{\top}\otimes\boldsymbol{I}_d)\ \mathrm{vec}(\boldsymbol{W}),$$

这使得 $\boldsymbol{y}$ 对 vec($\boldsymbol{W}$) 的依赖关系变得明确。

关于 vec($\boldsymbol{W}$) 的雅可比矩阵。由上述表达式，$\boldsymbol{y}$ 关于 vec($\boldsymbol{W}$) 的雅可比矩阵为

$$\boldsymbol{J}^{\boldsymbol{y}}_{\mathrm{vec}(\boldsymbol{W})}\ \triangleq\ \frac{\partial\boldsymbol{y}}{\partial\,\mathrm{vec}(\boldsymbol{W})}=\boldsymbol{x}^{\top}\otimes\boldsymbol{I}_d\ \in\ \mathbb{R}^{d\times(md)}.$$

关于 vec($\boldsymbol{W}$) 的海森矩阵。由于映射 $\boldsymbol{W}\mapsto\boldsymbol{y}$ 是线性的，它不贡献二阶项。因此，根据二阶链式法则，损失函数关于 vec($\boldsymbol{W}$) 的海森矩阵为

$$\nabla^2_{\mathrm{vec}(\boldsymbol{W})}\ell=\boldsymbol{J}^{\boldsymbol{y}}_{\mathrm{vec}(\boldsymbol{W})}{}^{\top}\,\boldsymbol{H}_{\boldsymbol{yy}}\,\boldsymbol{J}^{\boldsymbol{y}}_{\mathrm{vec}(\boldsymbol{W})}.$$

代入显式雅可比矩阵可得

$$\nabla^2_{\mathrm{vec}(\boldsymbol{W})}\ell=(\boldsymbol{x}\otimes\boldsymbol{I}_d)\,\boldsymbol{H}_{\boldsymbol{yy}}\,(\boldsymbol{x}^{\top}\otimes\boldsymbol{I}_d).$$

克罗内克积简化。利用克罗内克积的标准恒等式，海森矩阵可简化为

$$\nabla^2_{\mathrm{vec}(\boldsymbol{W})}\ell=(\boldsymbol{x}\boldsymbol{x}^{\top})\otimes\boldsymbol{H}_{\boldsymbol{yy}}. \tag{12}$$

该结果表明，相对于权重矩阵的海森矩阵可分解为依赖输入的项 $\boldsymbol{x}\boldsymbol{x}^{\top}$ 与输出空间曲率项 $\boldsymbol{H}_{\boldsymbol{yy}}$ 的克罗内克积。等价地，将其视为一个分块矩阵，其中包含 $m\times m$ 个大小为 $d\times d$ 的分块时，第（$i,\ j$）个分块为

$$\left[\nabla^2_{\mathrm{vec}(\boldsymbol{W})}\ell\right]_{(i,j)\text{ block}}=x_ix_j\,\boldsymbol{H}_{\boldsymbol{yy}},$$

其中，x_i 表示 $\boldsymbol{x}$ 的第 i 个条目。

C. $\nabla_{\boldsymbol{\gamma}}\ell$、 $\nabla^2_{\boldsymbol{\gamma}}\ell$、 $\nabla_{\boldsymbol{W}}\ell$ 和 $\nabla^2_{\mathrm{vec}(\boldsymbol{W})}\ell$ 的导数，其中 $\boldsymbol{y}=\boldsymbol{\gamma}\odot\sqrt{d}\,\frac{\boldsymbol{W}\boldsymbol{x}}{\|\boldsymbol{W}\boldsymbol{x}\|_2}$。

我们考虑如下映射

$$\boldsymbol{y}=\boldsymbol{\gamma}\odot\sqrt{d}\,\frac{\boldsymbol{W}\boldsymbol{x}}{\|\boldsymbol{W}\boldsymbol{x}\|_2},$$

其中 $\odot$ 表示逐元素乘法， $\ell=\ell(\boldsymbol{y})$ 是一个标量损失。定义中间变量

$$\boldsymbol{z}=\boldsymbol{W}\boldsymbol{x},\qquad s=\|\boldsymbol{z}\|_2,\qquad \boldsymbol{u}=\frac{\boldsymbol{z}}{s},\qquad \boldsymbol{P}=\boldsymbol{I}-\boldsymbol{u}\boldsymbol{u}^\top,\qquad \boldsymbol{D}=\mathrm{Diag}(\boldsymbol{\gamma}).$$

我们有

$$\boldsymbol{y}=\sqrt{d}\,\boldsymbol{D}\boldsymbol{u},\qquad \ell=\ell(\boldsymbol{y}),\qquad \boldsymbol{g}_{\boldsymbol{y}}:=\nabla_{\boldsymbol{y}}\ell,\qquad \boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}}:=\nabla^2_{\boldsymbol{y}\boldsymbol{y}}\ell.$$

回顾两个有用的雅可比矩阵，

$$\frac{\partial\boldsymbol{u}}{\partial\boldsymbol{z}}=\boldsymbol{A}=\frac{1}{s}\boldsymbol{P},\qquad \frac{\partial\boldsymbol{y}}{\partial\boldsymbol{z}}=\sqrt{d}\,\boldsymbol{D}\boldsymbol{A}=\frac{\sqrt{d}}{s}\boldsymbol{D}\boldsymbol{P}.$$

我们将反复　　通过路径 $\boldsymbol{W}\to\boldsymbol{z}\to\boldsymbol{u}\to\boldsymbol{y}\to\ell$ 应用链式法则， $\boldsymbol{\gamma}\to\boldsymbol{D}\to\boldsymbol{y}\to$ 亦然　　ℓ.

