arXiv:2602.01480v1 [cs.LG] 1 Feb 2026

Rod Flow：梯度下降在稳定性边缘的连续时间模型

Eric Regis* Sinho Chewi*

摘要

我们如何理解非凸景观上的基于梯度的训练？Cohen 等人 [2021], 提出的稳定性边缘现象表明，答案并非如此简单：具体而言，采用较大步长的梯度下降（GD）往往会偏离梯度流。在这一机制下，Cohen 等人 [2025],最近提出的“中心流”为多种架构上的梯度下降动力学提供了精确的常微分方程近似。在本文中，我们提出$Rod\ Flow$，一种替代的常微分方程近似，它具有以下优势：（1）它基于一种有原则的推导，源于将梯度下降迭代视为延展的一维物体——一根“杆”——的物理图景；（2）它能更好地刻画简单玩具示例中的梯度下降动力学，并在代表性神经网络架构上达到与中心流相当的精度；（3）形式显式且计算成本低。理论上，我们证明 Rod Flow 能够正确预测临界尖锐度阈值，并解释四次势函数中的自稳定。我们通过一系列数值实验验证了该理论。

1 引言

机器学习模型通过在非凸损失景观上进行优化来训练。凸优化中的传统分析假设损失景观表现良好——具体而言，海森矩阵的最大特征值（即尖锐度）被 $2/\eta$ 所限制，其中 η是学习率。这一假设是通过下降引理获得经典收敛保证的基础。

然而，这一假设在实践中经常被违反。Cohen 等人 [2021] 观察到，使用全批量梯度下降训练神经网络时，尖锐度不会始终保持在 $2/\eta$以下。相反，它会在训练过程中增加（渐进式尖锐化），直至达到 $2/\eta$，随后在这一阈值附近徘徊（稳定性边缘）。这一现象极为稳健，在各种架构、数据集和学习率下都会出现。

稳定性边缘给连续时间近似带来了根本性挑战。理解梯度下降的经典方法是研究梯度流，即常微分方程 $\dot{w} = -\nabla L(w)$。但要用梯度流近似梯度下降，需要轨迹缓慢变化；而稳定性边缘处的梯度下降会沿海森矩阵最尖锐的方向呈现快速振荡。

已有多项研究尝试对稳定性边缘动力学进行建模 [Arora 等人，2022、Damian等人，2023、Cohen 等人， 2025], 但现有方法要么缺乏形式化论证，要么无法准确追踪离散迭代点。中心流 [Cohen 等人， 2025] 是一种先前提出的连续时间模型，旨在捕捉振荡优化器的时间平均轨迹。它追踪

*耶鲁大学。电子邮件： {eric.regis, sinho.chewi}@yale.edu

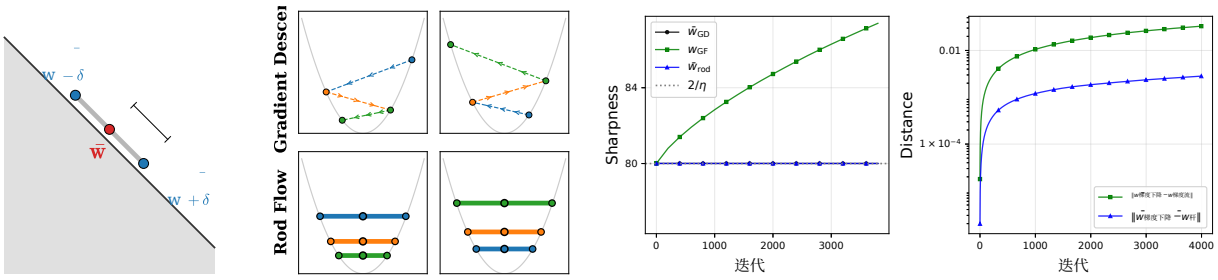


图 1：**Rod Flow 示意图。左：**一根杆沿平坦的损失景观向下移动，其中两个端点表示连续的 GD 迭代点，中点表示它们的平均值。**左中：**在收敛（左列）和发散（右列）步长下，二次损失景观中梯度下降（上排）与 Rod Flow（下排）的比较。**右：**三层多层感知机上梯度下降、梯度流和 Rod Flow 的比较。左图显示随时间变化的锐度：梯度下降和 Rod Flow 保持在稳定性边缘，而梯度流则持续变得更加尖锐。右图显示 GD 迭代点与两种流之间的距离，表明 Rod Flow 比梯度流更接近 GD 轨迹。

振荡的“中心” $\bar{w}$ 以及受半定互补问题约束的协方差矩阵 Σ 。然而，其推导依赖于启发式近似。

我们引入 **Rod Flow**，这是一组将梯度下降建模为扩展的一维物体——“杆”的常微分方程，其特征由中心 $\bar{w}$ 和展度 Σ决定。这一物理图像十分直观：我们不再追踪一个快速振荡的点，而是追踪一个扩展对象，其端点在 $\bar{w}+\delta$ 和 $\bar{w}-\delta$处对损失景观进行采样。关键洞见在于，尽管单个迭代点 w_t 会快速振荡，但连续迭代点的平均值 $\bar{w}_t$ 以及连续迭代点半差的外积 $\delta_t \otimes \delta_t$ 都会逐渐演化，因此适合进行常微分方程近似。

