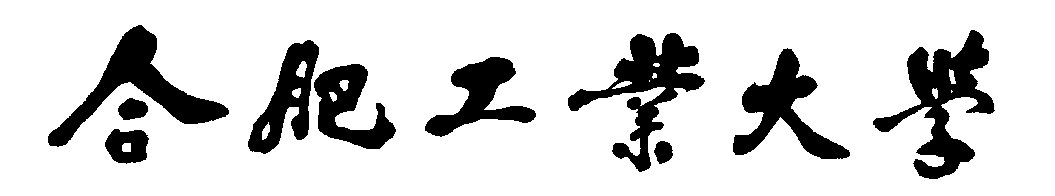
****

计算机与信息学院

人工智能原理期末报告

|  |  |
| --- | --- |
| 专 业 班 级 | 计科21-3班 |
| 学生姓名及学号 | 陈嘉乐 2021218152 |
| 任 课 教 师 | 张赞 |
| 选题类型 | 一篇 |
| 阅读论文来源 | ICML 2023 |
| 阅读论文题目 | Second-order regression models exhibit progressive sharpening to the edge of stability |
|  |  |
| 2023 ~2024学年第 一 学期 | |

说 明

期末报告是关于该课程教学内容、过程及效果的一种记录和总结，因此，应注意以下事项和要求：

1．期末报告要求：**格式规范，语言表达清楚，数据和程序真实。并能够理论联系实际，认真分析研究中出现的问题与现象，总结经验。**

2．每位同学应独立完成期末报告的撰写，严禁抄袭或拷贝，否则，一经查实，按作弊论取，并取消成绩。

3. 可根据实际需要调整每个单元格的篇幅，但是不建议超过40页。

4. 请按照要求填写期末报告。算法源代码请放置在附录中。

本报告翻译的主要论文

ICML：《Second-order regression models exhibit progressive sharpening to the edge of stability》

论文总结：

（1）.研究主题和现状

二阶回归模型呈现向稳定边缘逐步锐化的现象

最近在深度学习理论理解方面的趋势集中在线性化区域，其中神经切向核（NTK）控制学习动态。NTK描述了所有网络在足够短的时间范围内的学习动态，并且可以描述宽网络在较大时间范围内的动态。在NTK的区域中，存在一个函数空间ODE，允许对网络输出进行明确的表征。这种方法已被广泛应用以深入了解宽神经网络，但它存在一个主要限制：该模型在参数上是线性的，因此它描述了一种相对琐碎的动态区域，无法捕捉特征学习，并且不能准确地表示实践中经常观察到的复杂训练现象。

（2）.主要内容

 本文针对二阶回归模型在机器学习中的潜在价值展开了研究，提出了二阶回归模型呈现向稳定边缘逐步锐化的现象。作者首先介绍了二阶回归模型呈现了NTK特征值向一个与稳定边缘稍有不同的值逐渐尖锐化的特性，并表明了二阶回归模型在机器学习中的潜力。为了实现这一目标，他们通过二阶泰勒展开的算法转换、低维动力学和高维动力学方面的研究和建立模型进行训练并在多个基准数据集上进行了全面的定量和定性实验证明了他们提出的现象的正确性和有效性。

（3）.主要发现

在特定的低维设置中，我们证明了最大的NTK特征值收敛到（接近）稳定边缘，并且我们通过实验证明，在大数据点、大模型的极限情况下，通常会发生朝稳定边缘的逐渐尖锐化。最后，我们对一个真实神经网络的性质进行了数值分析，并利用我们理论分析的工具，展示了“在实际中”的稳定边缘行为显示出与理论模型相似的一些模式。

（4）.未来展望

未来研究的一个方向是在大D和P的二次回归模型中定量地理解渐进尖锐化和EOS行为。特别是在边缘稳定性区域，将最终偏差作为函数进行计算仍然是一个有趣的开放问题。了解更高阶的项如何影响训练动力学也很有用，特别是关于围绕y = 2的振荡的详细信息。

本报告的主要贡献

1.二阶泰勒展开的算法转换：我们首先对任意模型进行二阶泰勒展开，将其转换为二阶回归模型。任何模型都可以通过对任意可微分点进行二阶泰勒展开转化为二次回归模型。即使是随机的、非结构化的二次回归模型也会导致稳定边缘（EOS）行为。这种算法转换为我们提供了研究的基础，使得我们能够从数学的角度深入探讨模型的动态行为。假设我们有一个复杂的神经网络模型，我们可以通过对其进行二阶泰勒展开，将其转换为一个简单的二阶回归模型。这就好比将一个复杂的问题转化为一个更易于解释和理解的问题，类似于在复杂数据集上进行拟合的简化。这种创新点在于将复杂的深度学习模型简化为具有清晰数学形式的模型，从而更容易理解和分析。

2.对逐步锐化现象的深入数学剖析：我们通过对二阶回归模型的深入数学分析，揭示了在训练过程中逐步锐化现象的原因。为了研究稳定边缘行为，我们需要一些初始化条件，使得曲率T(0)随时间增加这一现象被称为逐渐尖锐化。逐渐尖锐化在机器学习模型中被证明是普遍存在的，因此任何有用的现象学模型都应该显示出这一现象。如果我们考虑一个简单的拟合问题，例如一个二次曲线回归模型，通过调整学习率和初始参数，我们可能会观察到模型在训练过程中逐步锐化的现象。这就类似于模型逐渐调整自身以更好地拟合数据。这涉及到对参数、残差和曲率等关键概念的数学理解。通过深入剖析，我们发现了这种行为与模型参数、学习率等因素之间的复杂交互关系，这为模型动态的数学建模提供了新的视角。