C.1. 第一部分：关于 $\boldsymbol{\gamma}$ 的导数

(1) 一阶导数。 由于 $\boldsymbol{y}_i=\sqrt{d}\,\boldsymbol{\gamma}_i\boldsymbol{u}_i$， $\boldsymbol{y}$ 关于 $\boldsymbol{\gamma}$ 的雅可比矩阵是对角矩阵：

$$\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{\gamma}}=\sqrt{d}\,\mathrm{Diag}(\boldsymbol{u}).$$

应用链式法则 $\boldsymbol{g}_{\boldsymbol{\gamma}}=\left(\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{\gamma}}\right)^\top\boldsymbol{g}_{\boldsymbol{y}}$，可得

$$\boldsymbol{g}_{\boldsymbol{\gamma}}=\nabla_{\boldsymbol{\gamma}}\ell=\sqrt{d}\,\mathrm{Diag}(\boldsymbol{u})\,\boldsymbol{g}_{\boldsymbol{y}}=\sqrt{d}\,(\boldsymbol{u}\odot\boldsymbol{g}_{\boldsymbol{y}}).\tag{13}$$

(2) 二阶导数。 因为 $\boldsymbol{y}$ 关于 $\boldsymbol{\gamma}$、 $\frac{\partial^2\boldsymbol{y}}{\partial\boldsymbol{\gamma}^2}=0$ 是线性映射，因此海森矩阵完全来自 $\boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}}$：

$$\boldsymbol{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}=\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell=\Big(\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{\gamma}}\Big)^\top\boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}}\Big(\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{\gamma}}\Big)=d\;\mathrm{Diag}(\boldsymbol{u})\,\boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}}\,\mathrm{Diag}(\boldsymbol{u}).\tag{14}$$

C.2. 第二部分：关于 $\boldsymbol{W}$ 的导数

(1) 一阶导数。步骤 1（$\boldsymbol{y}\to\boldsymbol{z}$）：对 $\boldsymbol{z}$ 求梯度

根据链式法则，

$$\boldsymbol{g}_{\boldsymbol{z}}=\Big(\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{z}}\Big)^\top\boldsymbol{g}_{\boldsymbol{y}}=\Big(\frac{\sqrt{d}}{s}\boldsymbol{D}\boldsymbol{P}\Big)^\top\boldsymbol{g}_{\boldsymbol{y}}.$$

利用 $\boldsymbol{P}^\top=\boldsymbol{P}$ 和 $\boldsymbol{D}^\top=\boldsymbol{D}$，可简化为

$$\boldsymbol{g}_{\boldsymbol{z}}=\frac{\sqrt{d}}{s}\,\boldsymbol{P}\,\boldsymbol{D}\,\boldsymbol{g}_{\boldsymbol{y}}.$$

步骤 2（$\boldsymbol{z}=\boldsymbol{W}\boldsymbol{x}\to\boldsymbol{W}$）：对 $\boldsymbol{W}$ 求梯度

我们将微分写作 $d\boldsymbol{z}=d\boldsymbol{W}\,\boldsymbol{x}$，因此

$$d\ell=\boldsymbol{g}_{\boldsymbol{z}}^\top d\boldsymbol{z}=\boldsymbol{g}_{\boldsymbol{z}}^\top(d\boldsymbol{W}\,\boldsymbol{x})=\langle\boldsymbol{g}_{\boldsymbol{z}}\boldsymbol{x}^\top,\,d\boldsymbol{W}\rangle_F.$$

在弗罗贝尼乌斯内积下匹配系数，得到

$$\boldsymbol{g}_{\boldsymbol{W}}=\nabla_{\boldsymbol{W}}\ell=\boldsymbol{g}_{\boldsymbol{z}}\boldsymbol{x}^\top=\frac{\sqrt{d}}{s}\,(\boldsymbol{P}\boldsymbol{D}\boldsymbol{g}_{\boldsymbol{y}})\,\boldsymbol{x}^\top.\tag{15}$$

如果需要向量化映射，

$$\mathrm{vec}(\boldsymbol{z})=\mathrm{vec}(\boldsymbol{W}\boldsymbol{x})=(\boldsymbol{x}^\top\otimes\boldsymbol{I})\,\mathrm{vec}(\boldsymbol{W}).$$

(2) 二阶导数。我们首先推导 $\boldsymbol{H_{zz}} := \nabla^2_{\boldsymbol{zz}}\ell$，然后将其提升为 $\mathrm{vec}(\boldsymbol{W})$。

步骤 1：关于 $\boldsymbol{z}$ 的海森矩阵

二阶链式法则给出

$$\boldsymbol{H_{zz}} = \Big(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}}\Big)^\top \boldsymbol{H_{yy}} \Big(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}}\Big) + \sum_{k=1}^{d} (\boldsymbol{g_y})_k \, \nabla^2_{\boldsymbol{zz}} y_k.$$

第一项是高斯-牛顿部分：

$$\Big(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}}\Big)^\top \boldsymbol{H_{yy}} \Big(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}}\Big) = \frac{d}{s^2}\, \boldsymbol{P}\, \boldsymbol{D}\, \boldsymbol{H_{yy}}\, \boldsymbol{D}\, \boldsymbol{P}.$$

对于第二项，我们使用归一化海森矩阵的双线性形式：

$$\nabla^2 \boldsymbol{u}(\boldsymbol{z})[\boldsymbol{a}, \boldsymbol{b}] = -\frac{1}{s^2}\Big((\boldsymbol{u}^\top \boldsymbol{a})\, \boldsymbol{P}\boldsymbol{b} + (\boldsymbol{u}^\top \boldsymbol{b})\, \boldsymbol{P}\boldsymbol{a} + (\boldsymbol{a}^\top \boldsymbol{P}\boldsymbol{b})\, \boldsymbol{u} \Big).$$

由于$\boldsymbol{y} = \sqrt{d}\bar{\boldsymbol{D}}\boldsymbol{u}$,

$$\sum_{k=1}^{d} (\boldsymbol{g_y})_k \, \nabla^2 y_k[\boldsymbol{a}, \boldsymbol{b}] = \boldsymbol{g_y}^\top \nabla^2 \boldsymbol{y}[\boldsymbol{a}, \boldsymbol{b}] = \sqrt{d}\, (\boldsymbol{D}\boldsymbol{g_y})^\top \nabla^2 \boldsymbol{u}[\boldsymbol{a}, \boldsymbol{b}].$$

代入可得

$$\sum_{k=1}^{d} (\boldsymbol{g_y})_k \, \nabla^2 y_k[\boldsymbol{a}, \boldsymbol{b}] = -\frac{\sqrt{d}}{s^2}\Big((\boldsymbol{u}^\top \boldsymbol{a})(\boldsymbol{D}\boldsymbol{g_y})^\top \boldsymbol{P}\boldsymbol{b} + (\boldsymbol{u}^\top \boldsymbol{b})(\boldsymbol{D}\boldsymbol{g_y})^\top \boldsymbol{P}\boldsymbol{a} + (\boldsymbol{u}^\top \boldsymbol{D}\boldsymbol{g_y})\, \boldsymbol{a}^\top \boldsymbol{P}\boldsymbol{b} \Big).$$