1.1 贡献

我们的主要贡献如下：

- 通过精确差分方程和后向误差分析，从离散动力学原则性地推导出 Rod Flow（第 4 节）
- 理论分析表明，Rod Flow 如何正确刻画 $2/\eta$ 稳定性阈值，并在四次势函数中表现出自稳定（第 5 节）
- 对 Rod Flow 与中心流的详细比较，表明中心流如何作为 Rod Flow 的泰勒近似而出现（第 6 节）
- 实证表明，Rod Flow 在玩具问题上具有高精度，并且在代表性深度学习架构上与中心流具有相当的精度（第 7 节）

2 相关工作

稳定性边缘。 Cohen 等人 [2021] 记录了不同架构中的渐进式尖锐化和稳定性边缘现象，观察到尖锐度上升至 $2/\eta$ 后便徘徊不变。理论解释包括自稳定机制 [Damian 等人、 2023], 两步分析 [Ahn 等人，2023年，Chen 和 Bruna、 2023], 不稳定区域中的收敛保证 [Ahn 等人，2022，Arora

等人， 2022], 以及分岔理论 [Song 和 Yun， 2023]。Ahn 等人 [2023], Zhu 等人 [2023]提供了极简示例，而 Agarwala 等人 [2023] 展示了二次回归中的渐进式尖锐化。弹射机制 [Lewkowycz 等人， 2020] 解释了较大学习率如何促使模型逃离尖锐初始化。其他相关研究包括损失景观分析[Ma 等人， 2022], 较大学习率下的平衡效应 [Wang 等人， 2023], 梯度下降中的混沌动力学 [Chen 等人， 2024], 以及稳定性边缘处的 NTK演化 [Jiang 等人， 2025]。

连续时间模型。 梯度流是梯度下降的经典连续极限，但无法捕捉振荡，因此在稳定性边缘处会失效。Rosca 等人 [2023] 研究了梯度流模型中的不稳定性。与我们工作最接近的是 Cohen 等人 [2025], 提出的中心流，它在临界特征向量构成的子空间中跟踪一个受半定互补问题（SDCP）约束的协方差矩阵 Σ 。我们的 Rod Flow 则提供了一种更简单、原理更清晰的替代方案，该方案由后向误差分析推导而来。

后向误差分析。 由后向误差分析得到的修正方程是数值分析中的标准方法 [Wilkinson, 1963，1965，Hairer 等人， 2006]。Barrett 和 Dherin [2021] 以及 Smith 等人 [2021] 运用这一框架理解梯度下降中的隐式正则化。Shi 等人[2021] 为动量方法推导了高分辨率常微分方程，而 Li 等人 [2017]则为随机梯度下降推导了随机修正方程。Di Giovacchino 等人 [2024] 为将后向误差分析应用于优化算法提供了方法论验证。我们拓展了这一框架，以推导能够捕捉稳定性边缘动力学的常微分方程。

3 梯度下降动力学

我们考虑使用学习率为 η 的梯度下降法来最小化损失函数 $L\colon\ \mathbb{R}^p \to \mathbb{R}$

$$w_{t+1} = w_t - \eta \nabla L(w_t)\,. \tag{1}$$

经典的收敛分析依赖于下降引理；该引理保证了当损失足够光滑时，优化过程能够取得进展。

引理3.1（下降引理）。 如果 L 的梯度是 S-利普希茨连续的（$\|\nabla^2 L\| \le S$ 处处成立），那么对于 $\eta < 2/S$:

$$L(w_{t+1}) \le L(w_t) - \eta\left(1 - \frac{\eta S}{2}\right)\|\nabla L(w_t)\|^2\,. \tag{2}$$

当尖锐度（海森矩阵的最大特征值）超过 $2/\eta$时，这一保证便不再成立，梯度下降可能会发散——事实上，在二次目标函数上确实会发生这种情况。然而，神经网络训练的实际表现则更为复杂。我们将训练过程分为三个不同阶段（图 2）：

第一阶段：稳定下降。 训练初期，尖锐度保持在 $2/\eta$ 以下。损失单调下降，梯度流能够准确地近似梯度下降的连续时间演化。这一阶段的特点是渐进式尖锐化：尖锐度稳步上升，趋近临界阈值 [Cohen 等人，2021年，Agarwala 等人， 2023]。

第二阶段：过渡至稳定性边缘。 当尖锐度首次达到 $2/\eta$ 时，训练进入一个动荡的过渡阶段，其特征是梯度范数出现大幅尖峰和波动，同时损失与尖锐度也随之波动。这些事件具有 Lewkowycz 等人 [2020] 在研究“弹射现象”时所发现的自稳定机制。不同于训练初始化时发生的单次大规模弹射，此处渐进式尖锐化驱动尖锐度反复超过阈值，在动力学趋于稳定之前产生多次瞬态不稳定性。