3.对模型稳定性边缘的理论探讨：我们不仅关注逐步锐化现象，还将注意力扩展到模型接近稳定性边缘时的行为。模型稳定性边缘的理论探讨是深入研究模型在参数空间中的边缘情况，尤其是关注模型在临界状态下的行为。这方面的理论分析对于理解模型的鲁棒性、性能极限和优化行为具有重要意义。分析模型参数的改变对边缘解的影响，特别是在参数空间的边缘情况下。这可能涉及到参数的灵敏度分析，以及在不同条件下边缘解的稳定性和性能。探讨在参数空间的边缘情况下，优化算法的收敛性和效率。这包括分析在参数边缘附近是否存在梯度消失或梯度爆炸等现象，以及这些现象如何影响优化过程。通过理论探讨，我们推断出与模型在这一阶段的行为相关的因素，例如曲率的正则化和学习率的临界值。这些洞见为模型在边缘行为的理论解释提供了新的线索。

4.数学推导和定量分析：我们在文中进行了详细的数学推导和定量分析，以支持对模型动态行为的理解。例如在梯度下降一节中论证了D = 1时损失景观L(·)作为参数θ的函数，其中P = 2，E = 0，Q具有特征值1和-0.1。GD轨迹收敛到具有比初始化时更大曲率的最小值，因此显示了逐步尖锐化。两步动态，其中仅考虑偶数迭代次数，表现出在稳定边缘附近的较少振荡。这种严密的数学分析是对深度学习领域研究方法的一种创新，强调了理论解释和数学建模在解释复杂现象中的重要性。

5.吸取经验：像渐进尖锐化和边缘稳定性这样的行为可能是高维梯度训练非线性模型的常见特征。实际上，在没有任何与深度学习模型的联系的简单设置中，我们的简化模型，对应于1个数据点和2个参数，经过轻微调整可以证明显示EOS行为。这与对CIFAR模型的分析一起表明，这一普遍机制可能具有低维描述。

6.不足之处：没有涉及模型的特征学习方面。在二次回归模型中，特征学习编码之间的关系中，特别是在特征结构之间的关系中。如何调解这变量的动力学可能为理解特征学习提供一个与现有理论方法互补的定量基础。

感想、体会、建议

在学习二阶回归模型的过程中，我深入了解了机器学习中的优化问题。分析模型的动力学行为，尤其是关注模型向稳定边缘逐步锐化的现象，帮助我更好地理解训练过程中出现的各种现象。这种深入分析为我提供了一种更深层次理解优化算法设计和理论分析的途径。

在研究二阶回归模型的学习收获中，我学会了从复杂的模型中提炼出简化的形式。将神经网络等复杂模型近似为二阶回归模型的能力，培养了我在建模和分析方面的技能。这对于理论研究者和算法工程师来说，是一项至关重要的能力。

人工智能课程的学习体会涉及到广泛的领域，其中包括机器学习、深度学习、自然语言处理等。在这个过程中，我发现实践至关重要。通过参与项目和处理实际案例，我能够将课堂理论知识应用到实际场景中，加深对人工智能领域的理解。

课程教授的知识涵盖了从基础概念到前沿技术的全面范围。在学习过程中，我意识到保持对领域内最新研究和发展的关注是至关重要的，因为人工智能领域变化迅速，持续学习是跟上技术趋势的关键。

综合而言，学习人工智能和深度学习是一项挑战而又充满乐趣的任务。通过理论学习和实际项目实践，我逐渐深入了解这个快速发展的领域，为未来的研究和职业生涯做好了充分准备。

二阶回归模型呈现向稳定边缘逐步锐化的现象

陈嘉乐、2021218152

计算机与信息学院、计算机科学与技术、21级3班

摘 要：近期对使用较大步长的梯度下降的研究表明，通常存在一种情境，即损失Hessian的最大特征值一开始呈现增加趋势（逐渐尖锐化），随后稳定在最大值附近，从而促使收敛（稳定边缘）。这些现象本质上是非线性的，且在常数神经切向核（NTK）情境下的模型中并不存在，因为在这种情况下，预测函数在参数上近似为线性。因此，我们考虑下一个最简单的预测模型类别，即在参数上是二次的模型，我们称之为二阶回归模型。对于二维的二次目标，我们证明了这个二阶回归模型呈现了NTK特征值向一个与稳定边缘稍有不同的值逐渐尖锐化的特性，而我们对这个值进行了明确的计算。在更高维度下，即使没有神经网络的特定结构，该模型通常表现出类似的行为，这表明逐渐尖锐化和稳定边缘的行为并不是神经网络的唯一特征，而可能是高维非线性模型中离散学习算法的更普遍属性。

关键词：二阶回归模型；梯度下降；神经网络；神经正切核；特征值；逐步锐化；稳定边缘；非线性；

引言

在机器学习和深度学习领域，研究人员一直在努力理解模型训练过程中的动态行为，特别是关于模型锐化和稳定性边缘的现象。本论文聚焦于二阶回归模型，以探讨这类模型在训练过程中表现出的逐步锐化走向稳定边缘的行为。