等价地，相应的矩阵形式为

$$\sum_{k=1}^{d} (\boldsymbol{g_y})_k \, \nabla^2_{\boldsymbol{zz}} y_k = -\frac{\sqrt{d}}{s^2}\Big(\boldsymbol{P}\boldsymbol{D}\boldsymbol{g_y}\boldsymbol{u}^\top + \boldsymbol{u}^\top \boldsymbol{D}\boldsymbol{g_y}\boldsymbol{P} + \boldsymbol{u}\boldsymbol{g_y}^\top \boldsymbol{D}\boldsymbol{P} \Big).$$

将两项合并，

$$\boldsymbol{H_{zz}} = \frac{d}{s^2}\, \boldsymbol{P}\, \boldsymbol{D}\, \boldsymbol{H_{yy}}\, \boldsymbol{D}\, \boldsymbol{P} - \frac{\sqrt{d}}{s^2}\Big(\boldsymbol{P}\boldsymbol{D}\boldsymbol{g_y}\boldsymbol{u}^\top + \boldsymbol{u}^\top \boldsymbol{D}\boldsymbol{g_y}\boldsymbol{P} + \boldsymbol{u}\boldsymbol{g_y}^\top \boldsymbol{D}\boldsymbol{P} \Big), \qquad s = \|\boldsymbol{W}\boldsymbol{x}\|_2.$$

使用 $\mathrm{vec}(\boldsymbol{z}) = (\boldsymbol{x}^\top \otimes \boldsymbol{I})\, \mathrm{vec}(\boldsymbol{W})$,

$$\begin{aligned} \boldsymbol{H}_{\mathrm{vec}(\boldsymbol{W})\,\mathrm{vec}(\boldsymbol{W})} &= (\boldsymbol{x}\boldsymbol{x}^\top) \otimes \boldsymbol{H_{zz}} \\ &= (\boldsymbol{x}\boldsymbol{x}^\top) \otimes \left[\frac{d}{s^2}\, \boldsymbol{P}\, \boldsymbol{D}\, \boldsymbol{H_{yy}}\, \boldsymbol{D}\, \boldsymbol{P} - \frac{\sqrt{d}}{s^2}\Big(\boldsymbol{P}\boldsymbol{D}\boldsymbol{g_y}\boldsymbol{u}^\top + \boldsymbol{u}^\top \boldsymbol{D}\boldsymbol{g_y}\boldsymbol{P} + \boldsymbol{u}\boldsymbol{g_y}^\top \boldsymbol{D}\boldsymbol{P} \Big) \right]. \end{aligned} \tag{16}$$

D. 证明定理 4.1

证明。 回顾一下 SimpleNorm：

$$\boldsymbol{y} = \boldsymbol{\gamma} \odot \sqrt{d}\,\frac{\boldsymbol{W}\boldsymbol{x}}{\|\boldsymbol{W}\boldsymbol{x}\|_2}, \qquad \boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}, \qquad s := \|\boldsymbol{z}\|_2, \qquad \boldsymbol{u} := \frac{\boldsymbol{z}}{s}, \qquad \boldsymbol{P} := \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top.$$

假设 $\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{\gamma}) = \boldsymbol{I}$。令 $\ell = \ell(\boldsymbol{y})$并定义

$$\boldsymbol{g_y} := \nabla_{\boldsymbol{y}}\ell(\boldsymbol{y}) \in \mathbb{R}^d, \qquad \boldsymbol{H_{yy}} := \nabla^2_{\boldsymbol{yy}}\ell(\boldsymbol{y}) \in \mathbb{R}^{d\times d}.$$

根据链式法则，我们有如下分解：

$$\boldsymbol{H_{xx}} = \nabla^2_{\boldsymbol{x}}\ell = \underbrace{\boldsymbol{J_x^y}^\top \boldsymbol{H_{yy}}\, \boldsymbol{J_x^y}}_{\text{Gauss–Newton term}} + \underbrace{\boldsymbol{C}}_{\text{curvature term}} \tag{17}$$

此处雅可比矩阵为

$$\boldsymbol{J_x^y} = \frac{\sqrt{d}}{s}\,\boldsymbol{D}\boldsymbol{P}\boldsymbol{W} = \frac{\sqrt{d}}{s}\,\boldsymbol{P}\boldsymbol{W},$$

因此高斯–牛顿项为

$$\boldsymbol{L} = (\boldsymbol{J_x^y})^\top \boldsymbol{H_{yy}} \boldsymbol{J_x^y} = \frac{d}{s^2}\,\boldsymbol{W}^\top \boldsymbol{P}\,\boldsymbol{H_{yy}}\,\boldsymbol{P}\,\boldsymbol{W}. \tag{18}$$

曲率项（在条件 $\boldsymbol{D} = \boldsymbol{I}$下）可以写成

$$\boldsymbol{C} = -\frac{\sqrt{d}}{s^2}\,\boldsymbol{W}^\top\Big((\boldsymbol{P}\boldsymbol{g_y})\boldsymbol{u}^\top + (\boldsymbol{u}^\top \boldsymbol{g_y})\,\boldsymbol{P} + \boldsymbol{u}(\boldsymbol{g_y}^\top \boldsymbol{P})\Big)\boldsymbol{W}, \tag{19}$$

其中 $\boldsymbol{u}^\top \boldsymbol{g_y}$ 是一个标量。

对高斯–牛顿项进行界定。定义

$$\kappa := \frac{\sqrt{d}}{\|\boldsymbol{W}\boldsymbol{x}\|_2}\,\|\boldsymbol{W}\|_2$$

我们证明，在高有效秩和典型输入下， $\kappa = \Theta(1)$。令 $\boldsymbol{x} = \sqrt{d}\,\xi$ ，其中 ξ 是各向同性的（例如，在球面上均匀分布或次高斯）。则

$$\mathbb{E}\|\boldsymbol{W}\boldsymbol{x}\|_2^2 = d\,\mathbb{E}\,\xi^\top \boldsymbol{W}^\top \boldsymbol{W}\xi = \|\boldsymbol{W}\|_F^2,$$

