二阶回归模型表现出逐渐锐化到稳定边缘的趋势

阿蒂什-阿加尔瓦拉、费边-佩德雷戈萨和杰弗里-潘宁顿
谷歌研究，大脑团队
{thetish、pedregosa、jpennin}@google.com

摘要

最近对大步长梯度下降的研究表明，通常会出现这样一种情况：损失赫塞斯的最大特征值最初会增加（逐渐锐化），随后特征值会稳定在允许收敛的最大值附近（稳定边缘）。这些现象本质上是非线性的，不会发生在恒定神经切线核（NTK）机制下的模型中，因为在该机制下，预测函数与参数近似线性。因此，我们考虑下一类最简单的预测模型，即参数为二次方的模型，我们称之为二阶回归模型。对于二维的二次方目标，我们证明这种二阶回归模型的 NTK 特征值会逐渐锐化，其值与我们明确计算的稳定边缘值略有不同。在更高的维度上，即使没有神经网络的特定结构，该模型也普遍表现出类似的行为，这表明渐进锐化和稳定边缘行为并非神经网络的独有特征，而可能是高维非线性模型中离散学习算法的更普遍特性。

1 引言

对深度学习理论理解的最新趋势集中在*线性化*机制上，即由神经切线核（NTK）控制学习动态（Jacot 等人，2018 年；Lee 等人，2019 年）。NTK 可以描述所有网络在足够短的时间跨度内的学习动态，也可以描述宽网络在较大时间跨度内的动态。在 NTK 机制中，有一个函数空间 ODE，可以明确描述网络输出的特征（Jacot 等人，2018 年；Lee 等人，2019 年；Yang，2021 年）。这种方法已被广泛用于深入了解宽神经网络，但它有一个主要局限：模型的参数是线性的，因此它描述的机制具有相对琐碎的动态，无法捕捉特征学习，也不能准确地代表实践中经常观察到的复杂训练现象类型。

虽然其他大宽度缩放机制可以保留一定的非线性，并允许某些类型的特征学习（Bordelon & Pehlevan, 2022; Yang et al.与此相反，最近的实证工作强调了在训练大学习率的实用网络时，非线性离散动态所产生的一些重要现象（Neyshabur 等人，2017；Gilmer 等人，2022；Ghorbani 等人，2019；Foret 等人，2022）。特别是，许多实验表明，网络的曲率有*逐渐向稳定边缘锐化的*趋势，其中损失赫塞斯的最大特征值在训练过程中不断增加，直到稳定在一个大致等于学习率除以二的值，对应于梯度下降将收敛于二次潜能的最大特征值（Wu 等人，2018；Giladi 等人，2020；Cohen 等人，2022b;a）。

为了更好地理解这种行为，我们引入了一类模型，这些模型显示了所有相关现象，但又足够简单，可以进行数值和分析理解。特别是，我们提出了一个简单的*二次回归模型*和相应的四元损失函数，它同时满足了这两个目标。我们证明，在适当的条件下，这个简单模型既能显示渐进锐化行为，*也能*显示稳定边缘行为。然后，我们对一个

更*通用*的模型，在大数据点、大模型极限下显示出这些行为。最后，我们对一个真实神经网络的特性进行了数值分析，并利用理论分析中的工具说明 "野生 "稳定边缘行为显示了与理论模型相同的一些模式。

2 基本四次损失函数

2.1 模型定义

我们考虑优化二次损失函数 $L(\boldsymbol{\theta}) = z^2/2$，其中 z *是* $P \times 1$ 维参数向量 $\boldsymbol{\theta}$ 上的二次函数，$\mathbf{Q}$ 是 $P \times P$ 对称矩阵：

$$z = \frac{1}{2}\boldsymbol{\theta}^{\top}\mathbf{Q}\boldsymbol{\theta} - E. \tag{1}$$

这既可以解释为预测函数与输入参数成二次函数关系的模型，也可以解释为更复杂的非线性函数（如深度网络）的二阶近似。在这一目标中，具有缩放因子 η 的梯度流 (GF) 动态方程为

$$\dot{\theta} = -\eta\nabla_{\theta}L = \eta z\frac{\partial z}{\partial\boldsymbol{\theta}} = \frac{\eta}{2}\left(\boldsymbol{\theta}^{\top}\mathbf{Q}\boldsymbol{\theta} - E\right)\mathbf{Q}\boldsymbol{\theta}。 \tag{2}$$

用 $\tilde{z}$ 和 $1 \times P$ 维雅各布因子 $\mathbf{J} = \partial z/\partial\boldsymbol{\theta}$ 来重写动力学是有用的：

$$\dot{z} = -\eta(\mathbf{J}\mathbf{J}^{\top})z、\qquad \dot{\mathbf{J}} = -2\eta z\mathbf{Q}\mathbf{J}. \tag{3}$$

我们注意到，在这种情况下，神经切核（NTK）是一个由标量 $\mathbf{J}\mathbf{J}^{\top}$ 给出的标量。在这些坐标中，我们有 $E = \mathbf{J}\mathbf{Q}\,\mathbf{J}^{+\top} - 2z$，其中 $\mathbf{Q}^{+}$ 表示摩尔-彭罗斯伪逆。

GF 方程可以通过两种变换来简化。首先，我们变换为 $\tilde{z} = \eta z$ 和 $\tilde{\mathbf{J}} = \eta^{1/2}\mathbf{J}$。接下来，我们旋转 $\boldsymbol{\theta}$ 使 $\mathbf{Q}$ 对角。由于 $\mathbf{Q}$ 是对称的，这总是可能的。由于 NTK 由 $\mathbf{J}\mathbf{J}^{\top}$ 给出，因此这种旋转保留了曲率的动态性。让 $\omega_1 \ldots \omega_P$ 为 $\mathbf{Q}$ 的特征值，$\mathbf{v}_i$ 为相关的特征向量（在退化的情况下，可以选择任意基）。我们定义 $\tilde{J}(\omega_i) = \tilde{\mathbf{J}}\mathbf{v}_i$，即 $\tilde{J}$ 在第 i 个特征向量上的投影。那么梯度流方程可以写成

$$\frac{d\tilde{z}}{dt} = -\tilde{z}\sum_{i=1}^{P}\tilde{J}(\omega_i)^2、\qquad \frac{d\tilde{J}(\omega_i)}{dt} = -2\tilde{z}\omega_i\tilde{J}(\omega_i)^2. \tag{4}$$

第一个等式意味着，$\tilde{z}$ 在 GF 动力下不会改变符号。具有正 $\omega_i\tilde{z}$ 的模式则曲率减小，负 $\omega_i\tilde{z}$ 则曲率增大。

为了研究稳定边缘行为，我们需要初始化，使曲率（本例中为 $\mathbf{J}\mathbf{J}^{\top}$）随时间增加--这种现象称为*渐进锐化*。渐进锐化现象已被证明在机器学习模型中无处不在（科恩等，2022a），因此任何有用的现象学模型也应该显示这种现象。这个二次回归模型的一个初始化是 $\omega_1 = \omega, \omega_2 = \omega, \tilde{J}(\omega_1) = \tilde{J}(\omega_2)$。这种初始化（以及其他初始化）在任何时候都显示出渐进式锐化。

2.2 梯度下降

我们有兴趣了解该模型的稳定*边缘*（EOS）行为：NTK 的最大特征值 $\mathbf{J}\mathbf{J}^{\top}$ 保持在临界值

*$2/\eta$ 附近的*梯度下降（GD）轨迹。（注：我们根据 NTK 的最大特征值定义稳定边缘；对于用平方损失训练的任何二次微分模型，这等同于 Cohen 等人（2022a）在模型收敛到静止点时使用的损失 Hessian 的最大特征值（Jacot et al、2020).）

当 $\mathbf{Q}$ 同时具有正特征值和负特征值时，损失景观就是双曲抛物面的正方形（图 1 左）。梯度流分析表明，这会导致一些轨迹在收敛前增加曲率。这导致最终曲率取决于初始化和学习率。分析梯度下降（GD）过程中的一个挑战是

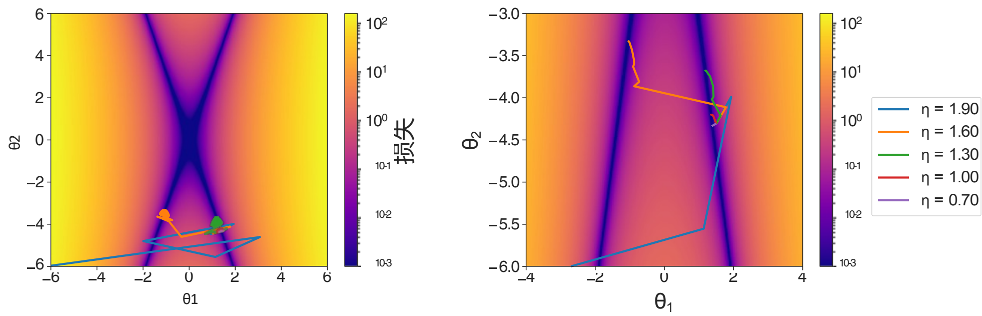


图 1：$D = 2$、$E = 0$ 和 $\mathbf{Q}$ 的特征值分别为 1 和 0.1 时，作为参数 $\boldsymbol{\theta}$ 函数的四次方损失景观（ ）。与初始化时相比，GD 轨迹收敛到曲率更大的极小值，因此呈现逐渐锐化的趋势（左图）。我们只考虑偶数迭代次数的两步动力学在稳定边缘附近的振荡较少（右图）。

在较大的学习率下，动态会在最小值附近快速剧烈振荡。缓解这一问题的方法之一是只考虑每隔一步（图 1 右）。我们将利用这一观察结果直接分析梯度下降（GD）动力学，以找到这些轨迹表现出稳定边缘行为的配置。

在特征基坐标中，梯度下降方程为

$$\tilde{z}_{t+1} - \tilde{z}_t = -\tilde{z}_t \sum_{i=1}^{P} \tilde{J}(\omega_i)_t^2 + \frac{1}{2}(\tilde{z}_t^2) \sum_{i=1}^{P} \omega_i \tilde{J}(\omega_i)_t^2 \tag{5}$$

$$\tilde{J}(\omega)_{i,t+1}^2 - \tilde{J}(\omega)_{i,t}^2 = -\tilde{z}_t \omega_i (2 - \tilde{z}_t \omega_i) \tilde{J}(\omega)_{i,t}^2 \text{ 对于所有 } 1 \le i \le P\text{。} \tag{6}$$

在下文中，我们会发现用下列各项的加权平均值来写动态过程会比较方便 $\tilde{J}(\omega)_i^2$ 而不是模式 $\tilde{J}(\omega_i)$：

$$T(\alpha) = \sum_{i=1}^{P} \omega_i^{\alpha} \tilde{J}(\omega_i^2)\,. \tag{7}$$

动力学方程变为

$$\tilde{z}_{t+1} - \tilde{z}_t = -\tilde{z}_t T_t(0) + \frac{1}{2}(\tilde{z}_t^2) T_t(1) \tag{8}$$

$$T_{t+1}(k) - T_t(k) = -\tilde{z}_t \left(2T_t(k+1) - \tilde{z}_t T_t(k+2)\right). \tag{9}$$

如果 $\mathbf{Q}$ 是可逆的，那么我们有 $E = T_t(-1)\, 2\tilde{z}_t$。请注意，根据定义，$T_t(0) = \eta \mathbf{J}_t \mathbf{J}_t^{\top}$ 是（重新标定的）NTK。稳定边缘行为对应于当 $\tilde{z}_t$ 变为 0 时保持 $T_t(0)$ 接近值 2 的动力学。

2.2.1 减少弹射器动力

如果 $\mathbf{Q}$ 的特征值为 ω 、$\{\omega\}$ 和 $E = 0$，模型就等同于具有一个训练数据点的单隐层线性网络（附录 A.1）--也称为弹射阶段动力学。该模型不会出现锐化或稳定边缘行为（Lewkowycz 等人，2020 年）。作为热身，我们将在 $\tilde{z}$ $T(0)$ 变量中分析该模型，以便分析确实显示锐化和稳定边缘的不同参数设置。

在不失一般性的前提下，我们假设特征值为 1，1--这{可以通}过重新标定 $\tilde{z}$ 来实现。这样，损失函数就是双曲抛物线的平方。由于只有两个变量，我们可以只用 $\tilde{z}$ 和曲率 $T(0)$ 来重写动力学（附录 B.1）：

$$\tilde{z}_{t+1} - \tilde{z}_t = - \tilde{z}\, T_{tt}(0) + \frac{1}{2}(\tilde{z}_t^2)(2\tilde{z}_t + E) \tag{10}$$

$$T_{t+1}(0) - T_t(0) = -2\tilde{z}_t(2\tilde{z}_t + E) + z_t T^2_t(0)\,. \tag{11}$$

对于 $E = 0$，我们可以看到 $\mathrm{sign}(\Delta T(0)) = \mathrm{sign}(T_t(0)-4)$，如 Lewkowycz 等人 (2020) 所述--因此收敛要求曲率严格递减。对于 $E = 0$，有一个曲率可以增加的区域（附录 B.1）。然而，仍然不存在稳定边缘行为--没有一组初始化从 λ_{max} 开始，远离 $2/\eta$，最终接近 $2/\eta$。相反，我们将证明非对称特征值会导致 EOS 行为。

2.2.2 稳定制度边缘

在本节中，我们将考虑 **Q** 有两个特征值的情况，其中一个特征值大且为正，另一个特征值小且为负。在不失一般性的前提下，我们假设 **Q** 的最大特征值为1。我们用 ϵ 表示第二个特征值，即 $0 < \epsilon \ll 1$。用这个符号，我们可以把动力学方程（附录 B.1) 写成

$$\tilde{z}_{t+1} - \tilde{z}_t = -\tilde{z}_t T_t(0) + \frac{1}{2}\tilde{z}_t^2((1-\epsilon)T_t(0) + \epsilon(2\tilde{z}_t + E)) \tag{12}$$

$$T_{t+1}(0) - T_t(0) = -2\tilde{z}_t(\epsilon(2\tilde{z}_t + E) + (1-\epsilon)T_t(0)) + \tilde{z}_t^2[T_t(0) + \epsilon(\epsilon-1)(T_t(0) - E - 2\tilde{z}_t)]\,. \tag{13}$$

对于较小的ϵ，存在这样的轨迹：λ_{max} 最初远离 $2/\eta$，但逐渐向它靠拢（图 2，左）--换句话说，这是 EOS 行为。我们使用了各种步长 η，但都是在成对初始化（$\eta z_0, \eta T_0(0)$），以显示 $\tilde{z}$-$T(0)$ 坐标的普遍性。

为了定量地理解渐进锐化和稳定边缘，研究两步动态是 非常有用的。研究两步动态的另一个动机来自于对大步长 λ 的线性最小二乘法（即线性模型）梯度下降的分析。对于每个坐标 $\tilde{\theta}$，一步动力学和两步动力学分别为

$$\tilde{\theta}_{t+1} - \tilde{\theta}_t = -\lambda\tilde{\theta}_t \text{和} \tilde{\theta}_{t+2} - \tilde{\theta}_t = (1-\lambda)^2\tilde{\theta}_t \quad \text{（二次电动势中的 GD）}. \tag{14}$$

当 $\lambda < 2$ 时，动力学收敛；而当 $\lambda > 1$ 时，一步动力学在接近最小值时发生振荡，而两步动力学保持 $\tilde{\theta}$ 的符号，轨迹没有振荡。

同样，在双参数模型中绘制每一次迭代图也能更清楚地展示这一现象。对于较小的ϵ，动力学表现出（Li 等人，2022 年）所描述的不同阶段：$T(0)$ 最初增大，$\tilde{z}$缓慢增大，然后 T（0）减小，最后$\tilde{z}$缓慢减小，而 T（0）保持在 2 附近（图 2，中间）。

不幸的是，方程 12 和 13 所定义的两步动力学更为复杂--它们在 $T(0)$ 中是三阶的，在 $\tilde{z}$ 中是九阶的；更详细的讨论见附录 B.2。为了理解 EOS 行为的机理，我们有必要了解两步动力学的*零线*。

$\tilde{z}$ 的 nullcline $f_{\tilde{z}}(\tilde{z})$ 和 $T(0)$ 的 nullcline $f_T(\tilde{z})$ 由以下隐式定义

$$(\tilde{z}_{t+2} - \tilde{z}_t)(\tilde{z}, f_{\tilde{z}}(\tilde{z})) = 0, (T_{t+2}(0) - T_t(0))(\tilde{z}, f_T(\tilde{z})) = 0 \tag{15}$$

其中，$\tilde{z}_{t+2}-\tilde{z}_t$ 和 $T_{t+2}(0)-T_t(0)$ 是上述 $\tilde{z}$ 和 $T(0)$ 的高阶多项式。由于这些多项式在 $T(0)$ 中是三次方，因此当 $\tilde{z}$ 变为 0 时有三种可能的解。我们尤其感兴趣的是经过 $\tilde{z} = 0, T(0) = 2$ 的解，即与 EOS 相对应的临界点。