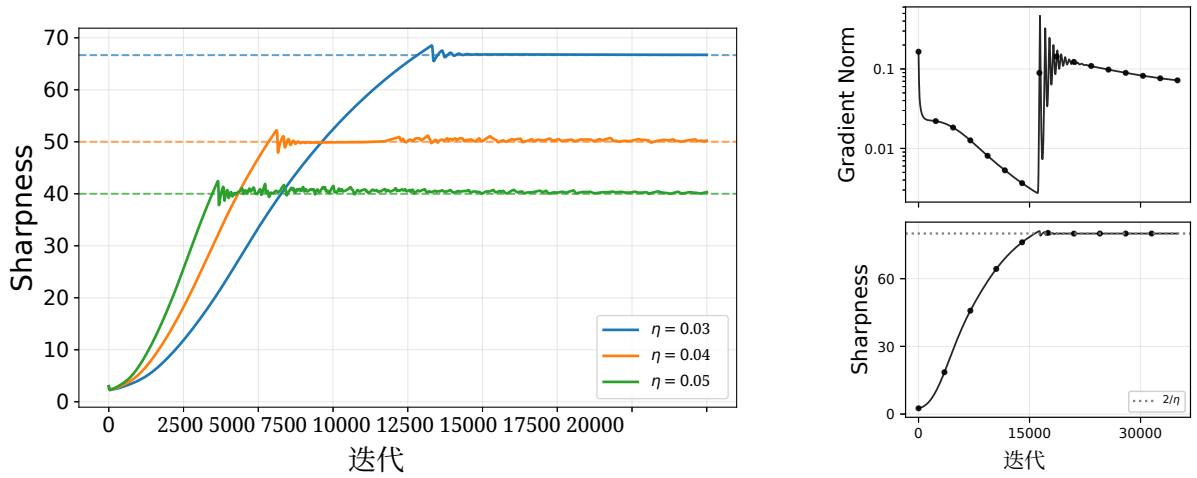


图 2：**稳定性边缘。左：**三种步长下呈现的稳定性边缘。对于所有步长，我们都观察到渐进式尖锐化，随后尖锐度稳定在 $2/\eta$。**右：**稳定性边缘的三个阶段。阶段 I（稳定下降）：尖锐度渐进式增加，同时梯度范数下降。阶段 II（进入稳定性边缘）：梯度范数骤增并剧烈波动。第三阶段（稳态稳定性边缘）：尖锐度保持在 $2/\eta$，而梯度范数变得平滑。

阶段 3：稳态稳定性边缘。 过渡阶段之后，训练进入一个尖锐度在 $2/\eta$ 附近徘徊的区域，但振荡幅度平滑变化，没有显著尖峰 [Cohen 等人，2021年，Damian 等人， 2023]。尽管局部存在非单调性，损失总体上仍持续下降，动力学呈现出规则且有界的振荡。第二阶段的瞬态不稳定性涉及梯度快速且大幅度的变化，难以通过任何连续时间近似捕捉。然而，阶段 3 的稳态振荡平滑且规则，因此适合采用连续时间模型进行描述。这正是 Rod Flow 旨在实现的目标：构建一种有理论依据的常微分方程近似，准确追踪稳态稳定性边缘区域中的梯度下降动力学。

4 Rod Flow

4.1 设置与符号

我们定义连续迭代点的中心 和半差：

$$\bar{w}_t = \frac{w_{t+1} + w_t}{2}, \quad \delta_t = \frac{w_{t+1} - w_t}{2}. \tag{3}$$

根据这些定义，我们可以还原原始迭代序列 $w_t = \bar{w}_t - \delta_t$ 和 $w_{t+1} = \bar{w}_t + \delta_t$。利用梯度下降更新规则 $w_{t+1} = w_t - \eta \nabla L(w_t)$，可得 $\delta_t = -\frac{\eta}{2} \nabla L(w_t)$。

令 $u \otimes v := uv^\top$ 表示外积。量 $\delta_t \otimes \delta_t$ 刻画了半差的大小和方向。其连续时间对应量记为 Σ。我们使用 Σ而非外积记法，因为 Σ 在演化过程中仅近似为秩-1（而且这种记法也与中心流的协方差矩阵相一致）。根据经验， Σ始终近似为秩-1，其最大特征值与次大特征值之比在 10^2–10^3 之间。因此，我们将 δ 视为

Σ 的主特征向量乘以其主特征值的平方根。为简洁起见，我们写作$L_+ := L(\bar{w} + \delta)$ 和 $L_- := L(\bar{w} - \delta)$，其中 $\nabla L_\pm$ 和 $\nabla^2 L_\pm$ 分别表示相应的梯度和海森矩阵。

4.2 Rod Flow方程

Rod Flow由关于中心 $\bar{w}$ 和展度 Σ的一组常微分方程定义：

$$\begin{aligned}\frac{d\bar{w}}{dt} &= -\frac{\eta}{2}\big[\nabla L_+ + \nabla L_-\big] \\ &\quad - \frac{\eta^2}{8}\big[\nabla^2 L_+ + \nabla^2 L_-\big]\big[\nabla L_+ + \nabla L_-\big]\,, \qquad (4)\\ \frac{d\Sigma}{dt} &= \frac{\eta^2}{4}\big[\nabla L_+ \otimes \nabla L_+ + \nabla L_- \otimes \nabla L_-\big] - 2\Sigma\,. \qquad (5)\end{aligned}$$

中心的演化遵循两个端点$\bar{w} \pm \delta$所受作用力的平均值。展度则根据端点处的梯度幅值增长，同时恢复力会促使其 Σ 趋向于零。

本节余下部分将推导这两个方程。我们先推导关于 $\bar{w}_t$ 和 $\delta_t \otimes \delta_t$的精确差分方程，再通过后向误差分析将其提升为常微分方程。

4.3 精确差分方程

4.3.1 关于 $\bar{w}_t$的差分方程

对于中心 $\bar{w}_t$，我们计算：

$$\begin{aligned}\bar{w}_{t+1} - \bar{w}_t &= \frac{w_{t+2} + w_{t+1}}{2} - \frac{w_{t+1} + w_t}{2} \\ &= \frac{w_{t+2} - w_{t+1}}{2} + \frac{w_{t+1} - w_t}{2} \\ &= -\frac{\eta}{2}\big[\nabla L_+ + \nabla L_-\big]\,. \qquad (6)\end{aligned}$$