最近的研究中，人们对于模型在稳定性边缘附近的动态行为产生了浓厚兴趣。我们的工作建立在对先前类似问题的研究的基础上，特别关注了二阶回归模型。这类模型的特点在于通过对参数进行二阶泰勒展开，将任意模型转化为二次回归模型。通过对这一模型的研究，我们得出了关于模型锐化和在接近稳定性边缘时的动态行为的新见解。

我们在文中阐述了二阶回归模型的基本数学框架，并展示了在训练过程中观察到的逐步锐化现象。与此同时，我们深入探讨了这种行为的原因，揭示了模型在较高维度下的一些独特性质。这不仅有助于理解二阶回归模型的动态特性，还可能为解释更复杂深度学习模型的行为提供一些启示。

通过对逐步锐化现象的深入研究，我们的目标是更好地理解模型训练的动态过程，为优化训练算法、提高模型性能提供更为准确的指导。本论文为深度学习领域的研究者和从业者提供了一个新的视角，推动了对于模型锐化和稳定性边缘现象的理论认识。

# 1问题背景和工作准备

1.1问题背景

最近在深度学习理论理解方面的趋势集中在线性化区域，其中神经切向核（NTK）控制学习动态[11][12]。NTK描述了所有网络在足够短的时间范围内的学习动态，并且可以描述宽网络在较大时间范围内的动态。在NTK的区域中，存在一个函数空间ODE，允许对网络输出进行明确的表征[11][12]。这种方法已被广泛应用以深入了解宽神经网络，但它存在一个主要限制：该模型在参数上是线性的，因此它描述了一种相对琐碎的动态区域，无法捕捉特征学习，并且不能准确地表示实践中经常观察到的复杂训练现象。尽管其他大宽度缩放区域可以保留一些非线性并允许某些类型的特征学习[2]，但这些方法往往专注于小学习率或连续时间动态。相反，最近的实证研究强调了在使用大学习率训练实际网络时由非线性离散动态引起的许多重要现象[6][7][9][15]。特别是，许多实验显示网络倾向于在稳定边缘附近逐渐尖锐化曲率的趋势，其中损失Hessian的最大特征值在训练过程中增加，直到它稳定在大约学习率的两倍的值，对应于在二次势能中梯度下降将收敛的最大特征值[3][4][8][18]。

为了更好地理解这种行为，我们引入了一类二次回归模型，展现了所有相关的现象学，同时足够简单以便进行数值和分析理解。在以前的研究中，已经研究过这类模型，以理解超出NTK区域发生的其他现象[17][22]。在特定的低维设置中，我们证明了最大的NTK特征值收敛到（接近）稳定边缘，并且我们通过实验证明，在大数据点、大模型的极限情况下，通常会发生朝稳定边缘的逐渐尖锐化。最后，我们对一个真实神经网络的性质进行了数值分析，并利用我们理论分析的工具，展示了“在实际中”的稳定边缘行为显示出与理论模型相似的一些模式。

1.2工作准备

1.2.1并发工作

几个同时进行的研究探讨了关于稳定边缘的类似问题，并提出了对我们在此发展的结论进行有效补充的见解。

Zhu等人[21][22]提出了一个极简模型，采用4次幂函数（或8次目标函数）的形式，并证明了尖锐度收敛到接近但不等于稳定边缘的值。虽然这些结果与我们在低维设置中的结果相似，但我们发现稳定边缘行为可以出现在2次幂函数中，并且最终的尖锐度与收敛阈值之间存在我们计算的精确问题相关的差异。

Damian等人提出了一个解释性模型，该模型基于Hessian的最大特征值的非线性自相互作用以及其与其他特征值的有效相互作用，用于解释在稳定边缘附近尖锐度的稳定。低维动力学在尖锐度接近稳定性阈值时的短时间内捕捉到完整的行为，并且与我们在模型的低维极限中发展的微分方程非常相似。我们的设置在某种程度上较为受限，因为它局限于二次模型，但它为我们提供了一些在Damian等人[5]中不可能获得的额外见解。特别是，我们能够证明渐进尖锐化在高维模型中普遍发生，至少在早期时期发生，并且我们发现最终曲率与收敛阈值之间存在可预测的差异。

# 2.二次回归模型

我们首先定义我们的基本二次回归模型。给定一个P维参数向量θ，D维输出f(θ)由以下公式给出：

(1)

这里y是一个D维向量，G是一个D×P维矩阵，Q是一个D × P × P维的张量，对最后两个指标是对称的 - 也就是说，Q(·, ·)以两个P维向量作为输入，并输出一个D维向量，满足Q(θ, θ)α = θ>Qαθ。如果Q = 0，该模型对应于线性回归。我们可以通过以下方式还原出D × P维的雅可比矩阵J = ：

(2)

任何模型都可以通过对任意可微分点进行二阶泰勒展开转化为二次回归模型。浅层MLP的二次展开已经在先前的研究中进行了研究[1][12]，而对于小Q的微扰理论在Roberts等人[17]中有所研究。其他相关模型的细节在附录A中说明。我们将提供证据表明，即使是随机的、非结构化的二次回归模型也会导致稳定边缘（EOS）行为。