附录 B.2 中的详细计算表明，两条空心线之间的距离是线性的，即
ϵ，因此当 ϵ *变为* 0 时，它们会变得接近（图 2，中间）。此外，轨迹保持在 $f_{\tilde{z}}$
- 这导致了 EOS 行为。这表明在空心线附近动力学速度很慢，轨迹似乎正在接近吸引子。我们可以将变量改为 $y_t = T_t(0) - f_{\tilde{z}}(\tilde{z}_t)$ - 与 $\tilde{z}$ nullcline 的距离，从而找到吸引子的结构。

在 $\tilde{z}$ 和 y 的最低阶，两步动力学方程变为（附录 B.3）：

$$\tilde{z}_{t+2} - \tilde{z}_t = 2y_t \tilde{z}_t + O(y_t \tilde{z}_t^2) + O(y_t^2 \tilde{z}_t) \tag{16}$$

$$y_{t+2} - y_t = -2(4 - 3\epsilon + 4\epsilon^2) y_t \tilde{z}_t^2 - 4\epsilon \tilde{z}_t^2 + \epsilon O(\tilde{z}_t^3) + O(y_t^2 \tilde{z}_t^2) \tag{17}$$

我们马上就会发现，当 $y = 0$ 时，$\tilde{z}_{t+2} - \tilde{z}_t = 0$。我们还可以看到，当 $y_t = 0$ 时，$y_{t+2} - y_t$ 的值为 $O(\epsilon)$，因此对于较小的ϵ，y 的动态变化很慢。

is slow too. Moreover, we see that the coefficient of the $\epsilon \tilde{z}_t^2$ term is negative - the changes in $\tilde{z}$ tend to drive y (and therefore $T(0)$) to decrease. The coefficient of the y_t term is negative as well; the dynamics of y tends to be contractive. The key is that the contractive behavior takes y to an $O(\epsilon)$ fixed point at a rate proportional to $\tilde{z}^2$, while the dynamics of $\tilde{z}$ are proportional to ϵ. This suggests a separation of timescales if $\tilde{z}^2 \quad \epsilon$, where y first equilibrates to a fixed value, and then $\tilde{z}$ converges to 0 (Figure 2, right). This intuition for the lowest order terms can be formalized, and gives us a prediction of $\lim_{t\to\infty} y_t = -\epsilon/2$, confirmed numerically in the full model (Appendix B.5).

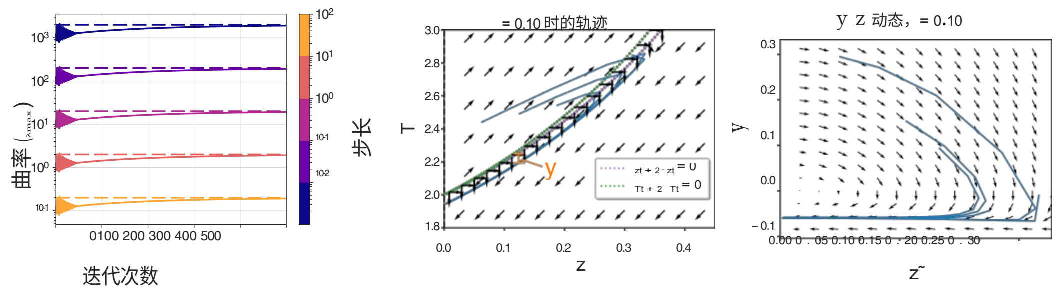


图 2：对于较小的ϵ，双特征值模型显示了不同步长的 EOS 行为（$\epsilon = 5\ 10^{-3}$，左图）。由于初始化时相应的重标定坐标 z˜ 和 $T(0)$ 相同，因此轨迹在缩放之前都是相同的。绘制每一次其他迭代时，我们会看到 z˜ - $T(0)$ 空间中的轨迹停留在空直线（$\tilde{z}, f_{\tilde{z}}(\tilde{z})$）附近，即 $\tilde{z}_{t+2} - \tilde{z}_t = 0$ 的曲线（中间）。将变量变为 $y = T(0) - f_{\tilde{z}}(\tilde{z})$，可以快速集中到一条近乎恒定、小的负 y 曲线上（右图）。

我们可以证明以下关于包含高阶项时 z˜ 和 y 的长时动态定理（附录 B.4）：

定理 2.1. *存在一个* $\epsilon_c > 0$，*使得对于* $E = 0$ *且特征值为* $\{-\epsilon, 1\}$ 的*二次回归模型*，$\epsilon \leq \epsilon_c$。*存在一个邻域* $U \subset \mathbb{R}^2$ *和区间* $[\eta_1, \eta_2]$，*这样对于初始* $\theta \in U$ *和学习率* $\eta \in [\eta_1, \eta_2]$，*模型表现出稳定边缘行为：*

$$2/\eta - \delta_\lambda \leq \lim_{t\to\infty} \lambda_{\max} \leq 2/\eta \tag{18}$$

对于 $O(\epsilon)$*的* δ_λ。

因此，与弹射器相位模型不同，小ϵ 可以证明具有 EOS 行为--其机理可以通过 z˜ - y 坐标变换很好地理解。

3 二次回归模型

3.1 一般模式

虽然等式 1 中定义的模型可以证明存在稳定边缘行为，但需要对 **Q** 的特征值进行调整才能证明。我们可以定义一个更通用的模型，它只需较少的调整就能显示稳定边缘行为。我们将*二次回归模型*定义如下。
给定 P 维参数向量 $\boldsymbol{\theta}$，D *维*输出向量 **z** 的计算公式为

$$\mathbf{z} = \mathbf{y} + \mathbf{G}^{\mathsf{T}} \boldsymbol{\theta} + \frac{1}{2} \mathbf{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}) \tag{19}$$

这里，**y 是**一个 D *维*向量，**G** 是一个 $D\ P$ *维*矩阵，**Q** 是一个 $D\ P\ P$ 维张量，在后两个

指数上对称--也就是说，$\mathbf{Q}(\ ,\ )$ 将两个 P 维向量作为输入，并输出一个 D 维向量验证 $\mathbf{Q}(\boldsymbol{\theta}, \boldsymbol{\theta})_\alpha = \boldsymbol{\theta}\ \boldsymbol{Q}^{\mathsf{T}}{}_\alpha\ \theta$。如果 $\mathbf{Q} = \mathbf{0}$，模型对应于线性化学习（如 NTK 机制）。当 $\mathbf{Q} = \mathbf{0}$ 时，我们得到对 NTK 系统的第一次修正。我们注意到

$$\mathbf{G}_{\alpha i} = \frac{\partial \mathbf{z}_\alpha}{\partial \boldsymbol{\theta}_i}\bigg|_{\theta=0},\ \mathbf{Q}_{\alpha ij} = \frac{\partial\, z^2{}_\alpha}{\partial \boldsymbol{\theta}_i\, \partial \boldsymbol{\theta}_j},\ \rightarrow \mathbf{J} = \mathbf{G} + \mathbf{Q}(\boldsymbol{\theta}, \text{-})\text{、} \tag{20}$$

为 $D \quad \times$ 对于 $D = 1$，我们将恢复方程 1 的模型。在本节的其余部分，我们将研究当 D 和 P 以固定比率 D/P 增加时的极限。

二次回归模型对应于参数变化二阶导数不变的模型，或者说是更复杂的 ML 模型的二阶展开。以前曾对浅层 MLP 的二次展开进行过研究（Bai 和 Lee，2020 年；Zhu 等，2022 年），但我们将提供证据，证明即使是随机的、非结构化的二次回归模型也会导致 EOS 行为。我们注意到，该模型与神经正切层次中的二阶扩展相关，但并不等同（Huang & Yau，2020）（详见附录 A.3）。

3.2 梯度流动力学

我们将重点关注损失平方 $L(\mathbf{z}) = \frac{1}{2}\sum_\alpha \mathbf{z}_\alpha^2$ 的训练。我们首先考虑动态梯度流（GF）下的力学：

$$\dot{\theta} = -\frac{\partial L(\mathbf{z})}{\partial \boldsymbol{\theta}} = -\mathbf{J}^\top z\text{。} \tag{21}$$

我们可以将输出空间 $\mathbf{z}$ 和雅各布 $\mathbf{J}$ 的动态关系写成

$$\dot{z} = J\dot{\theta} = -\mathbf{J}\mathbf{J}^\top z,\ \dot{J} = -\mathbf{Q}(\mathbf{J}^\top z, \cdot) \tag{22}$$

当 $\mathbf{Q} = \mathbf{0}$ 时（线性化/NTK 状态），$\mathbf{J}$ 为常数，动力学随 $\mathbf{z}$ 呈线性关系，并受 $\mathbf{J}\mathbf{J}^\top$ 的特征结构控制，即经验 NTK。在这种情况下，不存在 EOS 行为。

我们感兴趣的是在 GF 下发生渐进锐化的情况。我们可以研究随机初始化的早期 $\mathbf{J}\mathbf{J}^\top$ 最大特征值 λ_{max} 的动态。在附录 C.1 中，我们证明了以下定理：

定理 3.1. *让 $\mathbf{z}$、$\mathbf{J}$ 和 $\mathbf{Q}$ 分别以均值为零、方差为 σ_z^2, σ_J^2, 和 1 的 i.i.d. 元素初始化，其分布对数据和参数空间的旋转不变，并具有有限的第四矩。让 λ_{max} 成为 $\mathbf{J}\mathbf{J}^\top$ 的最大特征值。在大 D 和 P 的极限条件下，D/P 比值固定，初始化时我们有*

$$E[\dot{\lambda}_{max}(0)] = 0,\ E[\ddot{\lambda}_{max}(0)]/E[\lambda_{max}(0)] = \sigma_z^2 \tag{23}$$

其中，E 表示初始化时对 $\mathbf{z}$、$\mathbf{J}$ 和 $\mathbf{Q}$ 的期望。

与 $D = 1$ 的情况很相似，定理 3.1 表明，很容易找到显示渐进锐化的初始化--增加 σ_z 会使锐化更加突出。

3.3 梯度下降动力学

现在我们考虑有限步长梯度下降（GD）动力学。$\boldsymbol{\theta}$的动力学方程为

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\mathbf{J}_t^\top \mathbf{z}_t\text{。} \tag{24}$$

在这种情况下，动态方程可以写成

$$\mathbf{z}_{t+1} - \mathbf{z}_t = -\eta\mathbf{J}_t\mathbf{J}_t^\top\mathbf{z}_t + \frac{1}{2}\eta^2\mathbf{Q}(\mathbf{J}_t^\top\mathbf{z}_t, \mathbf{J}_t^\top\mathbf{z}_t) \tag{25}$$

$$\mathbf{J}_{t+1} - \mathbf{J}_t = -\eta\mathbf{Q}(\mathbf{J}_t^\top\mathbf{z}_t, \cdot)\,. \tag{26}$$

如果 $\mathbf{Q} = \mathbf{0}$，动力学就会简化为二次势能中的离散梯度下降--如果 $\lambda_{max} < 2/\eta$，它就会收敛。

一个直接的问题是：等式 25 中的η^2 什么时候会影响动力学？考虑到与第一项相比，η

和 $\mathbf{z}$ 的幂级数更大，我们可以推测这两项的大小之比 r_{NL} 与 $\mathbf{z}_2$ 和 η 成正比。附录 C.2 中的计算表明，对于随机旋转不变初始化，我们有

$$r_{NL} \equiv \left(\frac{\mathrm{E}[||\frac{1}{2}\eta^2 \, Q(\mathbf{J}_0^\top \mathbf{z}, \mathbf{J}_0^\top \mathbf{z})||_2^2]}{\mathrm{E}[||\eta \mathbf{J}_0 \mathbf{J}_0^\top \mathbf{z}||_2^2]} \right)^{1/2} = \frac{1}{2}\eta\sigma_z D \text{、} \tag{27}$$

这表明，提高学习率会增加动力学偏离 GF 的程度（这是显而易见的），但提高 $||z||$ *也会*增加偏离 GF 的程度。

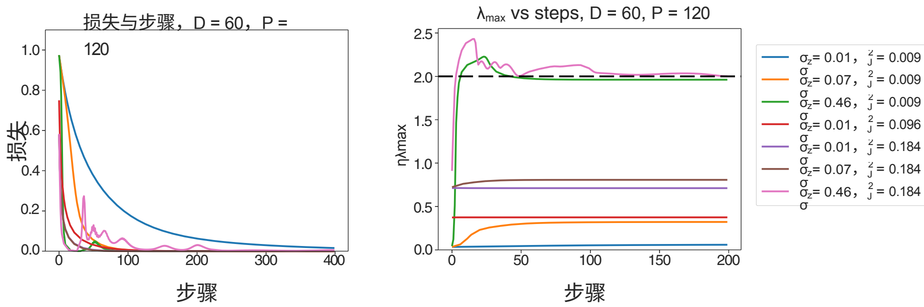


图 3：二次回归模型中的梯度下降动态。随着 **z** 初始化方差 σ^2 的增加，收敛后的曲率 λ_{max} 也在增加。随着锐化使 $\eta\lambda_{max}$ 接近 2，较大的 σ_z 允许非线性效应诱发稳定边缘行为（右图）。结果损失轨迹是非单调的，但仍趋近于 0（左图）。

我们可以从 GD 方程的动力学中看到这种现象（图 3）。这里我们绘制了定理 3.1 中随机初始化的不同轨迹，其中 $D = 60$，$P = 120$，$\eta = 1$。随着 σ_z 的增加，曲率 λ_{max} 也在增加（如定理 3.1 所示），当 σ_z 为 $O(1)$时，动力学是非线性的（如 r_{NL} 所预测），出现了 EOS 行为。这表明方程 25 中的第二项对于稳定 λ_{max} 至关重要。

我们可以通过在多个种子上对不同的 η、D、P、σ_z 和 σ_J 进行初始化，并绘制最终达到的 λ_{max} 的相图，从而更普遍地证实这一点。我们可以通过重新调整参数和初始化来简化绘图。例如

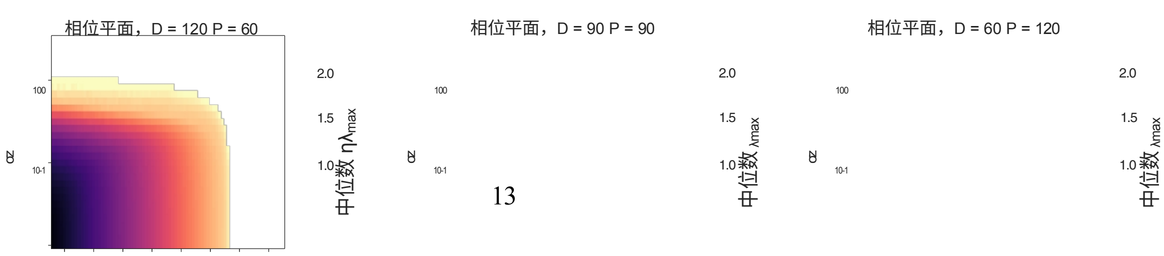
$$\tilde{\mathbf{z}} = \boldsymbol{\eta}\mathbf{z},\ \tilde{\mathbf{J}} = \eta^{1/2} J\text{、} \tag{28}$$

的动力学等价于方程 25 和 26，$\eta = 1$。与方程 8-9 中的 $\tilde{z}T(0)$ 模型一样，重标度坐标中的ηλ $T(0)$ 模型中的公式 8-9，重比例坐标中的λ_{max} 等价于非比例坐标中的$\eta\lambda$ 。 max
我们还可以为 **z** 和 **J** 定义重标度初始化。

$$\sigma_z = \tilde{\sigma}_z / D,\ \sigma_J = \tilde{\sigma}_J / (DP)^{1/4}\text{、} \tag{29}$$

那么我们就有 $r_{NL} = \tilde{\sigma}_z$ ，这样就可以更容易地在（D，P ）对之间进行比较。

利用这一初始化方案，我们可以绘制出 λ_{max} 的 最 终 值 与 $\tilde{\sigma}_z$ 和$\tilde{\sigma}$ 的函数关系图。$\tilde{\sigma}_J$ ，对每对$\tilde{\sigma}_z$ 、$\tilde{\sigma}_J$ ，进行 100 次独立随机初始化（图 4）。我们看到，关键在于 $r_{NL} = \tilde{\sigma}_z$ 是 $O(1)$--对应于初始化附近的渐进锐化和非线性动态。特别是，初始化时的小 $\tilde{\sigma}_J$ 值在 EOS 处收敛，对应的轨迹首先锐化，然后在 $\lambda_{max} = 2/\eta$ 附近稳定下来。大 $\tilde{\sigma}_z$ 和大$\tilde{\sigma}_J$ 动力发散。在很宽的$\tilde{\sigma}_z$ ，有一小段初始$\tilde{\sigma}_J$ ，它们的最终$\lambda_{max} \approx 2/\eta$；这些对应于在EOS附近初始化的模型，它们保持在EOS附近。