最后一行源于将梯度下降更新规则同时应用于 w_{t+1} 和 w_t。随后我们利用了 $w_{t+1} = \bar{w}_t + \delta_t$ 与 $w_t = \bar{w}_t - \delta_t$ 这一事实。

4.3.2 $\delta_t \otimes \delta_t$ 的差分方程

为了推导 $\delta_t \otimes \delta_t$ 的差分方程，我们首先加上并减去 $\delta_t \otimes \delta_t$：

$$\delta_{t+1} \otimes \delta_{t+1} - \delta_t \otimes \delta_t = \delta_{t+1} \otimes \delta_{t+1} + \delta_t \otimes \delta_t - 2\delta_t \otimes \delta_t\,. \qquad (7)$$

利用 $\delta_t = -\frac{\eta}{2}\nabla L(w_t)$，我们接着得到：

$$\begin{aligned}&\delta_{t+1} \otimes \delta_{t+1} - \delta_t \otimes \delta_t \\ &\quad = \frac{\eta^2}{4}\big[\nabla L_+ \otimes \nabla L_+ + \nabla L_- \otimes \nabla L_-\big] - 2\delta_t \otimes \delta_t\,. \qquad (8)\end{aligned}$$

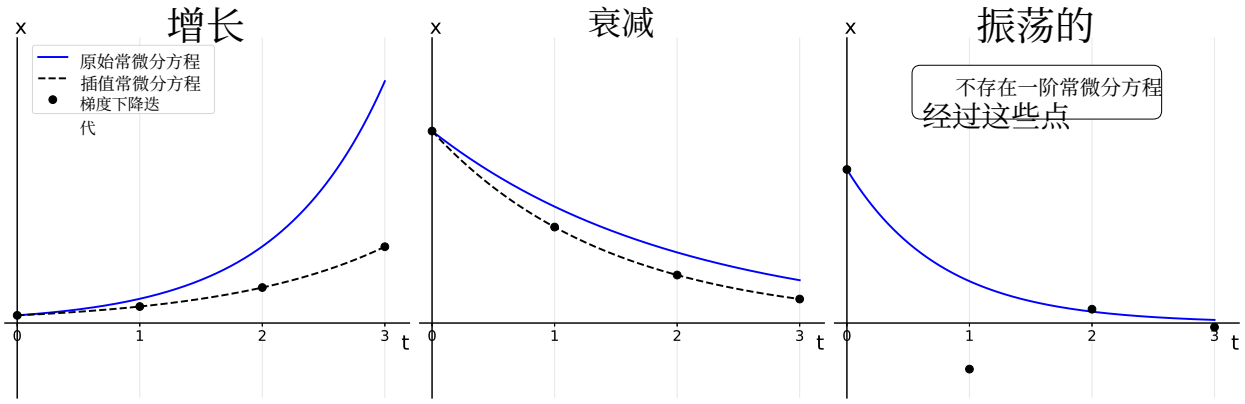


图 3：**后向误差分析**。 经过离散点的常微分方程，不同于由其离散化生成这些点的常微分方程。**左：** 指数增长。**中：** 指数衰减。**右：** 梯度下降迭代的符号交替变化，且不存在一阶常微分方程经过它们。

4.3.3 为什么追踪 $\delta \otimes \delta$ 而不是 δ?

一个自然的问题是：为什么跟踪 $\delta_t \otimes \delta_t$，而不直接跟踪 δ_t？

答案在于稳定性边缘处的行为。当尖锐度超过 $2/\eta$时，梯度下降迭代沿尖锐轴方向振荡，导致 δ_t 的相应分量符号交替。一阶常微分方程无法刻画振荡。因此，任何关于 δ_t 的常微分方程都无法刻画稳定性边缘。

然而，即使 δ_t 发生振荡，其幅度 $\|\delta_t\|$ 仍可平滑变化。考虑 $\delta_{t+1} = -\delta_t$时的极端情形。尽管 δ_t 在每个时间步都会翻转符号， $\delta_t \otimes \delta_t$ 仍保持不变。

4.3.4 差分方程的性质

方程（6）和方程（8）都是精确的——没有进行任何近似。我们只是用不同的变量重写了梯度下降递推。

有两项性质值得强调。首先，两个方程在 $\delta_t \mapsto -\delta_t$ 下都具有对称性，这一点至关重要，因为在我们的形式体系中， δ_t 的符号是任意的。其次，两个方程都只依赖于（$\bar{w}_t$， δ_t），从而构成一个独立于原始迭代点的封闭系统。

4.4 后向误差分析

现在，我们将差分方程提升为常微分方程。关键见解是：通过欧拉离散化定义递推关系 $x_{t+1} - x_t = V(x_t)$ 的常微分方程，并非穿过离散时间迭代点的常微分方程——必须考虑离散化误差。

4.4.1 一个启发性示例

考虑一个具体例子。常微分方程 $\dot{x} = kx$ 的解为 $x(t) = x_0 e^{kt}$。使用步长 $\Delta t = 1$ 进行离散化，得到：

$$x_{t+1} - x_t = kx_t \implies x_{t+1} = (1+k)x_t\,. \tag{9}$$

迭代点满足 $x_t = (1+k)^t x_0 = e^{\ln(1+k)\cdot t} x_0$。因此，经过这些离散点的常微分方程为 $\dot{x} = \ln(1+k)x$，而不是 $\dot{x} = kx$ （图 3）。