在这项工作中，我们专注于均方误差（MSE）损失设置。更明确地说，给定一个D维目标向量yˆ，损失L(θ)可以用残差z = f(θ) − yˆ的形式写成：

(3)

正如我们将展示的那样，在这个设置中，动态可以仅用z和J来在函数空间中表示。

# 3.低维动力学

我们首先关注D = 1的情况，即单个数据点的动态。不失一般性，我们可以将损失函数写为：

(4)

或者一个P × P矩阵Q和标量E˜，它们定义了这个问题。我们将分析与（标量）残差 z = 1/2θ>Qθ − E 以及 1 × P 维雅可比矩阵 J = 有关的动态。特别是，我们将关注标量曲率 （神经切向核，或NTK）的动态。

3.1梯度流

我们首先考虑相对于参数θ的损失L的梯度流（GF）动力学。给定一个缩放因子η，梯度流动力学可表示为：

(5)

在接近的情况下，z和J的动力学如下：

(7)

我们注意到T(0) = η - 也就是说，曲率被缩放因子归一化。动力学方程如下：

(8)

为了研究稳定边缘行为，我们需要一些初始化条件，使得曲率T(0)随时间增加 - 这一现象被称为逐渐尖锐化。逐渐尖锐化在机器学习模型中被证明是普遍存在的[3][4]，因此任何有用的现象学模型都应该显示出这一现象。在附录B.1中，我们确认该模型在一系列初始条件下都表现出逐渐尖锐化。

3.2梯度下降

我们将展示，对于学习率为η的梯度下降（GD）动力学，存在一类初始化，使得D = 1的二次回归模型表现出稳定边缘（EOS）行为。对于该模型，我们将EOS行为定义为一种情境，其中动力学导致NTK 的最大特征值λmax保持接近关键值2/η。这个关键值对应于动力学在损失L的最小值附近指数收敛的最大学习率。这个边界对应于T(0) = 2 - 因此T(0)被解释为经过缩放的曲率。

我们将展示存在一类模型，对于任何距离，都存在一个模型使得EOS收敛到距离稳定边缘的地方 - 对于初始化空间中的任何非平凡体积内的初始化。此外，我们将证明这种初始化体积在一组自然坐标中对于模型是均匀的。

我们将证明D = 1模型中EOS的这种形式，并发现这对于大的D同样成立的经验证据。请注意，我们将EOS的定义与最大损失Hessian特征值而非最大NTK特征值对齐；请参阅附录A.1以了解为什么这对于MSE损失是合适的讨论。

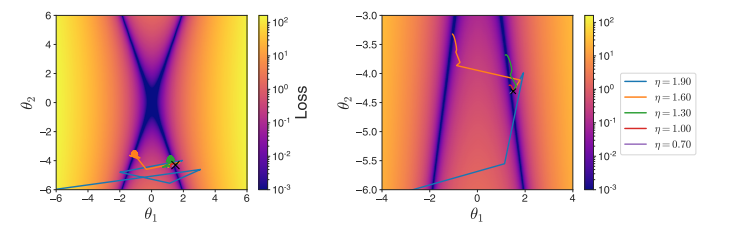
当Q具有正负特征值时，损失地形是一个双曲抛物面的平方（图1，左图）。正如梯度流分析所建议的，这会导致一些轨迹在收敛之前增加它们的曲率。最终的曲率自然取决于初始化和学习率。分析梯度下降动力学的一个挑战是，对于较大的学习率，它们会快速而且强烈地在最小值周围振荡。缓解这个问题的一种方式是仅考虑每隔一个步骤（图1，右图）。我们将利用这一观察结果直接分析梯度下降（GD）动力学，找到这些轨迹表现出稳定边缘行为的配置。

图1

图1显示了D = 1时损失景观L(·)作为参数θ的函数，其中P = 2，E = 0，Q具有特征值1和-0.1。图中的GD轨迹（在(1.5, -4.32)处标有x）收敛到具有比初始化时更大曲率的最小值，因此显示了逐步尖锐化（左图）。两步动态，其中仅考虑偶数迭代次数，表现出在稳定边缘附近的较少振荡（右图）。

设θt为第t步的参数。梯度下降方程如下：

(9)

在z˜ − T坐标系中，梯度下降方程变为（附录B.2）：

(10)

(11)

如果Q是可逆的，那么E˜ ≡ ηE = T(−1) − 2。根据定义，Tt(0) = ηJtJt = ηλmax,t是（经过缩放的）NTK特征值。在这些坐标中，EOS行为对应于在趋近于0时保持Tt(0)接近值2的动态。

在本节的其余部分，我们将专注于P = 2的情况。正如我们将看到的，这使我们能够仅通过z˜和T(0)（残差和曲率）来表示动态。

3.2.1减少弹射动力

如果Q的特征值是{-ω，ω}，且= 0，那么该模型变成了等效于一个单隐藏层线性网络，其中只有一个训练数据点（附录A.2） - 也被称为弹射阶段动态。这个模型不表现出尖锐化或稳定边缘行为[13][14]。我们可以在 − T(0)坐标中证明这一点。我们假设不失一般性，特征值为{-1, 1} - 这可以通过重新缩放z˜来实现。我们可以将动态重写为仅涉及z˜和曲率T(0)（附录B.3）：

(12)