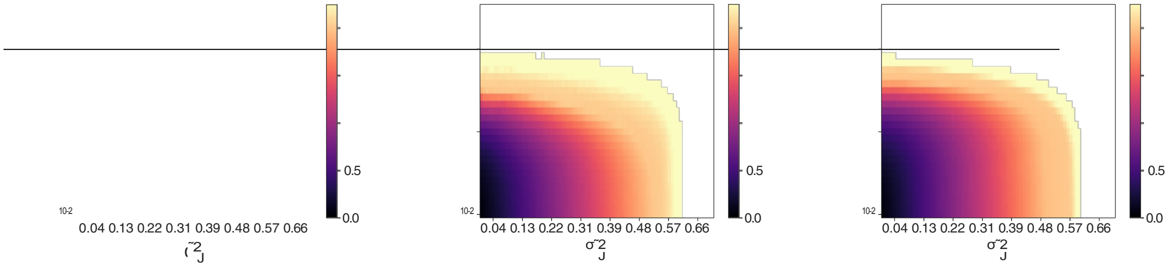
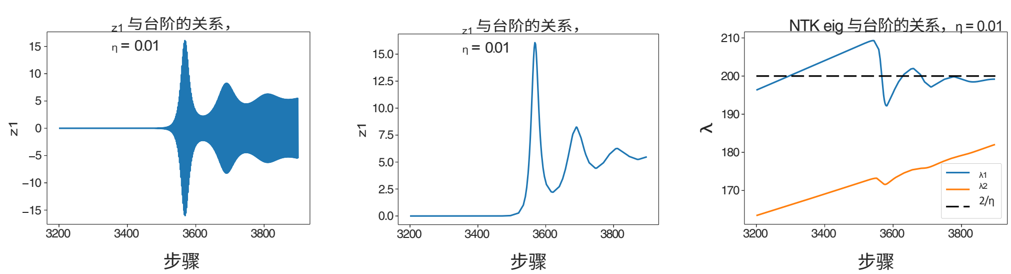


图 4：不同 D 和 P 条件下二次回归模型的 $\tilde{\sigma}_z$ / $\tilde{\sigma}^2_J$ 相平面。对每个 $\tilde{\sigma}_z$，$\tilde{\sigma}_J$ 对模型使用 100 个随机种子进行初始化，并迭代直到收敛。对于每一对 $\tilde{\sigma}_z$，$\tilde{\sigma}^2$，我们绘制 NTK $\mathbf{J}^\top J$ 的中值 $\lambda_{\max}$。对于同时出现锐化和非线性 $\mathbf{z}$ 动态的中间 $\tilde{\sigma}_z$，轨迹趋于收敛，因此 NTK 的 $\lambda_{\max}$ 接近 $2/\eta$（EOS）。

这表明，渐进锐化和稳定边缘并非神经网络模型的独特特征，而可能是高维非线性模型学习的更普遍特性。

4 与现实世界模型的联系

在本节中，我们将考察所提出的模型和所发展的理论对 "真实世界 "模型行为的代表性。根据 Cohen 等人（2022a）的研究，我们在 CIFAR10 的 5000 个示例上使用平方损失训练了一个 2 隐藏层 tanh 网络，学习率为 10^{-2}，这一设置显示了稳定边缘行为。在接近 EOS 开始时，我们使用 Lanczos 方法（Ghorbani 等人，2019；Novak 等人，2019）近似计算了 λ_1 、$\mathbf{J}\mathbf{J}^{\mathrm{T}}$ 的最大特征值及其相应的特征向量 $\mathbf{v}_1$ 。我们使用 $\mathbf{v}_1$ 计算 $z_1 = \mathbf{v}^{\mathrm{T}} \mathrm{z}$，其中 $\mathbf{z}$ 是神经网络函数 f、训练输入 $\mathbf{X}$、标签 $\mathbf{Y}$ 和参数 $\boldsymbol{\theta}$ 的残差向量 $f(\mathbf{X}, \vartheta)\ \mathbf{Y}$。NTK 中的 EOS 行为与 Cohen 等人（2022a）中定义的全 Hessian 的 EOS 行为类似（图 5，左和右）。同样，每隔一步绘制轨迹可以消除高频振荡（图 5，中间）。与 $D = 1$，$P = 2$ 模型不同的是，临界线 $\lambda_{\max} = 2/\eta$ 线有多次交叉。



z_1 ，训练集残差 $f(\mathbf{X}, \boldsymbol{\theta})\ \mathbf{Y}$ 在顶部 NTK 特征模式 $\mathbf{v}_1$ 上的投影，幅度增大并在 0 附近振荡（左图）。每两步绘制动态图可消除高频振荡（中）。最大特征值 λ_1 与稳定边缘多次交叉，但第二大特征值 λ_2 仍低于稳定边缘。

有证据表明，二次回归模型的低维特征可用来解释 EOS 行为的某些方面。我们根据经验通过自动微分计算输出 $f(\mathbf{x}, \boldsymbol{\theta})$ 的二阶导数。我们用 $\mathbf{Q}(\,,\,)$ 表示得到的张量。我们可以使用矩阵-矢量乘法计算矩阵 $\mathbf{Q}_1\ \mathbf{v}_1\ \mathbf{Q}(\,,\,)$ 的频谱，即 $\mathbf{Q}$ 在 $\mathbf{v}_1$ 方向的输出投影，而无需在内存中实例化 $\mathbf{Q}$（图 6 左）。该图显示，从 3200 步到 3900 步（我们绘图的范围），频谱没有太大变化。这说明在显示这些 EOS 动态时，**Q 并**没有发生太大变化。我们还可以看到，**Q 在** $\mathbf{v}_1$ 方向比随机方向要大得多。

让 y 定义为 $y = \lambda_1\ \eta\ 2$。将 z_1 与 *2yz 的*两步动力学关系绘制成图，我们会发现两者非常一致（图 6，中间）。这与我们简化模型中 $\tilde{z}$ 的动力学形式相同。在 $y = \lambda_1\ \eta\ 2$ 的雅各比固定的情况下，迭代方程 25 两次，并剔除 *η 中的*高阶项，也可以发现这一点。这表明，在这种特殊的 EOS 行为中，与我们的简化模型一样，特征值的动态变化比特征基础的旋转更为重要。

y 的动态变化更为复杂；$y_{t+2}\ y_t$ 与 z^2 是反相关的，但并不存在 y 和 z 的低阶函数形式$_1$（

附录 D.1）。我们可以通过绘制 $\eta\, Q^{2}{}_{1}\,(Jz\, v_{11}, Jz\, v_{11})$（来自 $\mathbf{v}_1$ 方向对 z_1 动态的非线性贡献）和 $\lambda\, z_{11}$（线性化贡献）的比率来了解稳定情况，并将其与 y 的动态进行比较（图 6，右）。在最初的锐化过程中，比值很小，但在曲率首次减小前不久，比值变为 $O(1)$。在其余的动态过程中，该比率一直保持为 $O(1)$。这表明，顶部特征模态动力学对自身的非线性反馈对于理解 EOS 动力学至关重要。

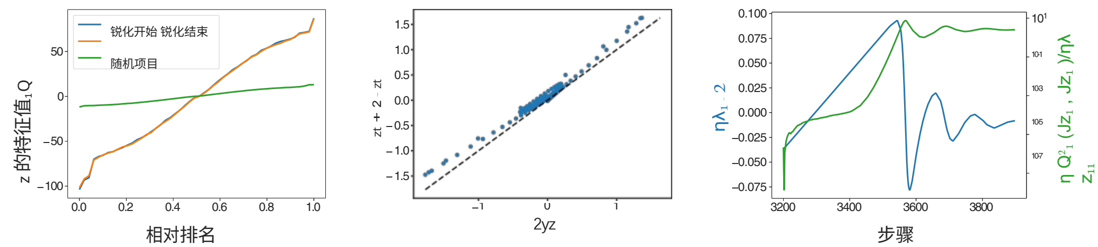


图 6：在 CI- FAR10（左）上训练的 FCN 的稳定边缘动力学过程中，**Q 值**近似恒定。投影到最大特征方向 $\mathbf{v}_1$ （蓝色和橙色）大于投影到随机方向（绿色）。两阶差分 $(z)_{1t+2}$ $(z)_{1t}$ 与 $2z_1$ y （中间）近似，是具有固定特征基的模型的前导阶项。非线性动力学影响 ηQ^2_1 $(Jz\, v_{11}, Jz\, v_{11})$ 在锐化过程中很小，但在顶部特征值减小之前立即变大（右图）--简单模型也是如此。

对于较小的模型，我们可以计算完整的 **Q 值**，然后直接对等式 25 和 26 进行数值积分。这相当于对完整模型进行二次泰勒展开训练。在附录 D.2 中，我们在一个两类 CIFAR 数据集上对一个全连接模型进行了这样的二次展开。在初始化时展开，我们可以看到在早期最大特征值很好地近似于二次模型，但错过了锐化机制（图 7 左）。在更接近锐化机制时，我们看到二次模型捕捉到了 EOS 的一些特征，尤其是第一次交叉（图 7，中），但 $y = 0$ 附近的振荡周期和振荡幅度没有被二次展开正确捕捉到。尽管如此，二次模型还是显示出在 $y = 0$ 上下收敛到一个稳定的双周期，平均值为负值（图 7，右）--这在完整模型和更简单的双参数模型中都可以看到。

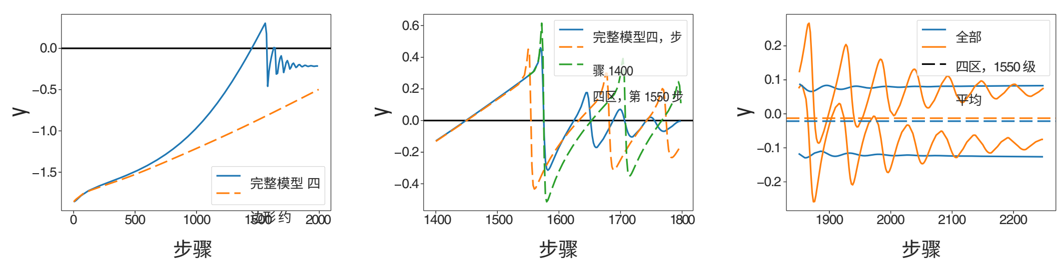


图 7：根据两类 CIFAR 训练的 FCN 模型初始化时的二次展开捕捉了完整模型的早期曲率动态（左图）。在接近第一个 $y = 0$ 交叉点时展开，可以看到两步动态中的多次振荡，但周期和幅度很快就与完整动态不同了（中图）。偶数步（上曲线）和奇数步（下曲线）的轨迹最终趋于稳定，两个模型的平均 y 都不为零（右图）。

5 讨论

5.1 二次回归模型的经验教训

从二次回归模型中学到的主要经验是，渐进锐化（对于 GF 和 GD）和稳定边缘行为（对于 GD）等行为可能是基于梯度的非线性模型高维训练的共同特征。事实上，这些现象可以在

与深度学习模型没有任何联系的简单设置中揭示出来：我们的简化模型对应 1 个数据点和 2 个参数，通过轻度调整就可以证明它显示出 EOS 行为。这与对 CIFAR 模型的分析相结合，表明一般机制可能具有低维描述。

实际模型的二次近似值可以定量地捕捉到 EOS 行为的早期特征（最初回归到 $\lambda_{max} < 2/\eta$），但不一定能捕捉到随后振荡的幅度和周期--这需要更高阶的项（附录 D.2）。然而

二次近似确实正确地描述了许多定性行为，包括 λ_{max} 收敛到在 $2/\eta$ 附近振荡的极限双周期，平均值*低于* 2/η。在简化的双参数模型中，可以通过分析预测收敛时的最终值，事实上我们发现它与 $2/\eta$ 值略有偏差。

本文研究的所有模型都有一个主要特点，即每隔一次迭代（两步动力学）都能极大地帮助我们从理论和经验上理解这些模型。在接近稳定边缘时，顶层特征模式的变化较小。在简化模型中，缓慢的 $\tilde{z}$ 动力学（以及相关的缓慢 T（0）动力学）使得详细的理论分析成为可能；而在 CIFAR 模型中，两步动力学在 z_1 和 $\lambda_{\max}$ 上都是缓慢变化的。对这些微小变化进行定量比较，可能有助于发现在其他系统和情景中解释 EOS 行为的任何普遍机制/典型形式。

5.2 未来的工作

未来工作的一个方向是定量了解大 D 和 P 二次回归模型中的渐进锐化和 EOS 行为。特别是，有可能预测稳定边缘机制中的最终偏差 $2\,\eta\lambda_{\max}$ *与* σ_z、σ_J 和 D/P *的*函数关系。了解高阶项如何影响训练动态也很有用。一种可能是，损失函数高阶导数的少量统计量足以让我们更好地定量理解 $y = 2$ 附近的振荡。

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，特征学习是由 **J** 和 **z** 之间的关系编码的，尤其是 **z** 和 **JJ** 的特征结构之间的关系$^{\mathsf{T}}$。了解 **Q** 如何介导这两个量的动态变化，可以为理解特征学习提供一个定量基础，与现有的理论方法相辅相成（Roberts 等人，2022 年；Bordelon & Pehlevan，2022 年；Yang 等人，2022 年）。

参考资料

Ben Adlam 和 Jeffrey Pennington.高维神经正切核：Triple Descent and a Multi-Scale Theory of Generalization.*第 37 届机器学习国际会议论文集》*，第 74-84 页。PMLR，2020 年 11 月。

Yu Bai 和 Jason D. Lee.超越线性化：关于宽神经网络的四阶和高阶逼近。*学习表征国际会议*，2020 年 3 月。

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal.调和现代机器学习实践与经典偏差-方差权衡。*美国国家科学院院刊》*，116（32）：15849-15854，2019 年 8 月。DOI：10.1073/pnas.1903070116。

Blake Bordelon 和 Cengiz Pehlevan.宽神经网络内核演化的自洽动态场理论》，2022 年 5 月。

Lin Chen、Yifei Min、Mikhail Belkin 和 Amin Karbasi。多重后裔：设计自己的泛化曲线*神经信息处理系统进展》*，第 34 卷，第 8898-8912 页。库兰联合公司，2021 年。

Jeremy Cohen、Simran Kaur、Yuanzhi Li、J. Zico Kolter 和 Ameet Talwalkar。神经网络上的梯度下降通常发生在稳定边缘。*国际学习表征会议*，2022 年 2 月a。

Jeremy M. Cohen、Behrooz Ghorbani、Shankar Krishnan、Naman Agarwal、Sourabh Medapati、Michal Badura、Daniel Suo、David Cardoze、Zachary Nado、George E. Dahl 和 Justin Gilmer。稳定边缘的自适应梯度方法》，2022年7月b。

Pierre Foret、Ariel Kleiner、Hossein Mobahi 和 Behnam Neyshabur。锐度感知最小化，有效提高泛化能力。*国际学习代表会议*，2022 年 4 月。

Behrooz Ghorbani、Shankar Krishnan 和 Ying Xiao.通过黑森特征值密度进行神经网络优化的研究。*第 36 届机器学习国际会议论文集》*，第 2232-2241 页。PMLR，2019 年 5 月。

Niv Giladi、Mor Shpigel Nacson、Elad Hoffer 和 Daniel Soudry。稳定边缘：如何调整超参数以保持神经网络异步训练中的最小值选择？*第八届学习表征国际会议，*2020 年 4 月。

Justin Gilmer、Behrooz Ghorbani、Ankush Garg、Sneha Kudugunta、Behnam Neyshabur、David Car- doze、George Edward Dahl、Zachary Nado 和 Orhan Firat。深度学习模型训练不稳定性的损失曲率视角。*国际学习代表会议，*2022 年 3 月。

Jiaoyang Huang and Horng-Tzer Yau.深度神经网络和神经切线层次的动力学。*第 37 届机器学习国际会议论文集》*，第 4542-4551 页。PMLR，2020 年 11 月。

Arthur Jacot、Franck Gabriel 和 Clement Hongler.神经切线核：神经网络中的收敛与泛化。*神经信息处理系统进展 31》*、
pp.8571-8580.Curran Associates, Inc.，2018 年。

Arthur Jacot、Franck Gabriel 和 Clement Hongler.DNN Hessian 在整个训练过程中的渐近谱。*学习表征国际会议，*2020 年 3 月。

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington.任意深度的宽神经网络在梯度下降过程中演化为线性模型。*神经信息处理系统进展》*第 *32 期，*第 8570-8581 页。Curran Associates, Inc., 2019.