并且根据标准的二次型集中性，

$$\|\boldsymbol{W}\boldsymbol{x}\|_2^2 \asymp \|\boldsymbol{W}\|_F^2 \quad \Longrightarrow \quad \|\boldsymbol{W}\boldsymbol{x}\|_2 \asymp \|\boldsymbol{W}\|_F \qquad \text{with high probability.}$$

Thus

$$\kappa = \frac{\sqrt{d}}{\|\boldsymbol{W}\boldsymbol{x}\|_2}\,\|\boldsymbol{W}\|_2 \asymp \sqrt{d}\,\frac{\|\boldsymbol{W}\|_2}{\|\boldsymbol{W}\|_F} = \sqrt{\frac{d}{r_{\text{eff}}(\boldsymbol{W})}}, \qquad r_{\text{eff}}(\boldsymbol{W}) := \frac{\|\boldsymbol{W}\|_F^2}{\|\boldsymbol{W}\|_2^2}.$$

若 $r_{\text{eff}}(\boldsymbol{W}) \asymp cd$，则以高概率有 $\kappa \asymp 1/\sqrt{c} = \Theta(1)$。

此外，定义

$$\tau := \frac{\big\|\widetilde{\boldsymbol{W}}^\top \boldsymbol{P}\,\boldsymbol{H_{yy}}\,\boldsymbol{P}\,\widetilde{\boldsymbol{W}}\big\|_2}{\|\boldsymbol{H_{yy}}\|_2} \in [0,1], \qquad \widetilde{\boldsymbol{W}} := \frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_2}.$$

在定理的“非病态对齐”假设下（即， $\boldsymbol{H_{yy}}$ 的主导谱模态不会因投影掉 $\mathrm{span}(\boldsymbol{u})$ 而被消除，并且由具有高有效秩的 $\boldsymbol{W}$ 的值域表示），我们有 $\tau = \Theta(1)$。

现在可得

$$\boldsymbol{L} = \frac{d}{\|\boldsymbol{W}\boldsymbol{x}\|_2^2}\,\boldsymbol{W}^\top \boldsymbol{P}\,\boldsymbol{H_{yy}}\,\boldsymbol{P}\,\boldsymbol{W} = \frac{\kappa^2}{\|\boldsymbol{W}\|_2^2}\,\boldsymbol{W}^\top \boldsymbol{P}\,\boldsymbol{H_{yy}}\,\boldsymbol{P}\,\boldsymbol{W} = \kappa^2\,\widetilde{\boldsymbol{W}}^\top \boldsymbol{P}\,\boldsymbol{H_{yy}}\,\boldsymbol{P}\,\widetilde{\boldsymbol{W}}$$

and

$$\|\boldsymbol{L}\|_2 = \tau\,\kappa^2\,\|\boldsymbol{H_{yy}}\|_2 \qquad \text{where } \tau = \Theta(1) \text{ and } \kappa = \Theta(1) \quad \text{[w.h.p]} \tag{20}$$

曲率项的界。由公式 19 和次乘性可得，

$$\|\boldsymbol{C}\|_2 \le \frac{\sqrt{d}}{s^2}\,\|\boldsymbol{W}\|_2^2 \left\|(\boldsymbol{P}\boldsymbol{g}_{\boldsymbol{y}})\boldsymbol{u}^\top + (\boldsymbol{u}^\top \boldsymbol{g}_{\boldsymbol{y}})\,\boldsymbol{P} + \boldsymbol{u}(\boldsymbol{g}_{\boldsymbol{y}}^\top \boldsymbol{P})\right\|_2.$$

现在使用 $\|\boldsymbol{u}\|_2 = 1$、$\|\boldsymbol{P}\|_2 = 1$ 和 $\|\boldsymbol{P}\boldsymbol{g}_{\boldsymbol{y}}\|_2 \le \|\boldsymbol{g}_{\boldsymbol{y}}\|_2$:

$$\|(\boldsymbol{P}\boldsymbol{g}_{\boldsymbol{y}})\boldsymbol{u}^\top\|_2 = \|\boldsymbol{P}\boldsymbol{g}_{\boldsymbol{y}}\|_2\|\boldsymbol{u}\|_2 \le \|\boldsymbol{g}_{\boldsymbol{y}}\|_2,$$

$$\|(\boldsymbol{u}^\top \boldsymbol{g}_{\boldsymbol{y}})\,\boldsymbol{P}\|_2 = |\boldsymbol{u}^\top \boldsymbol{g}_{\boldsymbol{y}}|\,\|\boldsymbol{P}\|_2 \le \|\boldsymbol{g}_{\boldsymbol{y}}\|_2,$$

$$\|\boldsymbol{u}(\boldsymbol{g}_{\boldsymbol{y}}^\top \boldsymbol{P})\|_2 = \|\boldsymbol{u}\|_2\,\|\boldsymbol{P}\boldsymbol{g}_{\boldsymbol{y}}\|_2 \le \|\boldsymbol{g}_{\boldsymbol{y}}\|_2.$$

根据三角不等式，中间的范数至多为 $3\|\boldsymbol{g}_{\boldsymbol{y}}\|_2$，因此

$$\|\boldsymbol{C}\|_2 \le \frac{3\sqrt{d}}{\|\boldsymbol{W}\boldsymbol{x}\|_2^2}\,\|\boldsymbol{W}\|_2^2\,\|\boldsymbol{g}_{\boldsymbol{y}}\|_2 = \frac{3\kappa^2}{\sqrt{d}}\,\|\boldsymbol{g}_{\boldsymbol{y}}\|_2. \tag{21}$$

$\boldsymbol{L}$ $\boldsymbol{C}$。结合（20）和（21），

$$\frac{\|\boldsymbol{C}\|_2}{\|\boldsymbol{L}\|_2} \le \frac{3}{\tau\sqrt{d}}\,\frac{\|\boldsymbol{g}_{\boldsymbol{y}}\|_2}{\|\boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}}\|_2}.$$