(13)

对于= 0，我们可以看到sign(∆T(0)) = sign(Tt(0)−4)，就像在[13]中一样 - 因此收敛意味着曲率严格减小。对于= 0，存在曲率可以增加的区域（附录B.3）。然而，仍然没有EOS行为 - 如果初始远离2/η，就没有机制将其稳定在边缘附近。

3.2.2稳定制度的边缘

在本节中，我们考虑Q具有两个特征值的情况 - 其中一个是大的正值，另一个是小的负值。不失一般性，我们假设Q的最大特征值是1。我们用 -（其中0 < ≤ 1）表示第二个特征值。使用这个符号，我们可以将动力学方程（附录B.3）写为：

(14)

(15)

对于小的ε，存在轨迹，其中λmax最初远离2/η但收敛于它（图2，左图） - 换句话说，EOS行为。我们使用了各种步长η，但在对(z0, T0(0))进行初始化时以(ηz0, ηT0(0))对来展示z˜-T(0)坐标的通用性。

为了定量地理解逐渐尖锐化和稳定边缘，查看两步动力学是有用的。研究两步动力学的另一个动机来自于对具有大步长λ的线性最小二乘（即线性模型）梯度下降的分析。对于每个坐标˜θ，一步和两步动力学为：

(16)

虽然在λ< 2时动力学会收敛，但如果λ>1，一步动力学在接近最小值时会交替变换符号，而两步动力学会保持˜θ的符号，轨迹不会出现振荡。

同样，在两参数模型中绘制每隔一个迭代更清晰地展示了这一现象。对于小的ε，动力学显示了在[14]中描述的明显阶段：T(0)的初始增加，z˜的缓慢增加，然后T(0)的减小，最后z˜缓慢减小，而T(0)保持在2附近（图2，中图）。

不幸的是，由方程14和15定义的两步动力学的版本更为复杂 - 它们对于T(0)是3阶的，对于z˜是9阶的；更详细的讨论请参见附录B.4。然而，我们可以在z˜趋近于0时分析动力学。为了理解EOS行为的机制，理解两步动力学的零等位线是有用的。零等位线是由初始化组成的曲线（˜z, fz˜(˜z))和（˜z, fT(˜z))，在这些初始化中，z˜和T(0)在两步之后分别保持不变。

我们定义多项式p (, T)和pT(, T)，使得+1 − z = p (t, Tt)和Tt+1(0) − Tt(0) = pT(t, Tt)。然后f ()和fT()遵循隐式方程：

(17)

这些表达式分别是f ()和fT()的三次多项式 - 因此随着趋近于0，存在三种可能的解。我们特别关注通过 = 0、T(0) = 2的解 - 也就是与EOS相对应的临界点。

附录B.4中详细说明的计算表明，两个零等位线之间的距离与ε呈线性关系，因此它们在ε趋近于0时变得接近（图2，中图）。此外，轨迹保持在f附近 - 这导致EOS行为。这表明在零等位线附近，动力学是缓慢的，轨迹似乎在接近一个吸引子。我们可以通过改变变量yt ≡ Tt(0) − f (t) - 到零等位线的距离来找到吸引子的结构。我们可以通过在z˜和y的最低阶上展开动力学来建立直观。附录B.5中的直接计算给出了我们的近似：

(18)

(19)

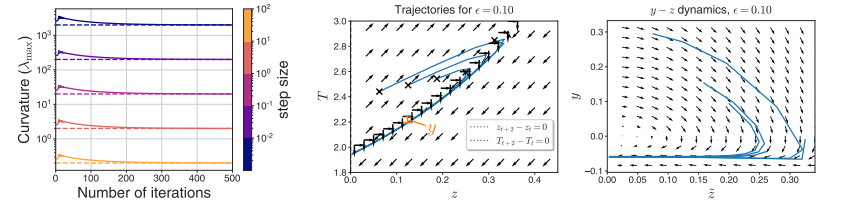


图2

图 2.左图：绘制 D = 1、P = 2 模型的曲率图，显示不同步长 ( = 5 - 10-3) 的 EOS 行为。

). 中间：绘制每隔一次迭代，我们都可以看到各种初始化（黑色 x）、 - T(0)空间中的轨迹都保持在零线（, f ()）附近，即箭头代表动态流的方向。右图： 将变量改为 y = T(0) - f ()，可以看到曲线迅速集中到一条近乎恒定、小的负 y 曲线。

# 4.高维动力学

在本节中，我们分析⼤ D 的动态。我们关注随机初始化的Q，并证明渐进锐化是普遍存在的。然后，我们证明⼆次回归模型在⼀系列初始化范围内显⽰出稳 定性⾏为的边缘。

4.1梯度流动⼒学

MSE 损失函数L上的梯度流动⼒学通常可以写为

(21)

其中，J 是 D × P 维雅各布因子，z 是残差 f(θ) - yˆ 的 D 维向量。对于二次

对于二次回归模型，z 和 J 的动态再次接近：

(22)

当Q = 0（线性化区域）时，J是常数，动力学在z方面是线性的，并由D × D矩阵的特征结构控制，即经验NTK。我们关心的是在GF下发生渐进尖锐化的情境。我们可以研究在随机初始化的早期时期的最大特征值λmax的动力学。在附录C.1中，我们证明了以下定理：