4.4.2 一般后向误差分析

给定差分方程 $x_{t+\Delta t} - x_t = \Delta t\ V(x_t)$，我们希望找到一个常微分方程，使其解能够插值这些离散点。假设一条光滑曲线 $x(t)$ 经过这些迭代点，泰勒展开可得：

$$x(t+\Delta t) = x(t) + \Delta t\ \dot{x}(t) + \frac{\Delta t^2}{2}\ddot{x}(t) + O(\Delta t^3). \tag{10}$$

将上述内容与 $x(t+\Delta t) = x(t) + \Delta t\ V(x(t))$ 相匹配，并对自治常微分方程使用 $\ddot{x} = \nabla V(x)\ \dot{x}$，可得：

$$\dot{x} = V(x) - \frac{\Delta t}{2}\nabla V(x)\cdot V(x) + O(\Delta t^2)\,. \tag{11}$$

右侧第一项是朴素的连续极限。我们必须推导到 Δt 阶，才能得到后向误差分析给出的第一个修正项。由此得到修正向量场$\tilde{V}(x) = V(x) - \frac{1}{2}\nabla V\cdot V$。附录 A 对后向误差分析作了更详细的阐述。

4.4.3 对 $\bar{w}$ 的应用

方程（6）的形式为 $\bar{w}_{t+1} - \bar{w}_t = V(\bar{w}_t,\ \delta_t)$，其中：

$$V(\bar{w},\delta) = -\frac{\eta}{2}\big[\nabla L_+ + \nabla L_-\big]\,. \tag{12}$$

应用方程（11），雅可比矩阵为：

$$\nabla_{\bar{w}} V = -\frac{\eta}{2}\big[\nabla^2 L_+ + \nabla^2 L_-\big]\,. \tag{13}$$

因此，后向误差修正常微分方程为 $\frac{d\bar{w}}{dt} = V - \frac{1}{2}\nabla_{\bar{w}} V\cdot V$，由此得到方程（4）。

4.4.4 应用于 Σ

对于 Σ 方程，我们省略后向误差修正。其理由是时间尺度分离：方程（8）中的 -2Σ 衰减项以速率 2 变化，而 $\bar{w}$ 以速率 $O(\eta)$ 变化。当 $\eta \ll 1$ 时，这种分离意味着 Σ 会迅速达到由当前 $\bar{w}$决定的准稳态。由于 Σ 始终接近其瞬时平衡，后向误差修正的影响可以忽略不计。因此，我们直接将方程（8）提升为方程（5）。

5 理论分析

Rod Flow 的简单明了的结构使我们能够分析其在典型损失函数上的行为，并建立直观认识。我们的目标是理解（ $\bar{w},\Sigma$）系统的不动点及其稳定性。

5.1 平坦景观

考虑具有常数梯度 $\nabla L = -b$ 和零曲率的线性损失 $L(w) = -b\cdot w$。Rod Flow方程变为：

$$\frac{d\bar{w}}{dt} = \eta b\,, \tag{14}$$

$$\frac{d\Sigma}{dt} = \frac{\eta^2}{2} b\otimes b - 2\Sigma\,. \tag{15}$$

中心以恒定速度 ηb移动。令 $\frac{d\Sigma}{dt} = 0$ ，可得稳态范围:

$$\Sigma^* = \frac{\eta^2}{4} b \otimes b \,. \tag{16}$$

这与离散动力学完全吻合：相邻的梯度下降迭代点相距 $\eta\|b\|$，因此半差的大小为 $\|\delta_t\| = \frac{\eta}{2}\|b\|$。

5.2 二次损失

考虑二次函数 $L(w) = \frac{1}{2} S w^2$ ，其中 $S > 0$。为简便起见，我们将讨论一维情形。由于 Σ 是标量，因此有 $\delta = \sqrt{\Sigma}$。

显然，($\bar{w}^*$， Σ^*) =(0，0) 是 Rod Flow 的不动点。我们将通过考察以下初始条件来分析其稳定性: $\bar{w}(0) = 0$ 以及 $\Sigma(0) > 0$。Rod Flow方程变为:

$$\frac{d\bar{w}}{dt} = 0 \,, \tag{17}$$

$$\frac{d\Sigma}{dt} = \Big(\frac{\eta^2 S^2}{2} - 2\Big)\Sigma \,. \tag{18}$$

中心保持静止。对于 Σ，定义 $\beta = \frac{\eta^2 S^2}{4} - 1$。于是:

$$\frac{d\Sigma}{dt} = 2\beta\Sigma \implies \Sigma(t) = \Sigma(0) e^{2\beta t} \,. \tag{19}$$