假设 $\tau = \Theta(1)$（非病态对齐）和 $\|\boldsymbol{g}_{\boldsymbol{y}}\|_2/\|\boldsymbol{H}_{\boldsymbol{y}\boldsymbol{y}}\|_2 = O(1)$（梯度与曲率之比有界，如定理假设所述），我们得到

$$\frac{\|\boldsymbol{C}\|_2}{\|\boldsymbol{L}\|_2} = O(d^{-1/2}),$$

因此，在高维下，$\|\boldsymbol{L}\|_2 \gg \|\boldsymbol{C}\|_2$ 以高概率成立。因此，在典型的非病态情形下，高斯–牛顿项占主导地位于 $\boldsymbol{H}_{\boldsymbol{x}\boldsymbol{x}}$。

至此，定理 4.1的证明完成。 □

E. 定理 4.2 的证明

证明。假设 $\ell = \ell(\boldsymbol{y})$ 二阶可微。记

$$\boldsymbol{g_y} := \nabla_{\boldsymbol{y}} \ell, \qquad \boldsymbol{H_{yy}} := \nabla^2_{\boldsymbol{yy}} \ell,$$

我们有以下两个映射：

线性映射：$\boldsymbol{y}_1 = \boldsymbol{W}_1 \boldsymbol{x}, \qquad \boldsymbol{H}^{\text{lin}}_{\boldsymbol{xx}} = \boldsymbol{W}_1^\top \boldsymbol{H}_{\boldsymbol{y}_1 \boldsymbol{y}_1} \boldsymbol{W}_1$。

SimpleNorm：$\boldsymbol{y}_2 = \boldsymbol{D} \frac{\sqrt{d}\boldsymbol{W}_2 \boldsymbol{x}}{\|\boldsymbol{W}_2 \boldsymbol{x}\|_2}$，其中 $\boldsymbol{D} = \text{Diag}(\gamma)$，并定义

$$\boldsymbol{z} = \boldsymbol{W}_2 \boldsymbol{x}, \quad s = \|\boldsymbol{z}\|_2, \quad \boldsymbol{u} = \boldsymbol{z}/s, \quad \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top.$$

关于 $\boldsymbol{x}$ 的海森矩阵具有标准分解

$$\boldsymbol{H}^{\text{sn}}_{\boldsymbol{xx}} \;=\; \boldsymbol{L} + \boldsymbol{C}, \qquad \boldsymbol{L} = \boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}_2 \top} \boldsymbol{H}_{\boldsymbol{y}_2 \boldsymbol{y}_2} \boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}_2},$$

其中，$\boldsymbol{J}_{\boldsymbol{x}}^{\boldsymbol{y}_2}$是 $\boldsymbol{y}_2$ 关于 $\boldsymbol{x}$的雅可比矩阵，$\boldsymbol{C}$ 是由归一化引起的曲率项。

假设 定理 4.1 中陈述的高维条件$\|\boldsymbol{x}\|_2 = \sqrt{d}$、$\boldsymbol{D} = \boldsymbol{I}$、$\|\boldsymbol{P}\|_2 = 1$、$\boldsymbol{W}_2$ 具有较高的有效秩，不存在病态对齐，且 $\|\boldsymbol{g_y}\|_2 / \|\boldsymbol{H_{yy}}\|_2 = O(1)$）成立，定理 4.1 表明高斯-牛顿项占主导地位，超过曲率项，即 $\|\boldsymbol{L}\|_2 \gg \|\boldsymbol{C}\|_2$。因此，我们得到近似 $\|\boldsymbol{H}^{\text{sn}}_{\boldsymbol{xx}}\|_2 \approx \|\boldsymbol{L}\|_2$。

现在我们给出一个积分结果：$\boldsymbol{H}^{\text{lin}}_{\boldsymbol{xx}}$ 的谱范数与权重矩阵$\boldsymbol{W}$的谱范数成正比，而 $\boldsymbol{H}^{\text{sn}}_{\boldsymbol{xx}}$ 的谱范数与权重矩阵的谱范数无关。下文中，我们假设$\boldsymbol{W} = \boldsymbol{W}_1 = \boldsymbol{W}_2$ 和$\boldsymbol{H}_{\boldsymbol{y}_1 \boldsymbol{y}_1} = \boldsymbol{H}_{\boldsymbol{y}_2 \boldsymbol{y}_2} := \boldsymbol{H}_{\boldsymbol{yy}}$。

定义 $\widetilde{\boldsymbol{W}} := \frac{\boldsymbol{W}}{\alpha}$，其中 $\alpha := \|\boldsymbol{W}\|_2$。因此，$\|\widetilde{\boldsymbol{W}}\|_2 = 1$和 $\boldsymbol{W} = \alpha \widetilde{\boldsymbol{W}}$。现在，将 $\boldsymbol{H}^{\text{lin}}_{\boldsymbol{xx}}$和 $\boldsymbol{H}^{\text{sn}}_{\boldsymbol{xx}}$视为 α的函数，可得：

$$\|\boldsymbol{H}^{\text{lin}}_{\boldsymbol{xx}}(\alpha)\|_2 = \|\boldsymbol{W}_1^\top \boldsymbol{H}_{\boldsymbol{y}_1 \boldsymbol{y}_1} \boldsymbol{W}_1\|_2 = \alpha^2 \|\widetilde{\boldsymbol{W}}^\top \boldsymbol{H}_{\boldsymbol{yy}} \widetilde{\boldsymbol{W}}\|_2 \tag{22}$$

and

$$\|\boldsymbol{H}^{\text{sn}}_{\boldsymbol{xx}}(\alpha)\|_2 \;\approx\; \|\boldsymbol{L}(\alpha)\|_2 = \frac{d}{\|\boldsymbol{W}\boldsymbol{x}\|_2^2} \|\boldsymbol{W}^\top \boldsymbol{P}\, \boldsymbol{D}\, \boldsymbol{H_{yy}}\, \boldsymbol{D}\, \boldsymbol{P}\, \boldsymbol{W}\|_2 = \frac{d}{\|\widetilde{\boldsymbol{W}}\boldsymbol{x}\|_2^2} \|\widetilde{\boldsymbol{W}}^\top \boldsymbol{P}\, \boldsymbol{H_{yy}}\, \boldsymbol{P}\, \widetilde{\boldsymbol{W}}\|_2 \tag{23}$$