Aitor Lewkowycz、Yasaman Bahri、Ethan Dyer、Jascha Sohl-Dickstein 和 Guy Gur-Ari。深度学习的大学习率阶段：弹射机制。2020年3月

Zhouzi Li, Zixuan Wang, and Jian Li.沿广东轨迹分析锐度：渐进锐化与稳定边缘》，2022 年 7 月。

Behnam Neyshabur、Srinadh Bhojanapalli、David Mcallester 和 Nati Srebro。探索深度学习中的遗传。In *Advances in Neural Information Processing Systems 30*, pp.Curran Associates, Inc., 2017.

Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A. Alemi, Jascha Sohl-Dickstein, and Samuel S. Schoenholz.神经切线：*ArXiv:1912.02803 [cs, stat]*, December 2019.

丹尼尔-A-罗伯茨（Daniel A. Roberts）、矢田翔（Sho Yaida）和鲍里斯-哈宁（Boris Hanin）。*深度学习理论原理》*。DOI: 10.1017/9781009023405.

Lei Wu, Chao Ma, and Weinan E. How SGD Selects the Global Minima in Over-parameterized Learning：动态稳定性视角》。*神经信息处理系统进展》*，第 31 卷。Curran Associates, Inc., 2018.

格雷格-杨张量程序 I：*ArXiv:1910.12478 [cond-mat, physics:math-ph]*, May 2021.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao.张量程序五：通过零点超参数传输调整大型神经网络》，2022 年 3 月。

Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin.用于理解神经网络动态的二次模型》，2022 年 5 月。

A 与其他型号的连接

A.1 单隐层线性网络

考虑一个具有标量输出的单隐层网络：

$$f(\mathbf{x}) = \mathbf{v}\,\mathbf{U}\mathbf{x}^{\mathsf{T}} \tag{30}$$

其中，$\mathbf{x}$ 是长度为 N 的输入向量，$\mathbf{U}$ 是 $K \times N$ 维矩阵，$\mathbf{v}$ 是 K 维向量。我们注意到

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{v}_i\, \partial \mathbf{v}_j} = \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{U}_{ij}\, \partial \mathbf{U}_{kl}} = 0, \quad \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{v}_i\, \partial \mathbf{U}_{jk}} = \delta\, x_{ijk} \tag{31}$$

其中，δ_{ij} 是克罗内克三角洲（Kroenecker delta）。对于固定的训练集，二次导数是常数；因此，单隐层线性网络是第 3 节中研究的二次回归模型。

在单个数据点 $\mathbf{x}$ 的特殊情况下，我们可以计算 **Q** 矩阵的特征向量。设 $(\mathbf{w}, \mathbf{W})$ 为 **Q** 的特征向量，分别代表 $\mathbf{v}$ 和 $\mathbf{U}$ 分量。特征向量方程为

$$\omega \mathbf{w}_i = \mathbf{x}_m\, \delta\, W_{ijjm} \tag{32}$$

$$\omega \mathbf{W}_{jm} = \mathbf{x}_m\, \delta\, w_{iji} \tag{33}$$

简化后，我们得到

$$\omega \mathbf{w} = \mathbf{W}\mathbf{x} \tag{34}$$

$$\omega W = \mathbf{w}\mathbf{x}^{\mathsf{T}} \tag{35}$$

我们有两种情况。第一种情况是 $\omega = 0$。在这种情况下，我们有 $\mathbf{w} = 0$，而 $\mathbf{W}$ 是一个 $\mathbf{x}$ 在其无效空间的矩阵。后一种情况给我们提供了 $M \times N$ 个方程的 M 个约束条件--我们的 $M(N+1)$ 个总特征模中总共有 $M(N-1)$ 个特征模。

如果 $\omega \neq 0$，那么将方程合并，我们就得到了条件：

$$\omega^2\, \mathrm{w} = (\mathbf{x}\cdot\mathbf{x})\mathbf{w} \tag{36}$$

$$\omega^2\, \mathrm{W} = \mathbf{W}\mathbf{x}\mathbf{x}^{\mathsf{T}} \tag{37}$$

由此得出 $\omega = \sqrt{\mathbf{x}\cdot\mathbf{x}}$。从等式 35 中我们可以知道，$\mathbf{W}$ 是低秩的。因此，我们可以猜测以下形式的解

$$\mathbf{W}_{\pm,i} = \pm \mathbf{e}\, \mathbf{x}_i^{\mathsf{T}} \tag{38}$$

其中，$\mathbf{e}_i$ 是 M 坐标向量。这表明我们有

$$\mathbf{w}_{\pm,i} = (\sqrt{\mathbf{x}\cdot\mathbf{x}})\mathbf{e}_i \tag{39}$$

这样我们就得到了最终的 $2M$ 个特征模。

我们还可以分析 $\tilde{J}(\omega_i)$ 的初始值。雅各布的分量可以写成

$$(\mathbf{J})_{vi} \equiv \frac{\partial f(\mathbf{x})}{\partial \mathbf{v}_i} = \mathbf{U}\,\mathbf{x}_{imm} \tag{40}$$

$$(\mathbf{J}_U)_{jm} \equiv \frac{\partial f(\mathbf{x})}{\partial \mathbf{U}_{jm}} = \mathbf{v}\,\mathbf{x}_{jm} \tag{41}$$

根据这一形式，我们可以推断出 $\mathbf{J}$ 与 0 模式是正交的。我们还可以计算守恒量。设 J_+^2 为正特征模式的总权重，J_-^2 为正特征模式的总权重。负特征模式中的权重。直接计算表明

$$\omega^{-1}(J_+^2 - J_-^2) = 2f(\mathbf{x}) \tag{42}$$

这意味着 $E = 0$。

因此，一个数据点上的单隐层线√ear 模型等同于四次方损耗模型，$E = 0$，特征值为 $\pm$ **x - x**。

A.2 连接波德隆和佩赫莱万（2022 年）

由于单隐层线性模型具有常数 **Q**，因此 Bordelon 和 Pehlevan（2022 年）的 F.1 节中的模型属于二次回归类。在第 F.1.1 节方程 67 中，我们可以明确地将其映射为 $D = 1$ 模型。如果我们进行以下识别，则动力学等价于上述具有单一特征值 ω_0 的模型

$$\Delta = \tilde{z},\ H_y = J_0^2,\ \gamma_0 = \sqrt{2}\omega,\ y = -E/2 \tag{43}$$

A.3 连接到第 n

神经切线层次方程（NTH）通过构建控制非线性学习动态的高阶张量的无限序列，扩展了 NTK 动态，以考虑切线核的变化。NTH 方程的三阶截断与二次回归模型相关，但并不相同，我们将在此说明。

三阶 NTH 方程描述了切核 $\mathbf{J}\mathbf{J}^{\mathsf{T}}$ 的变化。考虑 $D \times D \times$ 维核 $\mathbf{K}_3$，其元素由以下公式给出

$$(\mathbf{K})_{3\alpha\beta\gamma} = \frac{\partial z^2_{\alpha}}{\partial\boldsymbol{\theta}_i\,\partial\boldsymbol{\theta}_j} J_{i\gamma} J_{j\beta} + \frac{\partial z^2_{\beta}}{\partial\boldsymbol{\theta}_i\,\partial\boldsymbol{\theta}_j} \mathbf{J}\,\mathbf{J}_{i\gamma j\alpha} \tag{44}$$

其中重复指数相加。在 NTH 中，对于平方损失，NTK $\mathbf{J}\mathbf{J}^{\mathsf{T}}$ 的变化是由

$$\frac{d}{dt}\,\mathbf{J}\mathbf{J}^{\mathsf{T}}_{\alpha\beta} = -\eta(\mathrm{K}3)_{\alpha\beta\gamma}\, z_{\gamma} \tag{45}$$

对于固定的 $\mathbf{Q} = \frac{\partial^2 \mathbf{z}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}$，该方程与二次方程中 NTK 的 GF 方程相同。回归模型。我们注意到，在二次回归模型下，$\mathbf{K}_3$ 并非恒定不变。反之亦然、对于固定的 $\mathbf{K}_3$，$\frac{\partial^2 \mathbf{z}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}$ 也不是常数。因此，这两种方法可以用来构建不同的动力学低阶扩展。

B 2 参数模型

B.1 $\tilde{z}$-$T(0)$ 公式的推导

我们可以利用守恒量 E，只用 $\tilde{z}$ 和 $T(0)$ 来写动力学。在不失一般性的前提下，假设特征值为 1 和 λ，其中 $-1 \leq \lambda \leq 1$。

$$\tilde{z}_{t+1} - \tilde{z}_t = -\tilde{z}\,T_{tt}(0) + \frac{1}{2}(\tilde{z}_t^2)T_t(1) \tag{46}$$

$$T_{t+1}(0) - T_t(0) = -\tilde{z}_t\,(2T_t(1) - \tilde{z}\,T_{tt}(2)) \tag{47}$$

我们将用 $\tilde{z}$ 和 $T(0)$ 来代替 $T(1)$ 和 $T(2)$。回顾一下

$$T(-1) = E + 2\tilde{z} \tag{48}$$

其中 E 在整个动力学过程中都是守恒的（实际上也是地貌的一个属性）。我们将利用这一定义来求解 $T(1)$ 和 $T(2)$。

由于 $P = 2$，我们可以写出 $T(-1) = bT(0) + aT(1)$，系数 a 和 b 对 $\tilde{J}$ 的所有组合都有效。如果 $\tilde{J}(\lambda) = 0$，则 $b = 1 - a$。如果 $\tilde{J}(1) = 0$，则 $1 = \lambda(1 - a) + \lambda^2$ a。解得

$$T(-1) = (1 - a)T(0) + aT(1) \text{ for } a = -\frac{1}{\lambda} \tag{49}$$

对 λ 的限制转化为 $a \in (-1, 1)$。根据守恒量 $E = T(1) - 2\tilde{z}$，我们有

$$T(-1) = E + 2\tilde{z} \tag{50}$$

为了转换动力学，我们需要根据 $T(0)$ 和 $\tilde{z}$ 求解 $T(1)$ 和 $T(2)$。我们有

$$T(1) = \frac{1}{a}(T(-1) + (a-1)T(0)) = \frac{1}{a}(E + 2\tilde{z} + (a-1)T(0)) \tag{51}$$

我们还有

$$T(2) = T(0) + \frac{1-a}{a^2}\left(T(0) - E - 2\tilde{z}\right) \tag{52}$$

这使我们

$$\tilde{z}_{t+1} - \tilde{z}_t = -\tilde{z}_t T(0)_t + \frac{1}{2a}(\tilde{z}_t^2)((a-1)T_t(0) + 2\tilde{z}_t + E) \tag{53}$$

$$T_{t+1}(0) - T_t(0) = -\frac{2}{a}\tilde{z}_t(2\tilde{z}_t + E + (a-1)T_t(0)) + \tilde{z}_t^2 T_t(0) + \frac{1-a}{a^2}\left(T_t(0) - E - 2\tilde{z}_t\right) \tag{54}$$

如果 $\lambda = -\epsilon$（即 $a = \epsilon^{-1}$），我们就可以得到正文中的方程。

Γ^2 的非负性为我们提供了 $\tilde{z}$ 和 T 值的约束条件。对于 $a > 1$（小的负第二特征值），约束条件为

$$T > 2\tilde{z} + E, \quad T > -(2\tilde{z} + E)/a \tag{55}$$

这是一个朝上的圆锥体，顶点位于 $\tilde{z} = E/2$ （图 9 左）。 $\tilde{}$ $E/2$（图 9 左）。对于 $a < 1$ 时，约束条件为

$$-(2\tilde{z}+E)/a < T < 2\tilde{z}+E \tag{56}$$

这是一个侧向圆锥，顶点位于 $\tilde{z} = E/2$（图 9 右）。我们看到，在这种情况下，T 的收敛值是有限的。事实上，在 $E = 0$ 的情况下，除了 $T(0) = 0$ 外，没有收敛。

我们还可以求解无效曲线，即 $\tilde{z}_{t+1} - \tilde{z}_t = 0$（图 9 中的蓝色），或 $T_{t+1}(0) - T_t(0) = 0$（图 9 中的橙色）。$\tilde{z}$ 的零线（$\tilde{z}$，$f_{\tilde{z}}(\tilde{z})$）由以下公式给出

$$f_{\tilde{z}}(\tilde{z}) = \frac{\tilde{z}(2\tilde{z}+E)}{2a - (a-1)\tilde{z}} \tag{57}$$

$T(0)$ 的零线（$\tilde{z}$，$f_T(\tilde{z})$）由以下公式给出

$$f_T(\tilde{z}) = -\frac{(a-1)\tilde{z} - 2a}{(a^2 - a + 1)\tilde{z} - 2a(a-1)}(2\tilde{z}+E) \tag{58}$$

直线 $\tilde{z} = 0$ 也是一条空直线。

对于对称模型 $\epsilon = 1$，零线的结构决定了是否存在渐进锐化。对于 $E = 0$，不存在锐化现象；相位图（图 8，左）证实了这一点，因为 $T_t(0)$ 中的空心线将空间分为两半，一半收敛，另一半不收敛。然而，当 $E = 0$ 时，空轴分开，出现了一个渐进锐化的小区域（图 8，中）。然而，在这种情况下仍然没有稳定边缘行为--没有轨迹聚集在 $\lambda_{\max} = 2/\eta$ 附近的区域（图 8，右）。

B.2 两步动力学

通过公式 12 和 13 的迭代，可以得出两步差分方程。我们有

$$\tilde{z}_{t+2} - \tilde{z}_t = p_0(\tilde{z}_t, \epsilon) + p_1(\tilde{z}_t, \epsilon)T_t(0) + p_2(\tilde{z}_t, \epsilon)T(0)_t^2 + p_3(\tilde{z}_t, \epsilon)T(0)_t^3 \tag{59}$$

$$T(0)_{t+2} - T_t(0) = q_0(\tilde{z}_t, \epsilon) + q_1(\tilde{z}_t, \epsilon)T_t(0) + q_2(\tilde{z}_t, \epsilon)T_t(0)^2 + q_3(\tilde{z}_t, \epsilon)T(0)_t^3 \tag{60}$$

这里的 p_i 和 q_i 是 $\tilde{z}$ 的多项式，最大为 $\tilde{z}$ 的 9 阶和 ϵ 的 6 阶。它们可以显式计算，但我们选

择暂时省略精确形式。

对于固定的 ϵ，我们可以利用卡达诺公式求解 $\tilde{z}$ 两步无效曲线（$\tilde{z}_{t+2}\ \tilde{z}_t = 0$）和 T 无效曲线（$T_{t+2}(0)\ T_t(0) = 0$），将 T 作为 $\tilde{z}$ 的函数。特别是，每个空环方程都有一个解，该解经过 $\tilde{z} = 0$，$T(0) = 2$，与 ϵ 无关。这就是我们要重点研究的解系列。

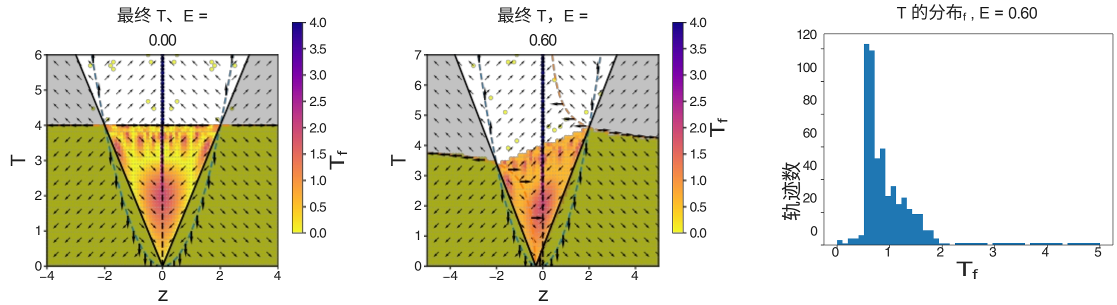


图 8：对称模型的相位图。箭头表示 z 和 T 的变化，灰色区域表示不允许的坐标。从均匀分布的初始化网格开始运行动力学，并记录曲率 $T(0)$ 的最终值。代表 $\tilde{z}_{t+1}\ \tilde{z}_t = 0$（蓝色）和 $T_{t+1}(0)\ T_t(0) = 0^{\text{-}}$（橙色）的零线取决于 E。轨迹显示逐渐锐化，但没有稳定边缘效应（右图）。