当 $\beta < 0$ 时，展度衰减；当 $\beta > 0$时，展度增长。要使 (0，0) 不动点稳定，必须满足:

$$\beta < 0 \implies S < \frac{2}{\eta} \,. \tag{20}$$

这重新得到了稳定性边界阈值：当尖锐度$S < 2/\eta$时，不动点 (0,0) 稳定；当 $S > 2/\eta$ 时，则不稳定。

5.3 四次势

二次情形的分析预测，当 $S > 2/\eta$时， Σ 中的增长将无界。实际上，振荡会在有限振幅处稳定下来。为刻画这一现象，我们加入高阶项。

考虑一维的偶四次势:

$$L(w) = \frac{S}{2} w^2 - \frac{Q}{4} w^4 \,, \tag{21}$$

其中， $S > 0$ 是原点处的曲率，而 $Q > 0$ 控制四次项。我们有$\nabla L(w) = Sw - Qw^3$。

在 $\bar{w} = 0$ 且满足 $\delta = \sqrt{\Sigma}$时，我们有:

$$\nabla L(\pm\delta) = \pm(S\sqrt{\Sigma} - Q\Sigma^{3/2}) \,. \tag{22}$$

再次得到 $\frac{d\bar{w}}{dt} = 0$。对于 Σ的常微分方程，我们有:

$$\frac{d\Sigma}{dt} = \Big(\frac{\eta^2 S^2}{2} - 2\Big)\Sigma - \eta^2 S Q \Sigma^2 + \frac{\eta^2 Q^2}{2}\Sigma^3 \,. \tag{23}$$

令 $\frac{d\Sigma}{dt} = 0$，即可求得 Σ 的稳态解。其中一个稳态解为 $\Sigma^* = 0$。非零解满足以下二次方程：

$$\frac{\eta^2 Q^2}{2}(\Sigma^*)^2 - \eta^2 S Q \Sigma^* + \frac{\eta^2 S^2}{2} - 2 = 0\,. \tag{24}$$

使用求根公式，我们可以求出非零根：

$$\Sigma^* = \frac{S}{Q} \pm \frac{2}{\eta Q}\,. \tag{25}$$

根 $\Sigma^* = S/Q + 2/\eta Q$ 是不稳定的，对应于四次势的失控区域。我们关注的根是 $\Sigma^* = S/Q - 2/\eta Q$。由于只有正值的 Σ才有定义，因此我们要求：

$$\frac{S}{Q} - \frac{2}{\eta Q} > 0 \implies S > \frac{2}{\eta}\,.$$

当 $S > 2/\eta$时，此解是稳定的。

这一分析从机制上解释了 Cohen 等人[2021]的实证观察。在训练过程中，随着迭代点接近最小值，尖锐度 S 往往会增加——这就是渐进式尖锐化。当 $S < 2/\eta$时，杆收缩：迭代点向最小值收敛，尖锐度随之增加。当 $S > 2/\eta$时，杆扩张：迭代点沿海森矩阵的最大特征值对应的特征向量发散，但这种扩张会采样到损失景观中高阶项降低有效曲率的区域。这就是自稳定化 [Damian 等人， 2023]：损失函数的非线性会在迭代点因不稳定性而发散时，通过降低尖锐度自然恢复稳定性。

5.4 不动点与稳定性

现在我们更一般地讨论 Rod Flow 的不动点。不动点是满足 $\frac{d\bar{w}}{dt} = 0$ 和 $\frac{d\Sigma}{dt} = 0$ 的一对 ($\bar{w}^*$， Σ^*)。

令 $S(\bar{w}^*)$ 表示 $\bar{w}^*$ 处的尖锐度。当 $\bar{w}^*$ 是损失的临界点（$\nabla L(\bar{w}^*) = 0$）时，存在两种类型的不动点：

静止不动点： $\Sigma^* = 0$。当 $S(\bar{w}^*) < 2/\eta$ 时，它们是稳定的。梯度下降会收敛到 $\bar{w}^*$，且不会发生持续振荡。

稳定性边缘不动点： $\Sigma^* > 0$。当 $S(\bar{w}^*) > 2/\eta$，且损失中的高阶项提供稳定作用时（如四次函数的例子），就会出现这类不动点。展度 Σ^* 反映了围绕 $\bar{w}^*$ 发生的持续振荡的振幅。

如果 $\bar{w}^*$ 是损失的临界点，那么 ($\bar{w}^*$,0) 始终是 Rod Flow 的不动点。梯度流与 Rod Flow 的关键区别在于，对于 Rod Flow，该不动点的稳定性取决于 $\Sigma = 0$处的尖锐度 $\bar{w}^*$。附录 B 对不动点进行了更详细的分析。