定理4.1

设z、J和Q的元素以零均值和方差a、b和1独立同分布初始化，分布对于数据和参数空间的旋转具有不变性，并且具有有限的四阶矩。设λmax是的最大特征值。在D和P足够大且D/P比值固定的极限情况下，初始时我们有

(23)

其中E表示在初始化时对z、J和Q的期望。

与D = 1的情况类似，定理4.1表明很容易找到显示渐进尖锐化的初始化 - 增加σz会使尖锐化更加明显。

4.2梯度下降动力学

我们现在考虑有限步长的梯度下降（GD）动力学。θ的动力学由以下方程给出：

(24)

在这种情况下，动力学方程可以写成：

(25)

(26)

如果Q = 0，动力学将简化为在二次势能中的离散梯度下降 - 当且仅当λmax < 2/η时收敛。一个直接的问题是：在方程25中，η^2什么时候影响动力学？鉴于它与η和z的高次幂成比例，我们可以猜想这两个项的幅度比rNL与||z||2和η成正比。附录C.2中的计算显示，对于随机旋转不变的初始化，我们有：

(27)

其中初始化统计量的定义如定理4.1所述。这证实了增加学习率和残差幅值||z||会增加动力学与GF的偏差，并揭示出非线性的程度对J不敏感。

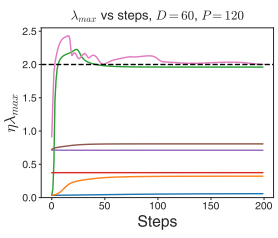
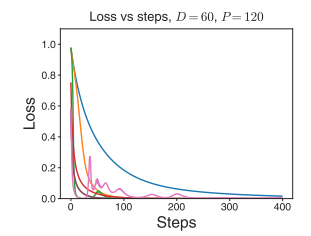


图3 图4

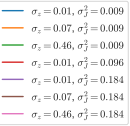


图 3、图4：二次回归模型中的 GD 动态。对于较小的 σz，损失函数是单调的（左图）。对于较大的 σz，最终 λmax 较高（右图）。

更高（右图）。由于锐化使 ηλmax 接近 2，非线性效应诱发了 EOS 行为

我们可以在GD方程的动力学中看到这种现象（图3）。在这里，我们绘制了不同轨迹，这些轨迹是根据定理4.1中的随机初始化得到的，其中D = 60，P = 120，η = 1。随着σz的增加，曲率λmax也增加（如定理4.1所示），当σz为O(1)时，动力学是非线性的（如rNL所预测），并出现了EOS行为。这表明方程25中的第二项对于稳定λmax至关重要。我们可以通过在多个种子上以各种η、D、P、σz和σJ初始化，并绘制达到的最终λmax的相图，更普遍地确认这一点。我们可以通过对参数和初始化进行一些重新缩放来简化绘图。在重新缩放的变量中：

(28)

动力学等同于带有η = 1的方程25和26。与方程56-57的 − T(0)模型一样，在重新缩放的坐标中，λmax相当于未缩放坐标中的ηλmax。我们还可以为z和J定义重新缩放的初始化。如果我们设置：

(29)

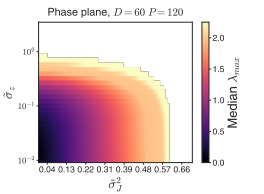
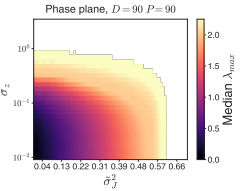
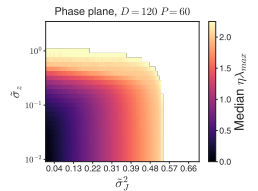


图5 图6 图7

图 5、图6、图7：不同 D 和 P 条件下二次回归模型的 a/b 相位平面。

每对、 都使用 100 个随机种子初始化模型，并迭代直到收敛。对于每一对、 我们绘制出 NTK . 对于中间的 、发生锐化和非线性 z 动态时，轨迹趋于收敛，因此 NTK 的 λmax 接近 2/η（EOS）。

那么我们有rNL = ，这样可以更容易地在(D, P)对之间进行比较。

使用这种初始化方案，我们可以绘制λmax的最终值作为和的函数，对于每个，对都有100次独立的随机初始化（图4）。我们看到关键是使rNL = ˜σz达到O(1) - 对应于初始附近的渐进尖锐化和非线性动力学。特别是，小值的初始化收敛到EOS的轨迹首先尖锐化，然后在λmax = 2/η附近稳定。大和大的动力学会发散。在一系列范围内有一个小范围的初始，其最终λmax ≈ 2/η；这些对应于接近EOS的初始化模型，它们保持在其附近。

这表明渐进尖锐化和稳定边缘并不是神经网络模型的独特特征，而可能是高维非线性模型中学习的更一般性质。

# 5.与现实世界模型的连接

在这一部分，我们在“真实世界”模型中进行数值实验，并将其行为与我们在简化模型上的理论进行比较。沿用(Cohen et al., 2022a)的方法[4]，我们使用二次损失在来自CIFAR10的5000个示例上训练了一个具有2个隐藏层的tanh网络，学习率为10^{-2}，这个设置显示出边缘稳定性行为。在EOS的开始附近，我们使用Lanczos方法[7][16]（近似计算了的最大特征值λ1及其相应的特征向量v1。我们使用v1计算z1 = v1^Tz，其中z是残差向量f(X, θ) − Y，f是神经网络函数，X是训练输入，Y是标签，θ是参数。NTK中的EOS行为类似于(Cohen et al., 2022a)中基于完整Hessian定义的EOS行为（图5，左和右）。再次绘制每隔一步的轨迹可消除高频振荡（图5，中）。与D = 1，P = 2模型不同，有多次穿越临界线λmax = 2/η。