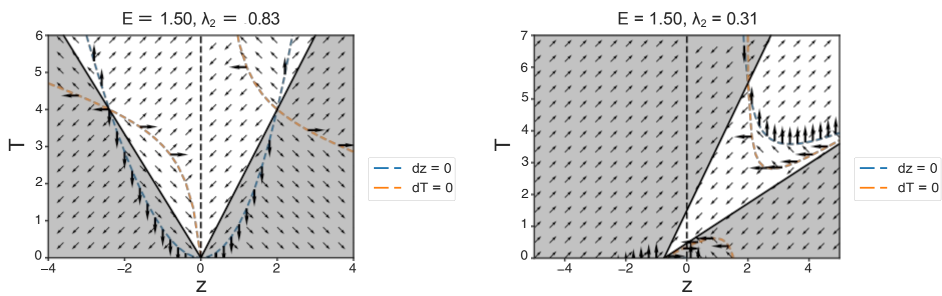


图 9：D=1，P=2 模型的相位平面。灰色区域对应于 $J\tilde{}(\omega)_i^2$ 的正向约束所禁止的参数。对于 $\lambda > 0$，允许区域较小，仅在很小范围内与 $\tilde{z} = 0$ 相交。可以用分析方法求解无效线。

设 $(\tilde{z}, f_{\tilde{z},\epsilon}(\tilde{z}))$为 $\tilde{z}$ 的 nullcline，设 $(\tilde{z}, f_{T,\epsilon}(\tilde{z}))$ 为 $T(0)$ 的 nullcline。我们将证明，作为 $\tilde{z}$ 和 ϵ 的函数，空环的 T 值在 $\tilde{z} = 0, \epsilon = 0$ 附近是可微分的。

无效线由隐式方程定义

$$0 = 6\tilde{z}^3\epsilon - 2T\tilde{z} - 3T\tilde{z}^2(\epsilon - 1) - \frac{T\tilde{z}^3(\epsilon+2)(2\epsilon+1)}{T} + {}^2\tilde{z} + \frac{7}{2}T\tilde{z}^2(\epsilon - 1) + \frac{1}{2}T^{2}\tilde{z}^{3}(9\epsilon - 10\epsilon + 9) - \frac{1}{2}T\tilde{z}^{2}(\epsilon - 1) - \frac{1}{2}T^{3}\tilde{z}^{3}(3\epsilon - 4\epsilon + 3) + O(z^4) \tag{61}$$

$$\begin{aligned}0 = {} & -8\tilde{z}^2\epsilon\ 12\tilde{z}^3(\epsilon)1)\epsilon + 4T\tilde{z}(\epsilon)\ 1) + 2T\tilde{z}^2\,3\epsilon^2 - \epsilon + 3 + 4T\tilde{z}^3(\epsilon\ 1)\epsilon^2 + 4\epsilon + 1 \\ & - 2T^2\tilde{z}(\epsilon)\ 1) - T^2\tilde{z}^2\,7\epsilon^2 - 8\epsilon + 7\ \ T^2\tilde{z}^3(\epsilon\ 1)\,9\epsilon^2 - \epsilon + 9 + T^3\tilde{z}^2\epsilon^2 - \epsilon + 1 \\ & + T^3\tilde{z}^3(\epsilon - 1)\,3\epsilon^2 - \epsilon + 3 + O(\tilde{z})^4\end{aligned} \tag{62}$$

我们暂时省略了高阶项，因为我们要在 $\tilde{z} = 0$ 处微分，以使用隐函数定理。除以 $\tilde{z}$，我们得到方程

$$0 = 6\tilde{z}^2\epsilon - 2T - 3T\tilde{z}(\epsilon - 1) - T\tilde{z}^2(\epsilon+2)(2\epsilon+1) + T^2 + \frac{7}{2}T\tilde{z}(\epsilon - 1) \tag{63}$$

$$+\frac{1}{2}T^2\tilde{z}^2\left(9\epsilon^2-10\epsilon+9\right)-\frac{1}{2}T^3\tilde{z}(\epsilon-1)-\frac{1}{2}T^3\tilde{z}^2\left(3\epsilon^2-4\epsilon+3\right)+O(\tilde{z})^3$$

$$\begin{aligned}0 = {} & -8\tilde{z}\epsilon-12\tilde{z}^2(\epsilon-1)\epsilon+4T(\epsilon-1)+2T\tilde{z}\left(3\epsilon^2-\epsilon+3\right)+4T\tilde{z}^2(\epsilon-1)\left(\epsilon^2+4\epsilon+1\right)\\ & -2T^2(\epsilon-1)-T^2\tilde{z}\left(7\epsilon^2-8\epsilon+7\right)-T^2\tilde{z}^2(\epsilon-1)\left(9\epsilon^2-\epsilon+9\right)+T^3\tilde{z}\left(\epsilon^2-\epsilon+1\right)\\ & +T^3\tilde{z}^2(\epsilon-1)\left(3\epsilon^2-\epsilon+3\right)+O(\tilde{z})^3\end{aligned} \tag{64}$$

我们马上就能看到，对于所有ϵ，$\tilde{z}=0, T=2$ 都能解出这两个方程。设 w（ϵ，$\tilde{z}$，T）和 v（ϵ，$\tilde{z}$，T）分别是等式 63 和 64 的右边。我们有

$$\left.\frac{\partial w}{\partial T}\right|_{(0,0,2)}=2,\quad \left.\frac{\partial v}{\partial T}\right|_{(0,0,2)}=4 \tag{65}$$

在这两种情况下，导数都是可逆的。因此，$f_{\tilde{z},\epsilon}(\tilde{z})$ 和 $f_{T,\epsilon}(\tilde{z})$在 0 的某个邻域内，在 $\tilde{z}$ 和 ϵ 中都是连续可微的。事实上，由于 w 和 v 在所有三个参数中都是解析的，因此 $f_{\tilde{z},\epsilon}(\tilde{z})$ 和 $f_{T,\epsilon}(\tilde{z})$ 也是解析的。

我们可以利用分析性来求解 nullclines 的低阶结构。计算导数值的一种方法是将 nullclines 定义为形式幂级数：

$$f_{\tilde{z}}(\tilde{z})=2+\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{j,k}\,\epsilon^j\,\tilde{z}^k \tag{66}$$

$$f_{T}(\tilde{z})=2+\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} b_{j,k}\,\epsilon^j\,\tilde{z}^k \tag{67}$$

然后，我们可以利用等式 63 和 64 求出数列的前几项。根据这一过程，我们可以得出

$$f_{\tilde{z},\epsilon}(\tilde{z})=2+2(1-\epsilon)\tilde{z}+2\left(1-\epsilon+\epsilon^2\right)\tilde{z}^2+O(\tilde{z})^3 \tag{68}$$

$$f_{T,\epsilon}(\tilde{z})=2-\frac{2-3\epsilon+2\epsilon^2}{1-\epsilon}\tilde{z}+\frac{1}{2}\left(4-\epsilon+4\epsilon^2\right)\tilde{z}^2+O(\tilde{z})^3 \tag{69}$$

两者之间的差值 $f_{\Delta,\epsilon}(\tilde{z})$ 是：

$$f_{\Delta}(\tilde{z})\equiv f_{\tilde{z}}(\tilde{z})-f_{T}(\tilde{z})=-\frac{\epsilon}{1-\epsilon}\tilde{z}-\frac{3}{2}\epsilon\tilde{z}^2+O(\tilde{z})^3 \tag{70}$$

随着 ϵ 的减小，低阶项的空轴距离也会减小。我们可以证明，随着 ϵ 的减小，两者之

间的距离也会减小。$\epsilon=0$ 时的一步动力方程为

$$\tilde{z}_{t+1}-\tilde{z}_t=-\tilde{z}\,T_{tt}(0)+\frac{1}{2}\tilde{z}_t^2 T_t(0) \tag{71}$$

$$T_{t+1}(0)-T_t(0)=-2\tilde{z}\,T_{tt}(0)+z_t T(0)^2_t \tag{72}$$

因此，$\Delta\tilde{z}=2\Delta T$。这意味着，一步和两步无效线都是相同的。由于 $f_{\tilde{z},0}(\tilde{z})=f_{T,0}(\tilde{z})$，且二者关于$\epsilon$都是可微分的，因此我们有：

$$f_{\tilde{z},\epsilon}(\tilde{z})-f_{T,\epsilon}(\tilde{z})=\epsilon f_{\Delta,\epsilon}(\tilde{z}) \tag{73}$$

对于某个函数 $f_{\Delta,\epsilon}(\tilde{z})$，该函数在$\epsilon$和 $\tilde{z}$ 周围 (0, 0) 的邻域中是解析的。

B.3 y 的两步动态变化

在坐标（$\tilde{z}$，y）中定义动力学方程是非常有用的，其中 y 是 "$\tilde{z}$"和 "y "之间的差值。$T(0)$ 和 $\tilde{z}$ nullcline：

$$y\equiv T(0)-f_{\tilde{z},\epsilon}(\tilde{z}) \tag{74}$$

对 $\tilde{z}$ 和 ϵ 的最低阶，我们有

$$y=T(0)-2-2(1-\epsilon)\tilde{z}-2\left(1-\epsilon+\epsilon^2\right)\tilde{z}^2+O(\tilde{z}^3) \tag{75}$$

我们注意到，在 $\tilde{z}=0$ 时，$y=0$ 相当于 $T(0)=2$。对于正的 $\tilde{z}$，$y=0$ 意味着 $T(0)>2$。

我们可以写出 $\tilde{z}$ 和 y 的动态：

$$\tilde{z}_{t+2} \quad \tilde{z}_t = p_0(\tilde{z}_t, \epsilon) + p_1(\tilde{z}_t, \epsilon)(y_t + f_{\tilde{z},\epsilon}(\tilde{z}_t)) + p_2(\tilde{z}_t, \epsilon)(y_t + f_{\tilde{z},\epsilon}(\tilde{z})_t^2) + p_3(\tilde{z}_t, \epsilon)(y_t + f_{\tilde{z},\epsilon}(\tilde{z}))_t^3 \tag{76}$$

我们知道，这个方程的右边在 z˜、ϵ和（微不足道的）y 中也是解析的。通过计算 f 的多重连续导数，我们可以写出

$$\tilde{z}_{t+2} - \tilde{z}_t = 2y_t\,\tilde{z}_t + y_t^2\,\tilde{z}\,\mathrm{f}_{t1,\epsilon}(\tilde{z}_t, y_t) + y_t\,\tilde{z}\,\mathrm{f}^2_{2,\epsilon}(\tilde{z})_t \tag{77}$$

这里，$f_{1,\epsilon}$ 和 $f_{2,\epsilon}$ 在 0 附近的 $\tilde{z}$、ϵ 和 y 中是解析的。这意味着我们

有边界

$$|f_{1,\epsilon}(\tilde{z}, y)| < F_1,\ |f_{2,\epsilon}(\tilde{z}, y)| < F_2 \tag{78}$$

为 $(\tilde{z}, \epsilon, y) \in [\tilde{z}_d, \tilde{z}]_d \times [0, \epsilon]_d \times [y_d, y_d]$ 对于一些非负常量 F_1 和 F_2。注意，这个约束与 ϵ 无关。

现在我们考虑 y 的动态变化：

$$y_{t+2} - y_t = T_{t+2}(0) - T_t(0) - f_{\tilde{z},\epsilon}(\tilde{z}_{t+2}) + f_{\tilde{z},\epsilon}(\tilde{z})_t \tag{79}$$

由于 $\lim_{\tilde{z}\to 0, y\to 0} \tilde{z}_{t+2} = 0$，$f_{\tilde{z},\epsilon}(\tilde{z}_{t+2})$ 在 $(0, 0, 0)$ 的某个邻域内是解析的。因此 $y_{t+2} - y_t$ 也是解析的。代入可得

$$\begin{aligned} y_{t+2} - y_t = {} & q_0(\tilde{z}_t, \epsilon) + q_1(\tilde{z}_t, \epsilon)[y + f_{\tilde{z},\epsilon}(\tilde{z})] + q_2(\tilde{z}_t, \epsilon)[y + f_{\tilde{z},\epsilon}(\tilde{z})]^2 + q_3(\tilde{z}_t, \epsilon)[y + f_{\tilde{z},\epsilon}(\tilde{z})]^3 \\ & - f_{\tilde{z},\epsilon}(\tilde{z}_t + 2y_t\tilde{z}_t + y_t^2\tilde{z}\, f_{t1,\epsilon}(\tilde{z}_t, y_t) + y_t\tilde{z}_t^2 f_{2,\epsilon}(\tilde{z}_t)) + f_{\tilde{z},\epsilon}(\tilde{z})_t \end{aligned} \tag{80}$$

如果我们把 $f_{\tilde{z},\epsilon}(\tilde{z}) = f_{T,\epsilon}(\tilde{z}) + \epsilon f_{\Delta,\epsilon}(\tilde{z})$写出来，那么我们就可以写出：

$$\begin{aligned} y_{t+2} - y_t = {} & q_0(\tilde{z}_t, \epsilon) + q_1(\tilde{z}_t, \epsilon)[f_{T,\epsilon}(\tilde{z})] + q_2(\tilde{z}_t, \epsilon)[f_{T,\epsilon}(\tilde{z})]^2 + q_3(\tilde{z}_t, \epsilon)[f_{T,\epsilon}(\tilde{z})]^3 \\ & 2q_2(\tilde{z}_t, \epsilon)[f_{T,\epsilon}(\tilde{z})(y + \epsilon f_{\Delta,\epsilon}(\tilde{z}))] + 3q_3(\tilde{z}_t, \epsilon)[(f_{T,\epsilon}(\tilde{z}))(y + \epsilon f_{\Delta,\epsilon}(\tilde{z}))^2 + (f(\tilde{z}))_{T,\epsilon}^2(y + \epsilon f_{\Delta,\epsilon}(\tilde{z}))] \\ & q_0(\tilde{z}_t, \epsilon) + q_1(\tilde{z}_t, \epsilon)[y + \epsilon f_{\Delta,\epsilon}(\tilde{z})] + q_2(\tilde{z}_t, \epsilon)[y + \epsilon f_{\Delta,\epsilon}(\tilde{z})]^2 + q_3(\tilde{z}_t, \epsilon)[y + \epsilon f_{\Delta,\epsilon}(\tilde{z})]^3 \\ & - f_{\tilde{z},\epsilon}(\tilde{z}_t + 2y_t\tilde{z}_t + y_t^2\tilde{z}\, f_{t1,\epsilon}(\tilde{z}_t, y_t) + y_t\tilde{z}_t^2 f_{2,\epsilon}(\tilde{z}_t)) + f_{\tilde{z},\epsilon}(\tilde{z})_t \end{aligned} \tag{81}$$

根据 nullclines 的定义，前四项消失了。再次利用 nullclines 的可微性，以及 $f_{1,\epsilon}$ 和 $f_{2,\epsilon}$，我们可以用展开式重写动力学：

$$y_{t+2} - y_t = -2(4 - 3\epsilon + 4\epsilon^2)y_t\tilde{z}_t^2 - 4\epsilon\tilde{z}_t^2 + y^2\tilde{z}_t^2 g_{1,\epsilon}(\tilde{z}_t, y_t) + \epsilon\tilde{z}_t^3 g_{2,\epsilon}(\tilde{z})_t \tag{82}$$

这里，$g_{1,\epsilon}$ 和 $g_{2,\epsilon}$ 在 $\tilde{z}$、y 和 ϵ 中为零附近的解析。我们有以下边界

$$|g_{1,\epsilon}(\tilde{z}, y)| < G_1,\ |g_{1,\epsilon}(\tilde{z}, y)| < G_2 \tag{83}$$

为 $(\tilde{z}, \epsilon, y) \in [\tilde{z}_d, \tilde{z}]_d \times [0, \epsilon]_d \times [y_d, y_d]$ 对于一些非负常量 G_1 和 G_2。这个约束也与ϵ 无关。

我们可以用下面的 Lemma 来概括这些界限：

定理 B.1. *定义 $y = T - f_{\tilde{z}}(\tilde{z})$。$\tilde{z}$ 和 y 的两步动力学关系如下*

$$\tilde{z}_{t+2} - \tilde{z}_t = 2y_t\tilde{z}_t + y_t^2\tilde{z}\, f_{t1,\epsilon}(\tilde{z}_t, y_t) + y_t\tilde{z}^2 f_{2,\epsilon}(\tilde{z})_t \tag{84}$$

$$y_{t+2} - y_t = -2(4 - 3\epsilon + 4\epsilon^2)y_t\tilde{z}^2 - 4\epsilon\tilde{z}^2 + y^2\tilde{z}^2 g_{1,\epsilon}(\tilde{z}_t, y_t) + \epsilon\tilde{z}^3 g_{2,\epsilon}(\tilde{z}_t, y)_t \tag{85}$$

其中 $f_{1,\epsilon}, f_{2,\epsilon}, g_{1,\epsilon}, g_{2,\epsilon}$ 都在 $\tilde{z}$、y 和 ϵ 中解析。此外，存在正的 $\tilde{z}_c, y_c$, 和 ϵ_c 这样

$$|f_{1,\epsilon}(\tilde{z}, y)| < F_1,\ |f_{2,\epsilon}(\tilde{z}, y)| < F_2,\ |g_{1,\epsilon}(\tilde{z}, y)| < G_1,\ |g_{1,\epsilon}(\tilde{z}, y)| < G_2 \tag{86}$$

for all $(\tilde{z}, \epsilon, y) \in [\tilde{z}_d, \tilde{z}_d] \times [0, \epsilon_d] \times [y_d, y_d]$, 其中 F_1, F_2, G_1, 和 G_2 都是非负常量。