6 与 CentralFlow 的比较

中心流 [Cohen 等人， 2025] 是针对稳定性边缘提出的先前连续时间模型。我们比较这两种方法。

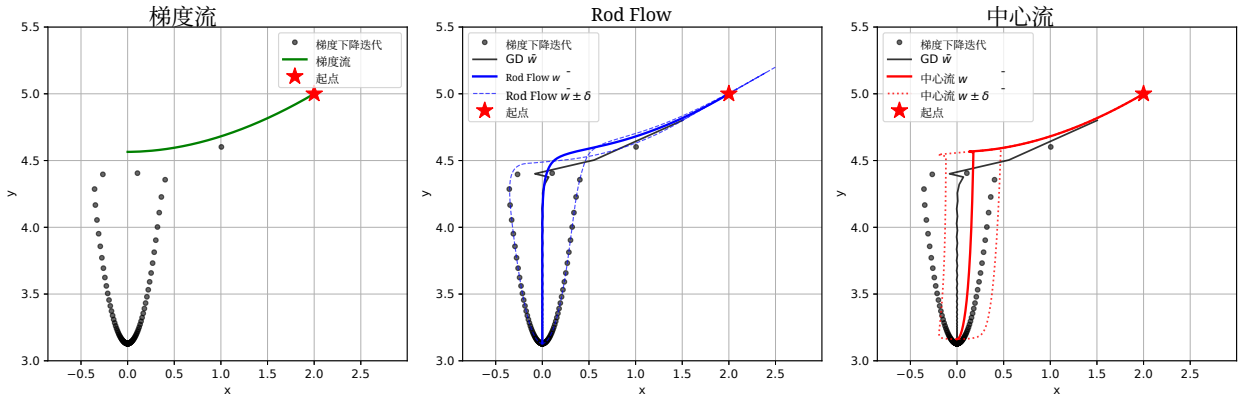


图 4：**二维平方根损失上的比较** $L(x, y) = \sqrt{1 + (xy)^2}$。该损失在坐标轴上存在极小值，且距离原点越远，尖锐度越高。当局部最小值过于尖锐时，梯度下降会不断反弹，直到尖锐度降至 $2/\eta$。无论尖锐度如何，梯度流都会收敛到局部最小值。Rod Flow 能精确追踪梯度下降。中心流则沿着锐度流形移动。

6.1 中心流概述

中心流的核心见解是，尽管精确的振荡动力学难以分析，但时间平均轨迹却易于处理。中心流为这条平滑轨迹推导出一个常微分方程，并将 $\bar{w}$ 视为振荡迭代点的局部平均值。它定义了一个矩阵 Σ，用于表示稳定性边缘处梯度下降迭代点振荡的协方差。 Σ被限制在临界特征向量所张成的子空间内——这些特征向量的特征值接近 $2/\eta$。

中心的演化遵循：

$$\frac{d\bar{w}}{dt} = -\eta \nabla L(\bar{w}) - \frac{\eta}{2} \operatorname{Tr}(\Sigma \nabla^3 L)\,, \tag{26}$$

其中，第一项是梯度流，第二项则说明振荡如何使运动偏向曲率较低的区域。

协方差 Σ 也是动态的，但与 $\bar{w}$不同，它的演化并非由常微分方程描述。相反，Σ 由每一时刻求解一个半定互补问题（SDCP）来确定。

中心流也有几何解释。它可以被视为投影梯度流：动力学沿梯度方向演化，但受到可行域 $\{\bar{w}$ 的约束： $S(\bar{w}) \leq 2/\eta\}$，其中Σ 作为施加该约束的拉格朗日乘子。关于中心流的更多细节见 D 节。

6.2 主要差异

We 简要介绍 Rod Flow 与中心流之间的一些主要差异。

推导。中心流的推导涉及临界子空间中的泰勒展开，以及用于稳定性约束的拉格朗日乘子。Rod Flow 则通过精确差分方程和标准后向误差分析推导得出。

简洁性。 由于 Σ 由 SDCP 确定，中心流的动力学较为复杂。相比之下，Rod Flow 给出了更简单、也更便于数学分析的常微分方程（正如我们在第 5 节所展示的）。

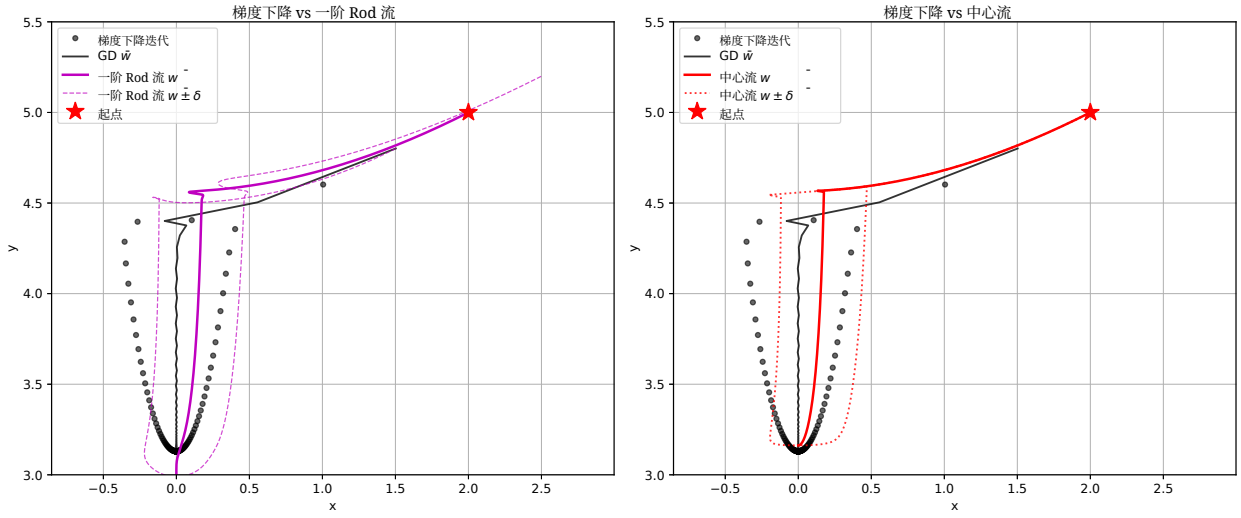


图5： **一阶 Rod 流与中心流的比较**。

物理解释。 中心流将协方差解释为围绕 $\bar{w}$ 的高斯扩散。Rod Flow 则提供了更直观的图景：杆的端点位于 $\bar{w} \pm \delta$，对应于迭代点的实际位置。

评估点。 中心流在中心 $\bar{w}$ 处计算梯度，并使用泰勒展开来考虑位移。Rod Flow 在杆的端点 $\bar{w} \pm \delta$处计算梯度。这使得 Rod Flow 的非局部性成为一种优势：它直接对迭代点实际经过的损失景观进行采样。