注意，$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top$ 与 α无关：

$$\boldsymbol{u}(\alpha) = \frac{\boldsymbol{W}\boldsymbol{x}}{\|\boldsymbol{W}\boldsymbol{x}\|_2} = \frac{\alpha \widetilde{\boldsymbol{W}}\boldsymbol{x}}{\|\alpha \widetilde{\boldsymbol{W}}\boldsymbol{x}\|_2} = \frac{\widetilde{\boldsymbol{W}}\boldsymbol{x}}{\|\widetilde{\boldsymbol{W}}\boldsymbol{x}\|_2}, \qquad \Longrightarrow \qquad \boldsymbol{P}(\alpha) = \boldsymbol{I} - \boldsymbol{u}(\alpha)\boldsymbol{u}(\alpha)^\top = \boldsymbol{P}(1).$$

因此，$\|\boldsymbol{H}^{\text{lin}}_{\boldsymbol{xx}}(\alpha)\|_2$ 关于 α 呈二次依赖，而 $\|\boldsymbol{H}^{\text{sn}}_{\boldsymbol{xx}}(\alpha)\|_2$ 在 α中具有 尺度不变性。

接下来，我们比较两个海森矩阵的大小。首先，回顾 κ 在 定理 4.1中给出的定义 ：

$$\kappa \;:=\; \frac{\sqrt{d}}{\|\boldsymbol{W}\boldsymbol{x}\|_2} \|\boldsymbol{W}\|_2 = \frac{\sqrt{d}}{\alpha \|\widetilde{\boldsymbol{W}}\boldsymbol{x}\|_2} \alpha \|\widetilde{\boldsymbol{W}}\|_2 = \frac{\sqrt{d}}{\|\widetilde{\boldsymbol{W}}\boldsymbol{x}\|_2}$$

根据定义，$\kappa^2 = \frac{d}{\|\widetilde{\boldsymbol{W}}\boldsymbol{x}\|_2^2}$，并且定理 4.1 意味着 $\kappa^2 = \Theta(1)$ 以高概率

此外，由于$\|\boldsymbol{W}\|_2 = \|\boldsymbol{P}\|_2 = 1$,

$$\|\boldsymbol{H}^{\text{sn}}_{\boldsymbol{xx}}(\alpha)\|_2 \approx \|\boldsymbol{L}\|_2 = \kappa^2 \|\widetilde{\boldsymbol{W}}^\top \boldsymbol{P}\, \boldsymbol{H_{yy}}\, \boldsymbol{P}\, \widetilde{\boldsymbol{W}}\|_2 \le \kappa^2 \|\boldsymbol{H_{yy}}\|_2 \tag{24}$$

对于线性模块，存在一个常数 $c_{\text{lin}} > 0$ 使得

$$\|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{lin}}(\alpha)\|_2 = \alpha^2 \|\widetilde{\boldsymbol{W}}^\top \boldsymbol{H}_{\boldsymbol{yy}} \widetilde{\boldsymbol{W}}\|_2 \geq \alpha^2 c_{\mathrm{lin}} \|\boldsymbol{H}_{\boldsymbol{yy}}\|_2.$$

~我们假设 Range($\boldsymbol{W}$) 不会与 $\boldsymbol{H}_{\boldsymbol{yy}}$ 的主特征子空间发生对抗性对齐；特别地，二者的重叠度为常数量级，因此可见性常数满足 $c_{\mathrm{lin}} = \Theta(1)$，且不会随 d 衰减。

综合上述所有结果，两个海森矩阵之间的最终比率满足：

$$\frac{\|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{lin}}(\alpha)\|_2}{\|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{sn}}(\alpha)\|_2} \geq \frac{\alpha^2 \|\widetilde{\boldsymbol{W}}^\top \boldsymbol{H}_{\boldsymbol{yy}} \widetilde{\boldsymbol{W}}\|_2}{\kappa^2 \|\boldsymbol{H}_{\boldsymbol{yy}}\|_2} \geq \alpha^2 \frac{c_{\mathrm{lin}}}{\kappa^2}$$

从经验上看，$\alpha = \|\boldsymbol{W}\|_2$ 在训练过程中通常会增长到几十或几百，因此 $\alpha^2 c_{\mathrm{lin}} \gg \kappa^2$ （回顾一下，$\kappa = \Theta$以高概率成立）。因此：

$$\|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{lin}}\|_2 \gg \|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{sn}}\|_2 \qquad \text{(with high probability)}.$$

总而言之，在非病态情况下，$\|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{lin}}\|_2$ 的梯度利普希茨常数远大于 $\|\boldsymbol{H}_{\boldsymbol{xx}}^{\mathrm{sn}}\|_2$的梯度利普希茨常数。

这就完成了 定理 4.2的证明。

□

F. 详细实验设置

我们的**SimpleGPT** 模型基于 Llama2、Llama3 和 nanoGPT（GPT-2 的一种实现）构建。我们对除嵌入层和分类层之外的所有 Transformer 块应用 SimpleNorm 算子。我们的实现基于 Adam-mini [1] (Zhang 等人，2024年）和 nanoGPT [2] (Karpathy，2022）。下面简要介绍每种骨干网络的主要架构和训练设置。

nanoGPT 是 GPT-2 架构的一种轻量且高效的实现。它采用 GELU 激活函数和与 GPT-2 一致的字节对编码（BPE）分词器（Gage，1994）（Radford 等人，2019年），词汇表扩展至包含 50,257 个词元。学习率预热采用 2,000 步。nanoGPT 模型的训练数据为 OpenWebText。

Llama2 在Shazeer，2020提出的 SwiGLU 激活函数用于前馈网络，从而提升表达能力和参数效率。位置信息通过旋转位置嵌入（RoPE）（Su 等人，2023）进行编码。Llama2 还引入了分组查询注意力（GQA），以降低推理时内存占用和计算成本。该模型采用基于 SentencePiece 的 BPE 分词器，词汇表大小为 32K 个词元。在我们的实验中，1% 的总步数用于学习率预热。Llama2 模型的训练数据为 C4。

Llama3 沿用 Llama2 的密集型 Transformer 设计，同时进行了几项有针对性的改动。它继续采用 GQA，并使用 8 个键值头，以提高解码效率并减小键值缓存的大小。一个主要区别在于分词器：Llama 3 采用了大得多的词汇表，包含 128K 个词元，结合了 tiktoken 分词器中的词元与额外的多语言词元，从而提高了压缩率和语言覆盖范围。为更好地支持长上下文，RoPE 基频提高至 500,000。在我们的实验中，学习率预热占总步数的 1%。Llama3 模型的训练数据为 C4。