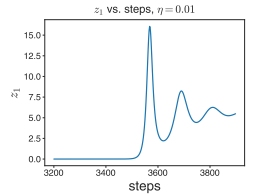
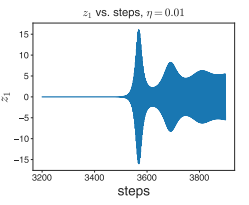


图9 图10

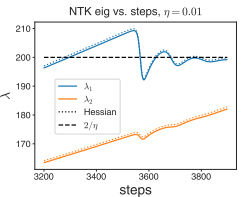


图11

图 9、图10、图11：以 CIFAR 为基础训练的 FCN 显示了多个周期的锐化和稳定边缘行为。z1 是训练集残差 f(X, θ) - Y 在顶部 NTK 特征模式 v1 上的投影，其幅度不断增大，并在 0 附近摆动（左图）。绘制动态消除高频振荡（中）。最大特征值 λ1 多次越过稳定边缘，但第二大特征值λ2 仍低于稳定边缘（右图）。赫氏特征值（虚线）的动态也类似。

有证据表明，二次回归模型的低维特征可以用来解释EOS行为的一些方面。我们通过自动微分计算输出f(x, θ)的二阶导数，用Q(·, ·)表示得到的张量。我们可以使用矩阵-向量乘积计算矩阵Q1的谱，其中Q1 ≡ v1 · Q(·, ·)，这是Q输出在v1方向的投影，而无需在内存中实例化Q（图6，左）。该图显示，从步骤3200到3900（我们绘图的范围）的谱基本上没有变化。这表明在显示这些EOS动态时，Q的变化不大。我们还可以看到，在v1方向上，Q要比随机方向大得多。定义y = λ1η − 2。绘制z1相对于2yz的两步动力学，我们看到了显著的一致性（图6，中）。这与我们简化模型中z˜动力学的形式相同。也可以通过在对y = λ1η − 2的雅可比矩阵保持不变的情况下两次迭代方程25，并且忽略掉与η高阶的项，找到这个形式。这表明在这种特定的EOS行为中，就像在我们的简化模型中一样，特征值的动力学比特征基中的任何旋转更重要。

y的动力学更为复杂；yt+2 − yt与z1^2呈反相关关系，但在y和z1方面没有低阶的函数形式（附录D.2）。通过绘制η^2Q1(Jz1v1, Jz1v1)（来自v1方向对z1动力学的非线性贡献）和λ1z1（线性化贡献）的比率，并将其与y的动力学进行比较（图6，右），我们可以对稳定化的机制进行一些了解。在初始尖锐化期间，这个比率很小，但在曲率第一次降低之前很快变为O(1)。在其余动力学中保持在O(1)。这表明来自最高特征模式对其自身动力学的非线性反馈对理解EOS动态至关重要。

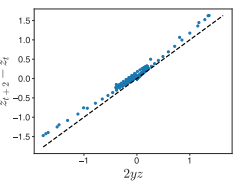
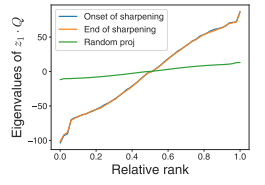


图12 图13

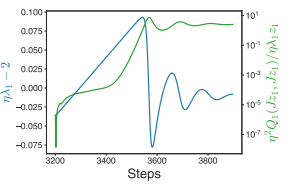


图14

图 12、图13、图14：在 CIFAR10 上训练的 FCN 的稳定边缘动力学过程中，Q 值近似恒定（左图）。投影到最大投影到最大特征方向 v1（蓝色和橙色）比投影到随机方向（绿色）大。两步差 (z1)t+2 - (z1)t的近似值为 2z1y（中间），这是固定特征基模型的前导阶项。非线性动力学贡献 η 2Q1(Jz1v1,Jz1v1)在锐化过程中很小，但在上特征值减小之前（右图）变得很大，这与简单模型的情况相同。

# 6.讨论

6.1从二次回归模型中吸取的教训

从二次回归模型中可以得到的主要教训是，像渐进尖锐化（对于GF和GD都是如此）和边缘稳定性行为（对于GD而言）这样的行为可能是高维梯度训练非线性模型的常见特征。实际上，在没有任何与深度学习模型的联系的简单设置中，我们的简化模型，对应于1个数据点和2个参数，经过轻微调整可以证明显示EOS行为。这与对CIFAR模型的分析一起表明，这一普遍机制可能具有低维描述。

对真实模型的二次逼近定量地捕捉到EOS行为的早期特征（最初返回λmax < 2/η），但不一定捕捉到后续振荡的幅度和周期 - 这需要更高阶的项（附录D.3）。尽管如此，二次逼近确实正确地描述了很多定性行为，包括λmax收敛到一个围绕2/η振荡的极限二周期，其平均值低于2/η。在简化的两参数模型中，可以在收敛时对最终值进行解析预测，事实上我们发现它与2/η的值略有偏差。此外，理论模型和现实模型都表明，曲率由低维反馈机制控制，通过Q1传递，这个机制由介导。