我们可以利用这个定理来分析小固定ϵ、小初始化 $\tilde{z}$、y 的动态。

B.4 定理 B.3 的证明

利用 Lemma B.1，$\tilde{z}$ 和 y 的动态变化可写成

$$\tilde{z}_{t+2} - \tilde{z}_t = 2y_t \tilde{z}_t + y_t^2 \tilde{z}_t f_{1,\epsilon}(\tilde{z}_t, y_t) + y_t \tilde{z}_t^2 f_{2,\epsilon}(\tilde{z}_t) \tag{87}$$

$$y_{t+2} - y_t = -2(4 - 3\epsilon + 4\epsilon^2) y_t \tilde{z}_t^2 - 4\epsilon \tilde{z}_t^2 + y_t^2 \tilde{z}_t^2 g_{1,\epsilon}(\tilde{z}_t, y_t) + \epsilon \tilde{z}_t^3 g_{2,\epsilon}(\tilde{z}_t, y_t) \tag{88}$$

让 $\epsilon < \epsilon_d$。然后，我们就可以利用 Lemma B.1 的约束条件来控制高阶项对动力学的贡献：

定理 B.2. *给定常数 $A > 0$ 和 $B > 0$，存在 $\tilde z_c$ 和 y_c，使得对于 $\tilde z \in [0, 2\tilde z_c]$、$y \in [-y_c, y_c]$，我们就有了边界：*

$$|y\tilde z^2 f_{1,\epsilon}(\tilde z, y) + y\tilde z^2 f_{2,\epsilon}(\tilde z)| \le A|2y\tilde z| \tag{89}$$

$$|y^2\tilde z^2 g_{1,\epsilon}(\tilde z, y)| \le \frac{B}{8}|2(4 - 3\epsilon + 4\epsilon^2)y\tilde z^2| \tag{90}$$

$$|\epsilon\tilde z^3 g_{2,\epsilon}(\tilde z, y)| \le \frac{B}{4}|4\epsilon\tilde z^2|4\epsilon\tilde z \tag{91}$$

证明。 我们从以下分解开始：

$$|y\tilde z^2 f_{1,\epsilon}(\tilde z, y) + y\tilde z^2 f_{2,\epsilon}(\tilde z)| \le |y\tilde z^2 f_{1,\epsilon}(\tilde z, y)| + |y\tilde z^2 f_{2,\epsilon}(\tilde z)| \tag{92}$$

根据 Lemma B.1，存在一个区域 $[\tilde z_d, \tilde z_d] \times [0, \epsilon] \times [y_d, y_d]$ 其中 $f_{1,\epsilon}$, $f_{2,\epsilon}$, $g_{1,\epsilon}$，和 $g_{2,\epsilon}$ 的大小分别以 F_1, F_2, G_1, 和 G_2 为界。

$$|y\tilde z^2 f_1(\tilde z, y) + y\tilde z^2 f_2(\tilde z)| \le F_1 y\tilde z^2 + F_2 y\tilde z^2 \tag{93}$$

$$|y^2\tilde z^2 g_1(\tilde z, y)| \le G_1 y^2\tilde z^2 \tag{94}$$

$$|\tilde z^3 g_2(\tilde z, y)| \le G_2\tilde z^3 \tag{95}$$

定义 $\tilde z_c$ 和 y_c 为

$$y_c = \min(A/F_1, B/2G_1, y_d), \quad \tilde z_c = \min(A/2F_2, B/2G, y)_{2c} \tag{96}$$

立即得出所需的边界。 □

我们考虑初始化（$\tilde z_0, y_0$），这样 $\tilde z_0$ $\tilde z_c$ 和 y_0 y_c，以及 y_0 $\tilde z^2$。有了定理 B.2，我们就可以对动力学进行分析。在第一阶段，$\tilde z$ 上升，y 下降。当 y 首次变为负值--达到 $O(\epsilon)$时，第一阶段结束。在第二阶段，$\tilde z$ 下降，y 保持负值且为 $O(\epsilon)$。

B.4.1 第一阶段

让 t_{sm} 成为这样的时间：对于 $t \le t_{sm}$，$\tilde z_t \le 2\tilde z_0$。（对于 $t \le t_{sm}$，利用 Lemma B.2，$\tilde z$的变化可以通过以下方式从下往上限定：$\tilde z \le 2\tilde z_0$。

$$\tilde z_{t+2} - \tilde z_t \ge 2y_t\tilde z_t(1 - A) \tag{97}$$

因此，在初始化时，$\tilde z$ 是递增的。直到 y_t 变成负值，或 $\tilde z_t \ge$ 时，它一直保持递增状态。$2\tilde z_0$。我们要证明，在 $\tilde z_t \ge 2\tilde z_0$ 之前，y_t 变成负值。

对于任意 $t \le t_{sm}$，Lemma B.2 给出了 $y_{t+2} - y_t$ 的如下上界：

$$y_{t+2} - y_t \le -(8 - B)y_t\tilde z_t^2 - (4 - B)\epsilon\tilde z_t^2 \tag{98}$$

假设 t_- 是 y_t 首次变为负值的时间。由于 $\tilde z_t$ 在 $t \le t_-$ 时是递增的，我们有

$$y_{t+2} - y_t \le -(8 - B)y_t\tilde z_0^2 - (4 - B)\epsilon\tilde z_0^2 \tag{99}$$

由此我们可以得出以下关于 y_t 的约束条件：

$$y_t \le y_0 e^{-(8-B)\tilde z_0^2 t} \tag{100}$$

对 $t \le t_-$ 和 $t \le t_{sm}$ 有效。

现在我们将证明 $t_{\_} < t_{sm}$。假设 $t\ t_{sm} \leq \_$。那么在 $t_{sm} + 2$ 时，$\tilde{z}_{t\,sm+2} > 2\tilde{z}_0$ 第一次出现。将方程 97 中的约束条件相加，我们得到

$$\tilde{z}_{t\,sm+2} - \tilde{z}_0 \leq \sum_{t=0}^{tsm} 2y_t \tilde{z}_t (1 + A) \leq 4\tilde{z}_0 (1 + A) \sum_{t=0}^{tsm} y_t \tag{101}$$

其中第二个约束来自 t 的定义$_{sm}$。利用我们对 y 的约束$_t$，我们得到

$$\tilde{z}_{tsm+2} - \tilde{z}_0 \leq 4\tilde{z}_0 (1 + A) \sum_{s=0}^{tsm} y_0 e^{-(8-B)\tilde{\epsilon}\, 2s} \leq \frac{(1 + A)}{2} \frac{\tilde{z}_0}{y_0} \tag{102}$$

由于 $y_0 \le \tilde{z}_0^2$, $\tilde{z}_{t_{sm}+2} \le 2\tilde{z}_0$。然而，根据假设 $\tilde{z}_{t_{sm}+2} \ge 2\tilde{z}_0$。我们得出一个矛盾；$t_{sm}$ 不小于或等于 t_-。

有三种可能：第一种可能是 t_- 定义良好，$t_- < t_{sm}$。另一种可能是 t_- 定义不佳，即 y_t 从未变为负数。在这种情况下，我们得出的边界

因此，利用等式 100，存在某个时间 t_e，此时 $y_{t_e} < (4-B)\epsilon\tilde{z}_0^2$。

那么，根据公式 99，我们可以得出 $y_{t_e+2} < 0$。因此，我们得出结论，t_- 是有限的，且小于 t_{sm}。

由于定义良好的值 $t_- < t_{sm}$，所以当 y 第一次变为负值时，$\tilde{z}_{t_-} \le 2\tilde{z}_0$。这意味着我们可以在下一阶段开始时继续应用 Lemma B.2 中的约束。在 $t = t_- - 2$ 时，应用 Lemma B.2 和 $\tilde{z}_{t_-} \le 2\tilde{z}_0$，我们有

$$y_{t_-} - y_{t_- -2} \ge -4(8+B)y_{t_- -2}\tilde{z}_0^2 - 4(4+B)\epsilon\tilde{z}_0^2 \tag{103}$$

得出 $y_{t_-} \ge -4(4+B)\epsilon\tilde{z}_0^2$。第一阶段到此结束。总结如下

$$-4(4+B)\epsilon\tilde{z}_0^2 < y_{t_-} \le 0, \tilde{z}_{t_-} \le 2\tilde{z}_0 \tag{104}$$

B.4.2 第二阶段

现在考虑动力学的第二阶段。我们将证明 y 保持负值且为 $O(\epsilon)$，$\tilde{z}$ 下降为 0。当 $y \ge -y_0$ 时，由 Lemma B.2 可知

$$\tilde{z}_{t+2} - \tilde{z}_t \le (1-A)2y_t\tilde{z}_t \tag{105}$$

因此，只要 $-y_0 \le y < 0$，$\tilde{z}_t$ 就会递减。如果随后的 t 都是如此，则 $\tilde{z}_0$ 将趋近于 0。

现在我们将证明 y 仍为负数且为 $O(\epsilon)$，从而结束证明。让 $y^* = -\epsilon \frac{\epsilon}{2-(3/2)\epsilon+2\epsilon^2}$.

我们可以将 y 的动力学方程改写为

$$y_{t+2} - y_t = -2(4-3\epsilon+4\epsilon^2)\tilde{z}_t^2(y_t - y^*) + y_t^2\tilde{z}_t g_1^2(\tilde{z}_t, y_t) + \tilde{z}_t g_2^3(\tilde{z}_t, y)_t \tag{106}$$

将 Lemma B.2 应用于高阶项，我们得到

$$y_{t+2} - y_t \le -2(4-3\epsilon+4\epsilon^2)\tilde{z}_t^2(y_t - y^*) + B(|y_t|+\epsilon)\tilde{z}_t^2 \tag{107}$$

$$y_{t+2} - y_t \ge -2(4-3\epsilon+4\epsilon^2)\tilde{z}_t^2(y_t - y^*) - B(|y_t|+\epsilon)\tilde{z}_t^2 \tag{108}$$

只要 $|y_t| < y_c$，这些不等式就是有效的。

在 t_- 时，$y^* < y_t < 0$。当 $y^* < y_t < 0$ 时，$|y_t| \le |y^*|$。注意$\epsilon < 2|y^*|$。由等式 107、我们有

$$y_{t+2} - y_t \le -2(4-3\epsilon+4\epsilon^2)\tilde{z}_t^2(y_t - y^*) + B(-y_t+\epsilon)\tilde{z}_t^2 \tag{109}$$

根据这个不等式，我们可以得出结论

$$y_{t+2} \le (1 - 2(4-3\epsilon+4\epsilon^2)\tilde{z}_t^2 - B)y_t + \tilde{z}_t^2[2(4-3\epsilon+4\epsilon^2)y^* + B\epsilon] \tag{110}$$

如果 $B < 1$，那么两个项都是负数。我们可以得出结论：如果 $y^* < y_t < 0$，则 $y_{t+2} < 0$。事实上，从最后一项我们可以得出结论：$y_{t+2} < -4\epsilon\tilde{z}_t^2$。

现在我们必须证明，当 $y^* < y_t < 0$ 时，y_{t+2} 不会变得太负（即小于$-y_c$）。利用等式 108，我们可以得出

$$y_{t+2} > y^*(1 + 3B\tilde{z}_0^2) \text{ if } y_t > y^*$$

(111) 这就是说，如果 y_t 开始大于 y^*，那么在下一步时，y_0^* 至多比 y^* 低 $3Bz^{-2}$ y 。因为 $B < 1, y_{t+2} > -y_c$ if $y^* < y_t < 0$.

最后，我们将证明，如果 $y^* (1 + 3B/(8 B)) < y_t < y^*$，则 $y^* (1 + 3B/(8 B)) < y_{t+2} < 0$。由于 y_{t-+2} 符合这一条件，我们可以得出结论：对于所有 $t > t\_$，y_t 均为负数，其大小自下而上受 $y^* (1 + 3B/(8 - B))$ 约束，并完成证明。

我们首先要证明 $y^* (1 + 3B/(8 B)) < y_t$ 意味着 $y^* (1 + 3B/(8 B)) < y_{t+2}$。让 $y_t = (1 + \delta_t)y^*$，因为 $\delta_t < 3B/(8 B)$。我们将证明 $\delta_{t+2} < 3B/(8 B)$。利用公式 108，我们可以得出

$$y_{t+2} \geq (1 + \delta_t)y^* - 8z^{-2}_t \delta y_t^* - Bz^{-2}_t (\epsilon - (1 + \delta_t)y)^* \tag{112}$$

将 $y_{t+2} = (1 + \delta_{t+2})y^*$ 代入，并将两边除以 y^*，得到

$$\delta_{t+2} - \delta_t \leq -(8 - B)\tilde{z}^2_t \delta^2_t + 3B\tilde{z}^2_t \tag{113}$$

如果 $\delta_t < 3B/(8 - B)$，那么我们就可以得到 $\delta_{t+2} < 3B/(8 - B)$。

最后，我们将证明 $0 < \delta_t < 3B/(8 - B)$意味着 $\delta_{t+2} > -1$ - 即 $[1 + 3B/(8 - B)]$。
$B)]y^* < y_t < y^*$ 意味着 $[1 + 3B/(8 - B)]y^* < y_{t+2} < 0$。 公式 107 意味着

$$y_{t+2} \leq (1 + \delta_t)y^* - 8\tilde{z}^2_t \delta y_t^* + B\tilde{z}^2_t (\epsilon - (1 + \delta_t)y)^* \tag{114}$$

这使我们

$$\delta_{t+2} - \delta_t \geq -(8 - B)\tilde{z}^2_t \delta^2_t - 3B\tilde{z}^2_t \tag{115}$$

如果 $\delta_t > 0$ 意味着

$$\delta_{t+2} > -3B\tilde{z}^2_t \tag{116}$$

如果 $3B\tilde{z}^2_0 < 1$，那么 $\delta_{t+2} > -1$。这意味着，如果 $[1 + 3B/(8 - B)]y^* < y_t < y^*$，则 $y_{t+2} < 0$。
最后，我们选择 B 和 $\tilde{z}_0$ 来保证收敛性。选择 $\tilde{z}^2_0 < 3/7$，并且
选择 $B < \frac{1}{2}$。总之，我们在第二阶段所展示的是

- 在阶段开始时（时间 t_-），$y^* < y_t < 0$。
- 如果 $y^* < y_t < 0, t > t_-$, $y^* (1 + 3B\tilde{z}^2_0) < y_{t+2} < -4\epsilon\tilde{z}^2_t$。
- 如果 $[1 + 3B/(8 - B)]y^* < y_t < y^*$, $t > t_-$, $[1 + 3B/(8 - B)]y^* < y_{t+2} < 0$。

通过选择 $\tilde{z}_0$ 和 B，我们知道 $[1 + 3B/(8 - B)]y^* < y^* (1 + 3B\tilde{z}^2_0)$.因此
的整个轨迹是由这些区域决定的，而 $[1 + 3B/(8 - B)]y^* < y_t < 0$
对于所有 $t > t_-$。此外，我们知道至少每两步一次，$y_t < -4\epsilon\tilde{z}^2_t$。这意味着
$\tilde{z}_t$ 的动态变化可以通过以下公式从上而下地限定

$$\tilde{z}_{t+2} - \tilde{z}_t \leq -2\epsilon^2 \tilde{z}^4_t \tag{117}$$

由此我们可以得出结论，$\tilde{z}_t$ 趋近于 0。

因此，对于 $\tilde{z}_0 \leq \tilde{z}_c$, $y_0 \leq y_c$, 和 $y_0 \leq \tilde{z}^2_0$ 的任何正初始化，我们有

$$\lim_{t\to\infty} \tilde{z}_t \to 0, \lim_{t\to\infty} y = -y_f \tag{118}$$

其中 $y_f = O(\epsilon)$

。

现在我们可以证明定理 2.1 的陈述。给定一个 ϵ ϵ_c 的模型，在 $\bar{\theta}\bar{\eta}$ 空间和 $\tilde{z}$ $\bar{y}$ 空间之间有一个连续的映射。由于 $\tilde{z}$ $\bar{y}$ 空间中的某个邻域表现出稳定边缘行为（$T_t(0)$ 收敛到 2 的 $O(\epsilon)$ 以内），因此该邻域的反像是θ $\bar{\eta}$ 空间中一个表现出稳定边缘行为的邻域。证明到此为止。

B.5 低阶动力学

为了预测 y 的最终值，并理解向定点的收敛，我们可以研究 $\tilde{z}$ 和 y 的低阶动态方程：

$$\tilde{z}_{t+2} - \tilde{z}_t = 2y_t \tilde{z}_t \tag{119}$$
$$y_{t+2} - y_t = -2(4 - 3\epsilon + 4\epsilon^2)y_t \tilde{z}^2 - 4\epsilon\tilde{z}^2 \tag{120}$$

t t

对于这些简化的动力学，我们可以证明如下：

定理 B.3. *对于等式 119 和 120 所定义的动力学，对于* ϵ 1，*对于正初始化* $\tilde{z}_0$ 1，y_0 1，*附加约束* $-\epsilon\log(\epsilon)$ $16\tilde{z}_0^2$ *和* $y_0 < 2\tilde{z}_0^2$，*我们有*

$$\lim_{t\to\infty}\tilde{z}_t = 0, \lim_{t\to\infty}y_t = -\epsilon/2 + O(\epsilon)^2 \tag{121}$$

*证明*证明区分了时间演化的两个阶段：

- 第 1 阶段：$\tilde{z}$ 开始为正值并上升，y 开始为正值并下降。在该阶段结束时，我们希望 $\tilde{z}_t \le 2\tilde{z}_0$，$y$ 为负值，但以 $-16\tilde{z}_0^2\epsilon$ 为界。
- 第 2 阶段：$\tilde{z}$ 缓慢减小，而 y 则（相对地）迅速稳定在定点上，误差可达 $O(\epsilon^2)$.