在所有实验中，我们均采用 AdamW 优化器（Kingma & Ba，2014；Loshchilov & Hutter，2019年），并使用 $\beta_1 = 0.9$和 $\beta_2 = 0.95$。我们知道，权重衰减与学习率相关，而我们的方法允许采用更大的学习率，因此相应地调整了权重衰减。除非另有说明，我们的实验均采用 0.1 的权重衰减值。其他超参数配置汇总于 表 1。

所有模型均使用 PyTorch（Paszke 等人，2019年）在 A800 GPU 上以 bfloat16 精度训练。所有训练过程均采用余弦学习率调度方案。

[1]https://github.com/zyushun/Adam-mini[2]
https://github.com/karpathy/nanoGPT

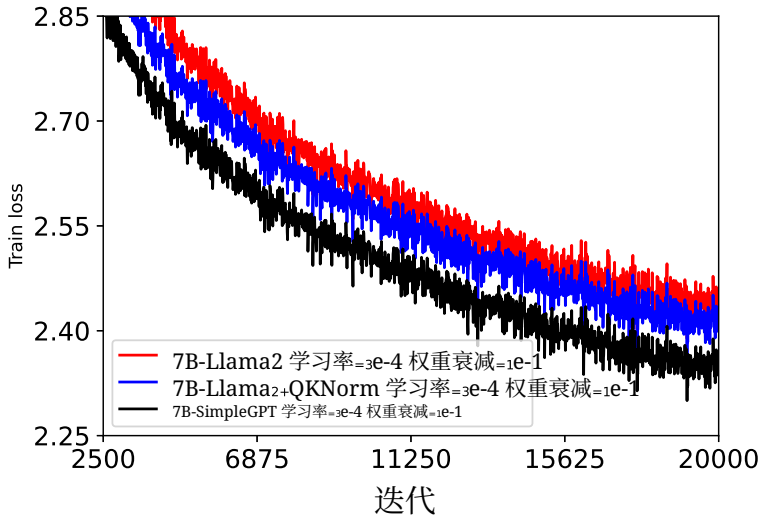
G. SimpleGPT 的参数和配置

表 *1*。 不同规模 SimpleGPT 的模型配置。1B 和 7B 模型基于 Llama 2，8B 模型基于 Llama 3，1.4B 模型基于 nanoGPT。

	SimpleGPT **10亿**	SimpleGPT **70亿**	SimpleGPT **80亿**	SimpleGPT **14亿**
源自	Llama2	Llama2	Llama3	nanoGPT（GPT-2）
层	18	32	32	48
模型维度	2,048	4,096	4,096	1,536
前馈网络维度	5,632	11,008	14,336	6,144
注意力头	16	32	32	24
键/值头	16	32	8	24
激活函数	SwiGLU	SwiGLU	SwiGLU	GeLU
词汇表大小	32,000	32,000	128,000	50,304
位置嵌入（旋转位置编码）	$\theta = 10,000$	$\theta = 10,000$	$\theta = 500,000$	No
批次大小	512×256	2048×192	2048×192	1024×512
训练步数	200K	20K/40K/60K	20K	100K
预热步数	1%	1%	1%	2000

H. SimpleGPT 70亿 的更多实验

此外，我们使用不同的学习率或权重衰减值评估 SimpleGPT 7B 模型。结果如图 8和图 9所示。

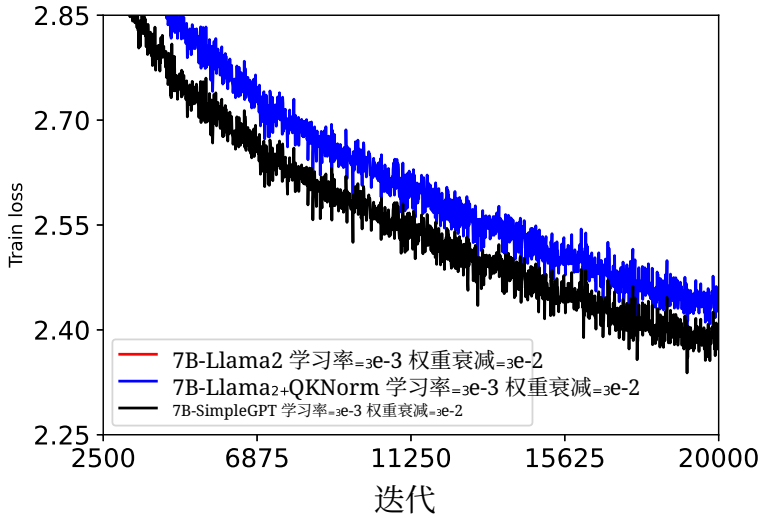


7B-Llama2 学习率=3e-4 权重衰减=1e-1：最后 = 2.431

7B-Llama2+QKNorm 学习率=3e-4 权重衰减=1e-1：最后 = 2.404

7B-SimpleGPT 学习率=3e-4 权重衰减=1e-1：最后 = 2.340

图 *8*。 Llama2 7B、使用 QKNorm 的 Llama2 7B 和 SimpleGPT 70亿 的训练损失曲线，学习率为 3e-4，权重衰减为 0.1。



7B-Llama2 学习率=3e-3 权重衰减=3e-2：最后 = 4.579

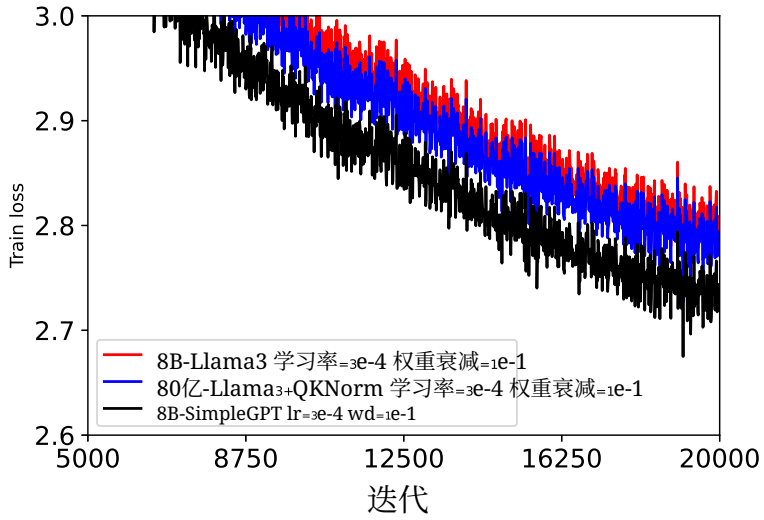
7B-Llama2+QKNorm 学习率=3e-3 权重衰减=3e-2：最后 = 2.430