本研究中研究的所有模型的一个关键特征是查看每隔一个迭代（两步动力学）在理论和实证上有助于理解模型。在稳定的边缘附近，这使得最大特征模式的变化很小。在简化模型中，缓慢的z˜动力学（以及相关的缓慢T(0)动力学）使得可以进行详细的理论分析；在CIFAR模型中，两步动力学在z1和λmax两者都缓慢变化。这些小变化的定量比较可能有助于揭示解释其他系统和情境中EOS行为的任何普遍机制/规范形式。

6.2未来的工作

未来研究的一个方向是在大D和P的二次回归模型中定量地理解渐进尖锐化和EOS行为。特别是在边缘稳定性区域，将最终偏差2−ηλmax作为、和D/P的函数进行计算仍然是一个有趣的开放问题。了解更高阶的项如何影响训练动力学也很有用，特别是关于围绕y = 2的振荡的详细信息。

最后，我们的分析没有涉及模型的特征学习方面。在二次回归模型中，特征学习编码在J和z之间的关系中，特别是在z和的特征结构之间的关系中。理解Q如何调解这两个量的动力学可能为理解特征学习提供一个与现有理论方法[10][17]互补的定量基础。

参考文献

1. Bai, Y. and Lee, J. D. Beyond Linearization: On Quadraticand Higher-Order Approximation of Wide Neural Networks. In International Conference on Learning Representations, March 2020.
2. Bordelon, B. and Pehlevan, C. Self-Consistent Dynamical Field Theory of Kernel Evolution in Wide Neural Networks, May 2022.
3. Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In International Conference on Learning Representations, February 2022a.
4. Cohen, J. M., Ghorbani, B., Krishnan, S., Agarwal, N., Medapati, S., Badura, M., Suo, D., Cardoze, D., Nado, Z., Dahl, G. E., and Gilmer, J. Adaptive Gradient Methods at the Edge of Stability, July 2022b.
5. Damian, A., Nichani, E., and Lee, J. D. Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability, September 2022.
6. Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware Minimization for Efficiently Improving Generalization. In International Conference on Learning Representations, April 2022.
7. Ghorbani, B., Krishnan, S., and Xiao, Y. An Investigation into Neural Net Optimization via Hessian Eigenvalue Density. In Proceedings of the 36th International Conference on Machine Learning, pp. 2232–2241. PMLR, May 2019
8. Giladi, N., Nacson, M. S., Hoffer, E., and Soudry, D. At Stability’s Edge: How to Adjust Hyperparameters to Preserve Minima Selection in Asynchronous Training of Neural Networks? In Eighth International Conference on Learning Representations, April 2020.
9. Gilmer, J., Ghorbani, B., Garg, A., Kudugunta, S., Neyshabur, B., Cardoze, D., Dahl, G. E., Nado, Z., and Firat, O. A Loss Curvature Perspective on Training Instabilities of Deep Learning Models. In International Conference on Learning Representations, March 2022.
10. Huang, J. and Yau, H.-T. Dynamics of Deep Neural Networks and Neural Tangent Hierarchy. In Proceedings of the 37th International Conference on Machine Learning, pp. 4542–4551. PMLR, November 2020.
11. Jacot, A., Gabriel, F., and Hongler, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In Advances in Neural Information Processing Systems 31, pp. 8571–8580. Curran Associates, Inc., 2018.
12. Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., SohlDickstein, J., and Pennington, J. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In Advances in Neural Information Processing Systems 32, pp. 8570–8581. Curran Associates, Inc., 2019.
13. Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., and Gur-Ari, G. The large learning rate phase of deep learning: The catapult mechanism. March 2020.
14. Li, Z., Wang, Z., and Li, J. Analyzing Sharpness along GD Trajectory: Progressive Sharpening and Edge of Stability, July 2022.
15. Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. Exploring Generalization in Deep Learning. In Advances in Neural Information Processing Systems 30, pp. 5947–5956. Curran Associates, Inc., 2017.
16. Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J., and Schoenholz, S. S. Neural Tangents: Fast and Easy Infinite Neural Networks in Python. arXiv:1912.02803 [cs, stat], December 2019.
17. Roberts, D. A., Yaida, S., and Hanin, B. The Principles of Deep Learning Theory. May 2022. doi: 10.1017/ 9781009023405.
18. Wu, L., Ma, C., and E, W. How SGD Selects the Global Minima in Over-parameterized Learning: A Dynamical Stability Perspective. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018.
19. Yang, G. Tensor Programs I: Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes. arXiv:1910.12478 [cond-mat, physics:mathph], May 2021.
20. Yang, G., Hu, E. J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., and Gao, J. Tensor Programs V: Tuning Large Neural Networks via ZeroShot Hyperparameter Transfer, March 2022.
21. Zhu, L., Liu, C., Radhakrishnan, A., and Belkin, M. Quadratic models for understanding neural network dynamics, May 2022a
22. Zhu, X., Wang, Z., Wang, X., Zhou, M., and Ge, R. Understanding Edge-of-Stability Training Dynamics with a Minimalist Example, October 2022b.