让 ϵ 1。考虑一个初始化（$\tilde{z}_0, y_0$），其中两个变量都是正数，这样 $\tilde{z}_0$ 1、$\epsilon\log(\epsilon)\tilde{z}_0^{-2}$，而 y_0 $\tilde{z}_0^{-2}$。从等式 119 和 120 中我们可以看出，y 的动态变化将取决于这两个术语的平衡。

最初，$\tilde{z}$ 增大，y 减小。我们假设 $\tilde{z}$ 固定不变，分析 y 的动态变化，然后计算修正量。

第 1 阶段。在初始化时，由于假设 $\epsilon\tilde{z}$……，动力学中的第一项占主导地位。$y_t \tilde{z}_t^2$。由于 $\tilde{z}_0^2$ 1，y 最初以指数形式递减，其衰减率从上而下限定为 $8\tilde{z}_0^2$。因此，在 $\log(-\epsilon/y_0)/8\tilde{z}_0^2$ 步内，$y < \epsilon$。

此时，y 的变化率至少为 $-4\epsilon\tilde{z}_0^2$。因此，在不超过 $1/4\tilde{z}_0^2$ 的情况下，步，y 变成负数。设 t_- 为 y 第一次变为负值的时间。我们注意到 $y_{t_-} \ge$ 根据这一分析，$-4\epsilon\tilde{z}_0^2$ - 等式 120 中第一项的大小小于 y_t，因此如果 y_t 为正值，则 y_{t+2} 的最小值为 $-4\epsilon\tilde{z}_0^2$。

现在我们可以理解由于 $\tilde{z}$ 的变化而产生的修正了。我们注意到，$e^{-8\tilde{z}_0^2 t}$ 是 y 的上限--因为 $\tilde{z}$ 是递增的，而 $-4\epsilon\tilde{z}^2$ 对 y 的减小速度要快于从第一次指数衰减开始的指数衰减。

项。由于 $\tilde{z}$ 是递增的，因此只要 y 保持正值（$t < t$），$y \ge e^{-8\tilde{z}^2 t}$ 只要 y 保持正值（$t < t_{sm}$）。让 t 是一个时间使得 $\tilde{z}_{t_{sm}} < 2\tilde{z}_0$ 我们可以约束 $\tilde{z}$ 的变化 $_t$ 为 $t < t_{sm}$.我们知道 $y_t \ge y_0 e^{-8\tilde{z}_0^2 t}$.
$\tilde{z}$ 的变化可由以下公式限定

$$\tilde{z}_{t_{sm}} - \tilde{z}_0 \le \sum_{t=0}^{t_{sm}} 2z y_{tt} \le 4\tilde{z}_0 \sum_{t=0}^{t_{sm}} y_t \le 4\tilde{z}_0 y_0 \sum_{t=0}^{t_{sm}} e^{-8\tilde{z}_0^2 t} \le \frac{1}{2} \frac{y_0}{\tilde{z}_0} \tag{122}$$

如果 $y_0 < 2\tilde{z}_0^2$，那么只要 y 的约束是正确的，约束的成立与 t_{sm} 的值无关。我们知道，在时间 t_- 之前，y 的约束是正确的；因此，$t_{sm} \ge t_-$。
第二阶段。这证明存在一个时间 t_-，使得 $\tilde{z}_{t_-} \le 2\tilde{z}_0$，并且 $-16\tilde{z}_0^2\epsilon \le y_{t_-} \le 0$。

为了理解动力学，我们将使用坐标变换。将方程 120 解为 $y_{t+2} - y_t = 0$ 对于 $\tilde{z}_t \neq 0$，我们有

$$y^* = -\frac{\epsilon}{2 - 3/2\epsilon + 2\epsilon^2} \tag{123}$$

现在考虑方程定义的坐标 δ_t

$$y_t = -(1 + \delta_t)\frac{\epsilon}{2 - 3/2\epsilon + 2\epsilon^2} \tag{124}$$

δ_t 的动态变化由以下公式给出

$$\delta_{t+2} = (1 - 2(4 - 3\epsilon + 4\epsilon^2)\tilde{z}_t^2)\delta_t \tag{125}$$

由于 $\tilde{z}_t$ 1，δ_t 的大小是严格递减的。我们可以通过以下方法对 δ_t 进行约束

$$|\delta_t| \le \exp\left(-8\sum_{s=t_-}^{t} \tilde{z}_s^2\right)|\delta_{t_-}| \tag{126}$$

由于 δ 开始时为负值，且大小递减，我们知道 $y_t > -\frac{\epsilon}{2-3/2\epsilon+2\epsilon^2}$.这

这意味着我们可以通过以下方法约束 $\tilde{z}_t$

$$\tilde{z}_t \geq 2e^{-\epsilon t}\tilde{z}_0 \tag{127}$$

代入后，δ_t 的约束条件如下：

$$|\delta_t| \leq \exp\left(-8\sum_{s=t-}^{t} 4e^{-2\epsilon s}\tilde{z}_0^{2}\right)|\delta_{t-}| \tag{128}$$

使用积分近似法求和，边界变为

$$|\delta_t| \leq \exp\left(-32\tilde{z}_0^2 \int_0^t e^{-2\epsilon s}\, ds\right) |\delta_{t_-}| = \exp\left(-16\tilde{z}_0^2/\epsilon\,(1-e^{-2\epsilon t})\right) |\delta_{t_-}| \tag{129}$$

根据前面的分析，我们知道 $-1 \leq \delta_{t_-} \leq 0$。在大 t 的极限，我们有

$$\lim_{t\to\infty} |\delta_t| \leq \exp\left(-16\tilde{z}_0^2/\epsilon\right) |\delta_{t_-}| \tag{130}$$

如果我们的条件是

$$16\tilde{z}_0^2/\epsilon \geq -\log(\epsilon) \tag{131}$$

则 $\lim_{t\to\infty} |\delta_t| \leq \epsilon^2$

。

如果我们希望 $\lim_{t\to\infty} y_t = -\epsilon/2 + O(\epsilon^2)$，那么我们需要的条件是

$$16\tilde{z}_0^2 \geq -\epsilon\log(\epsilon) \tag{132}$$

或等价地 $-\epsilon\log(\epsilon) < 16\tilde{z}_0^2$。在这些条件下，$\lim_{t\to\infty} \tilde{z}_t = 0$，$\lim_{t\to\infty} y_t = 0$。$-\epsilon/2 + O(\epsilon^2)$. □

通过运行各种初始化的动力学方程，计算特征值的中值（限制在 [1.9, 2.0] 范围内），并绘制与 ϵ 的关系图（图 10），可以在数值上证实这一结果。

$$\dot{\tilde{z}} = 2y\tilde{z} \tag{133}$$

$$\dot{y} = -2(4 - 3\epsilon + 4\epsilon^2)y\tilde{z}^2 - 4\epsilon\tilde{z}^2 \tag{134}$$

也得到了相同的极限（图 10）。从 ODE 可以看出，浓度取决于 y^0 和 y^1 项在 $\tilde{z}$ 中的等阶，以及时间尺度的分离--$\tilde{z}$ 以 ϵ 的速率收敛到 0，而 y_t 以 $\tilde{z}^2$ 的速率收敛到定点。在这两种情况下，偏离 $-\epsilon/2$
缩放为 $O(\epsilon^2)$（图 10，右）。

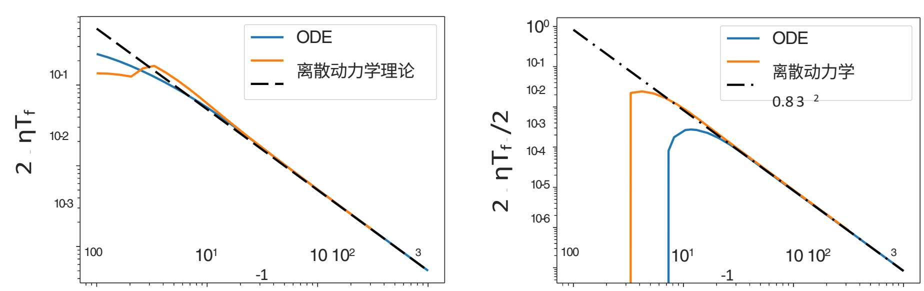


图 10：离散动力学和 ODE 近似法的 y 最终值、与临界值 $T(0) = 2$ 的归一化偏差。在很大范围内，偏差被 $\epsilon/2$ 很好地近似（左图）。与 $\epsilon/2$ 的偏差为 $O(\epsilon^2)$ (右图)。

C 二次回归模型动态

我们在本节中使用爱因斯坦求和符号--等式右侧的重复指数被视为求和，除非它们出现在

左侧。

C.1 定理 3.1 的证明

让 **z**、**J** 和 **Q** 以均值为 0、方差为 σ^2, σ^2, 的 i.i.d. 随机元素初始化。
z J
分别为 1。此外，让分布在数据空间和参数空间中都不随旋转变化，并具有有限的第 4 矩
。

In order to understand the development of the curvature at early times, we consider coordinates which convert **J** into its singular value form. In these coordinates, we can write:

$$J_{\alpha i} = \begin{cases} 0 & \text{如果}\ \alpha \neq i \\ \sigma_\alpha & \text{if}\ \alpha = i \end{cases} \tag{135}$$

奇异值 σ_α 是 NTK 矩阵奇异值的平方根。我们假设它们的大小从大（σ_1）到小排列。根据假设，在这种旋转下，**z** 和 **Q 的**统计量保持不变。

$t = 0$ 时的时间导数可直接在奇异值坐标中计算。一阶导数的计算公式为

$$\frac{d}{dt}\sigma_\alpha^2 = 2\sigma_\alpha \dot{\sigma}_\alpha \tag{136}$$

使用对角线坐标系，我们可以得出

$$\mathrm{E}\left[\frac{d}{dt}\sigma_\alpha^2\right] = \mathrm{E}[\mathbf{Q}_{\alpha\beta j}\mathbf{J}_{\beta j} z_\beta] = 0 \tag{137}$$

然而，平均二阶导数为正。计算得出

$$\frac{d^2}{dt^2}\sigma_\alpha^2 = 2(\dot{\sigma}_\alpha^2 + \sigma_\alpha \ddot{\sigma}_\alpha) \tag{138}$$

我们可以计算初始化时的平均值。我们有

$$\mathrm{E}[\dot{\sigma}_\alpha^2] = \mathrm{E}[\mathbf{Q}_{\alpha\beta j}\mathbf{J}_{\beta j} z_\beta \mathbf{Q}_{\alpha\delta k}\mathbf{J}_{\delta k} z_\delta] = \mathrm{E}[\delta_{\beta\delta}\delta_{jk} J_{\beta j} J_{\delta k} z_\beta z_\delta]。 \tag{139}$$

$$\mathrm{E}[\dot{\sigma}_\alpha^2] = \mathrm{E}[\mathbf{Q}_{\alpha\beta j}\mathbf{J}_{\beta j} z_\beta \mathbf{Q}_{\alpha\delta k}\mathbf{J}_{\delta k} z_\delta] = \mathrm{E}\Big[\sum_{\beta j}\mathbf{J}_{\beta j}^2 \mathbf{z}_\beta^2\Big] = DP\sigma_J^2\sigma_z^2 \tag{140}$$

为了计算第二项，我们计算 $\ddot{J}_{\alpha i}$：

$$\ddot{J}_{\alpha i} = -\mathbf{Q}_{\alpha ij}(\mathbf{J}_{\beta j}\dot{z}_\beta + \dot{\mathbf{J}}_{\beta j}\mathbf{z}_\beta) \tag{141}$$

扩大范围，我们有

$$\ddot{J}_{\alpha i} = \mathbf{Q}_{\alpha ij}(\mathbf{J}_{\beta j}\mathbf{J}_{\beta k}\mathbf{J}_{\delta k} z_\delta + \mathbf{Q}_{\beta jk}\mathbf{J}_{\delta k} \mathbf{z}_\delta \mathbf{z}_\beta) \tag{142}$$

在对角坐标 $\mathbf{J}_{\alpha\alpha} = \sigma_\alpha$ 中。由此可知

$$\mathrm{E}[\sigma_\alpha \ddot{\sigma}_\alpha] = \mathrm{E}[\sigma_\alpha Q_{\alpha\alpha j} Q_{\beta jk} J_{\delta k} z_\delta z_\beta]。 \tag{143}$$

将 **Q** 平均，我们得到

$$\mathrm{E}[\sigma_\alpha \ddot{\sigma}_\alpha] = P\,\mathrm{E}[\sigma_\alpha \delta_{\alpha\beta}\delta_{\alpha k} J_{\delta k} z_\delta z_\beta] = \mathrm{E}[\sigma_\alpha z_\alpha z_\delta J_{\delta\alpha}]。 \tag{144}$$

其评估结果为

$$\mathrm{E}[\sigma_\alpha \ddot{\sigma}_\alpha] = \sigma_z^2 P\,\mathrm{E}[\sigma_\alpha^2]。 \tag{145}$$

在大 D 和大 P 的极限条件下，对于固定的 D/P 比值，根据马琴科-帕斯图分布的统计数据，我们可以计算出最大特征模式的导数为

$$\mathrm{E}[\sigma_0 \ddot{\sigma}_0] = \sigma_z^2\sigma_J^2 P^2 D(1 + \sqrt{D/P})^2 \tag{146}$$

综上所述，我们可以得出

$$\mathrm{E}\left[\frac{d^2\lambda_{max}}{dt^2}\right] = \sigma_z^2\sigma_J^2 DP\left(P\left(1 + \sqrt{D/P}\right)^2 + 1\right) \tag{147}$$

我们在图 11 中用数字证实了这一预测。

That is, the second derivative of the maximum curvature is positive on average. If we normalize with respect to the eigenvalue scale, in the limit of large D and P we have:

$$\mathrm{E}\left[\frac{d^2\lambda_{max}}{dt^2}\right] / E[\lambda_{max}] = \sigma_z^2 \tag{148}$$

z

因此，增加 σ_z 会增加 λ_{max} 轨迹的相对曲率。这就是定理 3.1 的证明。

这一结果表明，随着 σ_z 的增加，逐渐锐化的程度也在增加。通过观察 GF 轨迹（图 12）可以证实这一点。σ_z 较小的轨迹曲率变化不大，损失以一定速度呈指数衰减。然而，当 σ_z 较大时，曲率最初会增加，然后稳定在一个较高的值上，从而可以更快地收敛到损失的最小值。