7B-SimpleGPT 学习率=3e-3 权重衰减=3e-2：最后 = 2.377

图 *9*。 Llama2 7B、使用 QKNorm 的 Llama2 7B 和 SimpleGPT 70亿 的训练损失曲线，学习率为 3e-3，权重衰减为 0.03。

I. 关于 SimpleGPT 80亿的更多实验

此外，我们使用不同的学习率或权重衰减值评估 SimpleGPT 8B 模型。结果如图 10和图 11所示。

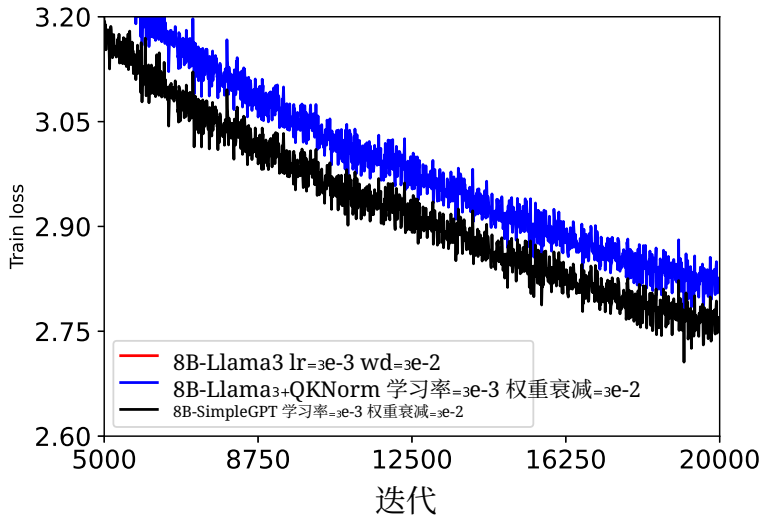


8B-Llama3 lr=3e-4 wd=1e-1：最后 = 2.779

8B-Llama3+QKNorm lr=3e-4 wd=1e-1：最后 = 2.763

8B-SimpleGPT lr=3e-4 wd=1e-1：最后 = 2.709

图 *10.* Llama3 8B、采用 QKNorm 的 Llama3 8B 和 SimpleGPT 8B 在学习率为 3e-4、权重衰减为 0.1 时的训练损失曲线。



8B-Llama3 lr=3e-3 wd=3e-2：最后 = 5.255

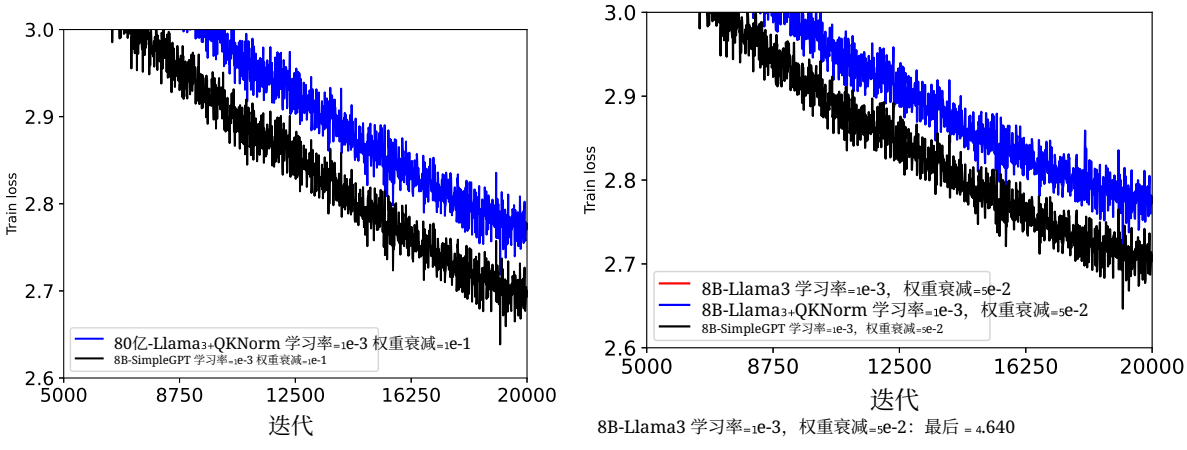
8B-Llama3+QKNorm lr=3e-3 wd=3e-2：最后 = 2.800

8B-SimpleGPT lr=3e-3 wd=3e-2：最后 = 2.738

图 *11.* Llama3 8B、采用 QKNorm 的 Llama3 8B 和 SimpleGPT 8B 在学习率为 3e-3、权重衰减为 0.03 时的训练损失曲线。

J. 权重衰减的更多实验

我们在 SimpleGPT 8B 模型上使用两种不同的权重衰减值进行实验，以评估其对正则化的鲁棒性。在所有测试设置中，SimpleGPT 8B 始终优于采用 QKNorm 的 LLaMA2 8B。这些结果表明，SimpleNorm 的优势不受权重衰减选择的影响，进一步证明了其在大规模训练中的鲁棒性。结果如图12所示。



80亿-Llama3+QKNorm 学习率=1e-3 权重衰减=1e-1：最后 = 2.752

8B-SimpleGPT 学习率=1e-3 权重衰减=1e-1：最后 = 2.669

*(a)*采用 wd=0.1 的 8B。

8B-Llama3 学习率=1e-3，权重衰减=5e-2：最后 = 4.640

8B-Llama3+QKNorm 学习率=1e-3，权重衰减=5e-2：最后 = 2.752

8B-SimpleGPT 学习率=1e-3，权重衰减=5e-2：最后 = 2.678

(b) 采用 wd=0.05 的 8B。

图*12*。总体共同　比较采用两种不同权重衰减值的 Llama3 8B（使用 QKNorm）和 SimpleGPT 1B。