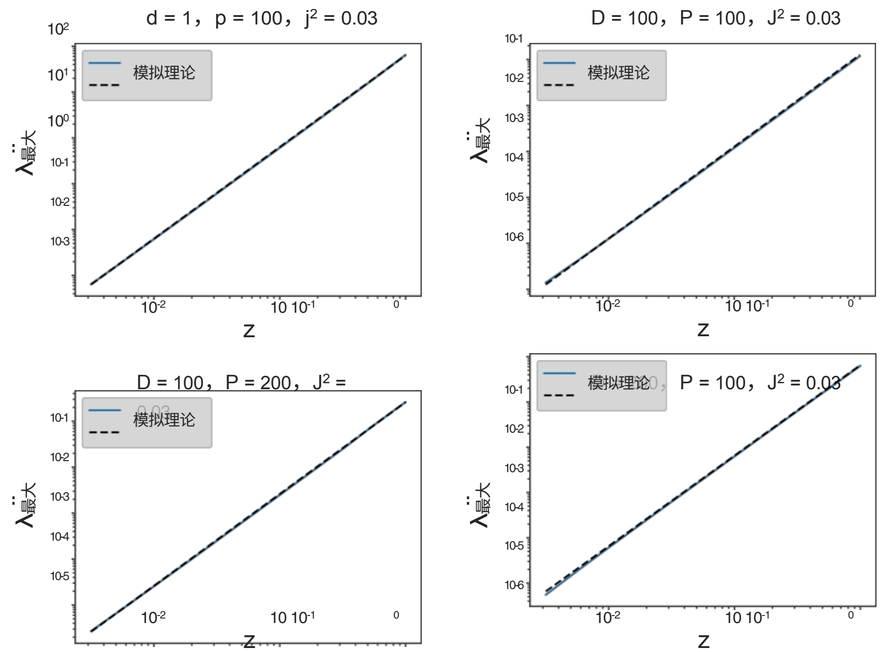


图 11：平均 $\lambda_{\max}(0)$ 与 σ_z 的关系，不同的 D 和 P（100 粒种子）。

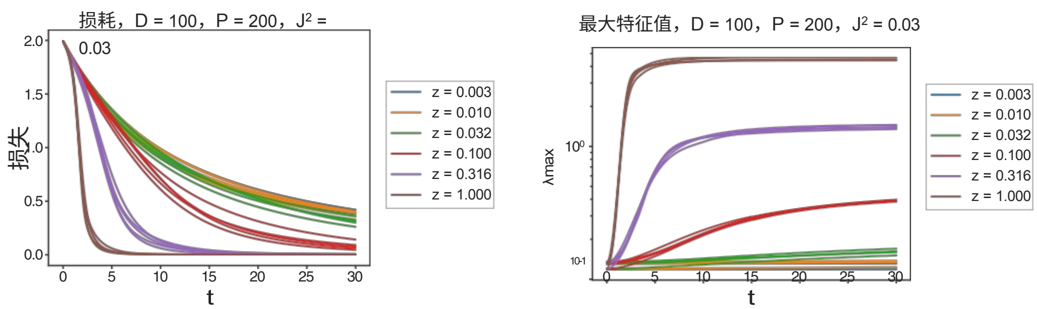


图 12：变化 σ_z 时二次回归模型的损失和最大 NTK 特征值的梯度流轨迹。随着 σ_z 的增大，$\lambda_{\max}$ 变化更快，且总体上呈增大趋势。在 GF 动力学中，σ_z 越高的模型收敛得越快。

C.2 梯度下降的时间尺度

考虑对 **z**、**J** 和 **Q** 进行随机初始化，其中各项均为 i.i.d.，均方差为零 σ_z^2, σ_J^2 和 1，以及有限第四矩。此外，假设 **z**、**J** 和 **Q** 为在输入和输出空间中都是旋转不变的。在这些条件下，我们希望计算

$$r_{NL}^2 \equiv \frac{\mathrm{E}[\|\frac{1}{2}\eta Q^2_{\alpha ij}(\mathbf{J})_{\beta i0}(\mathbf{z})_{\beta 0}(\mathbf{J})_{\delta j0}(\mathbf{z})_{\delta 0}\|^2]}{\mathrm{E}[/\eta(\mathbf{J})_{\alpha i0}(\mathbf{J})_{i\beta 0}(\mathbf{z})_{\beta 0}\|^2_2]}{}^2 = \frac{1}{4}\eta^2\sigma_z^2 D^{22} \tag{149}$$

分母的计算公式为

$$\mathrm{E}[\mathbf{J}\,\mathbf{J}_{\alpha i\beta i}(\mathbf{z}_\beta)\mathbf{J}\,\mathbf{J}_{\alpha j\delta j}(\mathbf{z}_\delta)] = \sigma_z^2\,\mathrm{E}[\mathbf{J}\,\mathbf{J}\,\mathbf{J}_{\alpha i\beta i\alpha j\delta j}\ J\delta_{\beta\delta}] = \sigma_z^2\,\mathbf{E}[\mathbf{J}\,\mathbf{J}\,\mathbf{J}\,\mathbf{J}]。_{\alpha i\beta i\alpha j\beta j} \tag{150}$$

评估为我们提供了

$$\mathrm{E}[\mathbf{J}\,\mathbf{J}_{\alpha i\beta i}\,(\mathbf{z}_\beta\,)\mathbf{J}\,\mathbf{J}_{\alpha j\delta j}\,(\mathbf{z}_\delta\,)] = \sigma_z^2\,(\sigma_J^4\,(P\,(P-1)D) + C_4\,DP\,) \tag{151}$$

其中 C_4 是 $\mathbf{J}_{\alpha i}$ 的第 4 矩。在 D 和 P 中达到最低阶

$$\mathrm{E}[\mathbf{J}\ \mathbf{J}_{\alpha i\beta i}\ (\mathbf{z}_\beta\ )\mathbf{J}\ \mathbf{J}_{\alpha j\delta j}\ (\mathbf{z}_\delta\ )] = \underset{z\ J}{\sigma\ \sigma^{24}}\ DP^2 + O(DP\ ) \tag{152}$$

对分子进行求值，我们得出

$$\mathrm{E}[Q\alpha ij\ J\beta iz\beta\ J\delta j\ z\delta\ Q\alpha mnJ\gamma mz\gamma\ J\nu nz\nu\ ] = \mathrm{E}[J\beta iz\beta\ J\delta j\ z\delta\ J\gamma mz\gamma\ \mathbf{J\nu nz\nu}\ ](\delta im\delta jn + (\mathrm{M4}\ \ -\ \ 1)\delta ijmn) \tag{153}$$

其中，M_4 是 $\mathbf{Q}_{\alpha ij}$ 的第 4 矩。由此可得

$$\frac{1}{D}\mathrm{E}[Q\alpha ij\ J\beta iz\beta\ J\delta j\ z\delta\ Q\alpha mnJ\gamma mz\gamma\ J\nu nz\nu\ ] = \mathrm{E}[J\beta iz\beta\ J\delta j\ z\delta\ J\gamma iz\gamma\ J\nu j\ z\nu\ ]+ \\ (M_4\ -\ 1)\mathrm{E}[\mathbf{J\ z\ J\ z\ J\ z\ J\ z}\ ]_{\beta i\beta\delta i\delta\gamma i\gamma\nu i\nu}。 \tag{154}$$

接下来，我们进行 $\mathbf{z}$ 平均。我们有

$$\begin{aligned}\frac{1}{D}\mathrm{E}[\mathbf{Q}_{\alpha ij}\ \mathbf{J}_{\beta i}\mathbf{z}_\beta\mathbf{J}_{\delta j}\mathbf{z}_\delta\mathbf{Q}_{\alpha mn}\ \mathbf{J}_{\gamma m}\mathbf{z}_\gamma\mathbf{J}_{\nu n}\mathbf{z}_\nu\ ] &= \sigma_z^4\,\mathrm{E}[\mathbf{J}_{\beta i}\mathbf{J}_{\delta j}\mathbf{J}_{\gamma i}\mathbf{J}_{\nu j}\ ]\ (\delta_{\beta\delta}\delta_{\gamma\nu} + \delta_{\beta\gamma}\delta_{\delta\nu} + \delta_{\beta\nu}\delta_{\delta\gamma}) \\ &+ (\mathrm{C4} - \sigma^4)\mathrm{E}[J\beta iJ\delta j\ J\gamma iJ\nu j\ ]\delta\beta\delta\gamma\nu \\ &+ (M_4 - 1)\sigma_z^4\,\mathrm{E}[\mathbf{J\ J\ J}_{\beta i\delta i\gamma i\nu i}\ ]\ \mathrm{J}(\delta\ \delta_{\beta\delta\gamma\nu} + \delta\ \delta_{\beta\gamma\delta\nu} + \delta_{\beta\nu}\ \delta\ )_{\delta\gamma} \\ &+ (M_4 - 1)(C_4 - \sigma^4\ )\mathrm{E}[\mathbf{J\ J\ J}_{\beta i\delta i\gamma i\nu i}\ ]\ J\delta_{\beta\delta\gamma\nu}\end{aligned} \tag{155}$$

其中 C_4 是 $\mathbf{z}$ 的第 4 矩：

$$\begin{aligned}\frac{1}{D}\mathrm{E}[\mathbf{Q}_{\alpha ij}\ \mathbf{J}_{\beta i}\mathbf{z}_\beta\mathbf{J}_{\delta j}\mathbf{z}_\delta\mathbf{Q}_{\alpha mn}\ \mathbf{J}_{\gamma m}\mathbf{z}_\gamma\mathbf{J}_{\nu n}\mathbf{z}_\nu\ ] &= \sigma_z^4\,(\mathrm{E}[\mathbf{J}_{\beta i}\mathbf{J}_{\beta j}\mathbf{J}_{\delta i}\mathbf{J}_{\delta j}]+ \mathrm{E}[\mathbf{J}_{\beta i}\mathbf{J}_{\delta j}\mathbf{J}_{\beta i}\mathbf{J}_{\delta j}\ ]+ \mathrm{E}[\mathbf{J}_{\beta i}\mathbf{J}_{\delta j}\mathbf{J}_{\delta i}\mathbf{J}_{\beta j}]) \\ &+ (C_4 - \sigma^4{}_z)\mathrm{E}[\mathbf{J\ J\ J\ J}\ ]_{\beta i\beta j\beta i\beta j}。 \\ &+ (M_4 - 1)\sigma_z^4\ (\mathrm{E}[\mathbf{J\ J\ J}_{\beta i\beta i\gamma i\gamma i}\ ]\ \ \mathrm{J}+ \mathrm{E}[\mathbf{J\ J}_{\beta i\delta i}\ \mathbf{J\ J}_{\beta i\delta i}\ ] + \mathrm{E}[\mathbf{J\ J\ J}_{\beta i\delta i\delta i\beta i} \\ &\mathrm{J}]) \\ &+ (M_4 - 1)(C_4 - \sigma^4\ )\mathrm{E}[\mathbf{J\ J\ J\ J}\ ]_{\beta i\beta i\beta i\beta i}。\end{aligned} \tag{156}$$

在 D 和 P 较大的情况下，最后三项都会逐渐小于第一项。对第一项进行前导求值，我们得到

$$\frac{1}{D}\mathrm{E}[\mathbf{Q}_{\alpha ij}\ \mathbf{J}_{\beta i}\mathbf{z}_\beta\mathbf{J}_{\delta j}\mathbf{z}_\delta\mathbf{Q}_{\alpha mn}\ \mathbf{J}_{\gamma m}\mathbf{z}_\gamma\mathbf{J}_{\nu n}\mathbf{z}_\nu\ ] = \sigma_z^4\ \sigma_J^4\ (2DP^2 + 2D^2\ \mathrm{P} + D\ P^{22}\ ) + O(D^2\ \mathrm{P} + DP\ )^2 \tag{157}$$

$$\mathrm{E}[\mathbf{Q\ J\ z\ J\ z\ Q\ J}_{\alpha ij\beta i\beta\delta j\delta\alpha mn\gamma m\gamma\nu n\nu}\ ]\ \ \mathrm{z\ J\ z}= \sigma_z^4\ \sigma_J^4\ D\ P^{32} + O(D^3\ \mathrm{P} + D\ P\ )^{22} \tag{158}$$

这样我们就

$$r^2_{NL} = \frac{1\ \sigma_z^4\ \sigma_J^4\ D\ P\ 1^{32}}{4\ \sigma_z^2\sigma_J^4\ DP^{\ 2}} = \frac{}{4}\sigma_z^{\ }D^{22} \tag{159}$$

在大 D 和大 P 的极限情况下，达到前导阶。

D　真实模型分析

D.1　CIFAR10 模型中 y 的动态变化

第 4 节分析的 CIFAR10 模型中 y 的动态变化比 z 的动态变化更复杂，[1]。从图 5 中我们可以看到，在 y 的两步变化中，有一个与 z_1 和 y 无关的部分。我们可以通过计算小 z_1 时 $y_{t+2} - y_t$ 的平均值（在这种情况下，取 $z_1 < 10^{-4}$）来近似估计这一变化 b。然后，我们可以从 $y_{t+2} - y_t$ 中减去 b，并绘制出余数与 z^2 的对比图（图 13 左）。我们可以看到，

y_{t+2} - y_t - b 与 z^2 负相关、 t t
尤其是对于较大的 z_t 。然而，y_{t+2} - y_t 显然不是 z_1 的简单函数。

两步模型动态可写成 $(ay + c)z^{-2}$ 。如果我们绘制 $(y_{t+2}\ y_t\ b)/z^2$ 与 y_t 的关系图，我们又没有一个单值函数（图 13，右）。因此，$_1$ y_{t+2} - y_t 的函数形式不是由 $b + ayz^2 + cz^2$ 给出的。

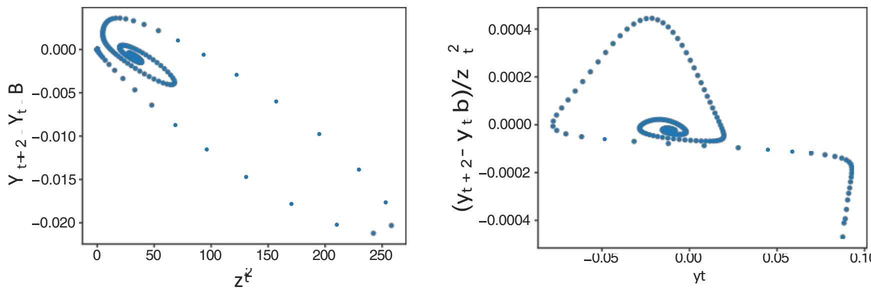


图 13：变化 σ_z 时二次回归模型的损失和最大 NTK 特征值的梯度流轨迹。随着 σ_z 的增大，λ_{max} 变化更快，且总体呈增大趋势。在 GF 动力学中，σ_z 越高的模型收敛得越快。

D.2 2 类 CIFAR 模型的二次展开

我们使用神经切线库（Novak 等人，2019 年）仅使用前两个类别的 5000 个数据点训练了一个 CIFAR 模型，通过该库，我们可以在任意参数下对模型进行二阶泰勒展开。模型为 2 隐藏层全连接网络，隐藏宽度为 256，Erf 为非线性。模型以 NTK 参数化初始化，权重方差为 1，偏差方差为 0。目标为标量值--第一类为 +1，第二类为 1。所有实验均使用 0.003204 的学习率。所有绘图均为使用浮点 64 精度。

在初始化时进行二次展开，我们可以看到，在这种情况下，损失在前 1000 步跟踪了完整模型（图 14 左），但错过了稳定边缘行为。我们使用神经切线来高效计算 NTK，从而得到顶部特征值 λ_1 （进而得到 y）。我们还可以通过计算相关特征向量 $\mathbf{v}_1$ 和投影残差 $\mathbf{z}$ 来计算 z_1。如果二次展开更接近稳定边缘，则 z_1 的动态与真实的 z_1 动态非常接近，直到与不同时间发生的 z_1 指数增长相关的偏移（图 14，中间）。我们看到，z_1 第一个峰值的形状在完整模型和二次模型中是一样的，但在完整模型中，随后的振荡更快，阻尼也更快。这表明，二次模型可能捕捉到了初始 EOS 行为，但详细的动力学需要了解高阶项。例如，三阶泰勒扩展改进了对振荡幅度和周期的预测，但仍然忽略了关键的定量特征（图 14，右）。

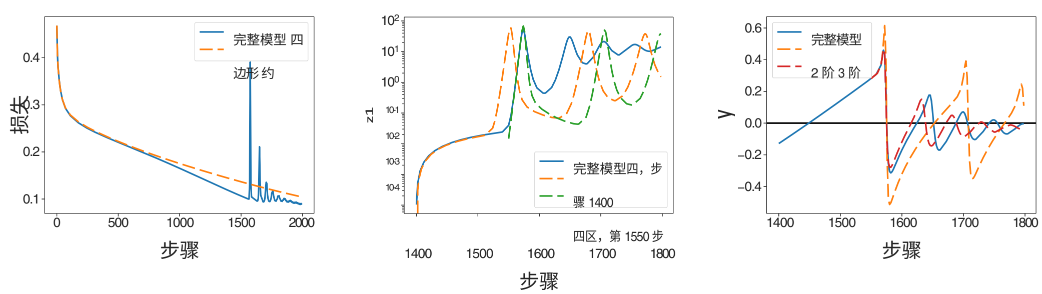


图 14：基于两类 CIFAR 训练的 FCN 模型的二次展开。在初始化时进行扩展，可以在 1000 步内很好地近似完整模型，之后完整模型会出现 EOS 行为， 而近似模型则不会（左图）。当 z_1 较小的时候，二次模型会跟踪完全模型；然而，在近似模型中，初始指数增长可能会更早（中）。与近似模型相比，完整模型中 z_1 的振荡幅度更大。三阶泰勒扩展更好地捕捉

到了振荡的幅度和周期，但仍然没有捕捉到定量特征（右图）。