计算成本。 中心流需要在每个时间步求解一个 SDCP——一个复杂的优化问题。Rod Flow 仅使用梯度评估和海森向量积，这两者都是 $O(p)$ 运算。Rod Flow 的 Σ 矩阵以低秩形式存储（秩为 3 即可）。有关 Rod Flow 计算实现的更多信息，请参见 C节。

6.3 流之间的联系

中心流可以看作 Rod Flow 的泰勒近似。将 $\nabla L_{\pm}$ 在 δ 中展开到二阶：

$$\nabla L(\bar{w} \pm \delta) \approx \nabla L(\bar{w}) \pm \nabla^2 L(\bar{w})\delta + \frac{1}{2}\nabla^3 L(\bar{w})[\delta, \delta]\,. \tag{27}$$

考虑关于 $\bar{w}$ 的常微分方程。如果我们舍弃后向误差分析项（因为其阶数为 η^2），并代入上述关于 $\nabla L_{\pm}$ 的表达式，则有：

$$\frac{d\bar{w}}{dt} \approx -\eta \nabla L(\bar{w}) - \frac{\eta}{2}\nabla^3 L(\bar{w})[\delta, \delta]\,. \tag{28}$$

如果 δ 与 $\nabla^2 L$ 的最大特征值对应的特征向量对齐，则 $\nabla^3 L(\bar{w})[\delta, \delta] \approx \|\delta\|^2 \nabla S$，从而恢复关于 $\bar{w}$ 的中心流常微分方程。

将 Σ 的常微分方程按 δ 展开至二阶，我们也可以推导出 Σ 的简化常微分方程。在此展开中，我们将忽略梯度项，仅保留包含海森矩阵的项：

$$\frac{d\Sigma}{dt} \approx \frac{\eta^2}{2}(\nabla^2 L \cdot \delta)(\nabla^2 L \cdot \delta)^\top - 2\Sigma \tag{29}$$

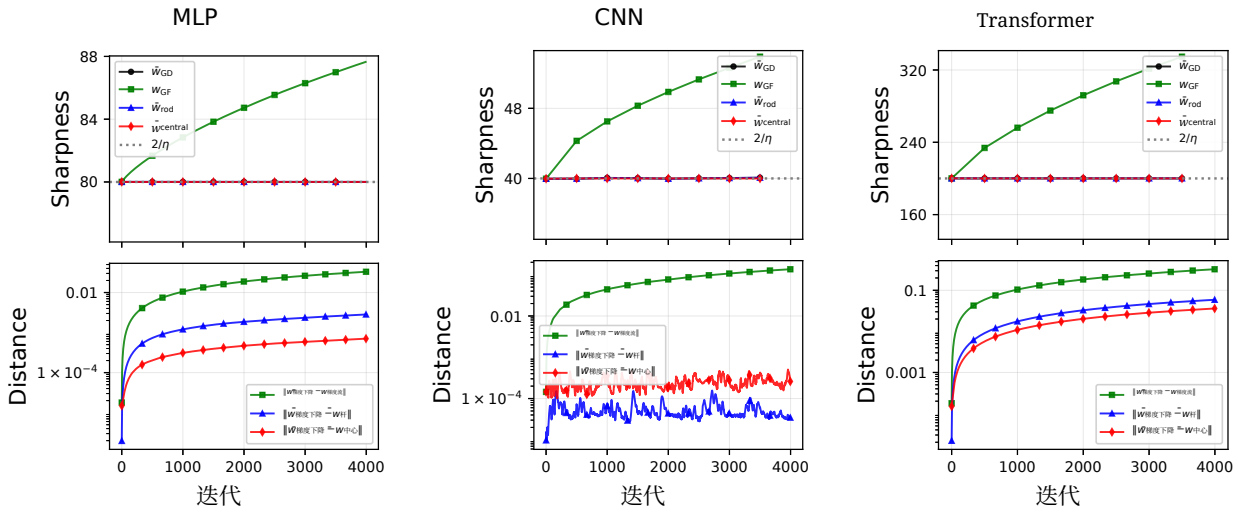


图6：**神经网络上的实验结果**。比较三种不同架构（多层感知机、卷积神经网络和 Transformer）上的梯度流、Rod Flow 和中心流。这三种流在经过梯度下降预热阶段后进行初始化。梯度流持续变尖锐，而 Rod Flow 和中心流则表现出自稳定性。就与梯度下降的距离而言，Rod Flow 和中心流的精度都比梯度流高出几个数量级。

我们将由此得到的常微分方程组称为“一阶 Rod 流”，因为这些方程是按 Σ 展开至一阶的。在玩具示例中，一阶 Rod 流的轨迹与中心流的轨迹十分接近，不过两种流的时间尺度有所不同。图5对此进行了展示。

7 实验

我们在玩具问题和具有代表性的神经网络架构中，对比梯度下降（GD）、梯度流（GF）、Rod Flow（RF）和中心流（CF）。

7.1 平方根损失（2D）

我们考虑以下二维损失函数，该函数由 Ahn 等人 [2023] 研究，作为稳定性边缘的一个最简示例（图4）。

$$L(x, y) = \sqrt{1 + (xy)^2}. \tag{30}$$

该损失函数在 x 轴和 y 轴上存在极小值（其中 $xy = 0$），且沿每条轴距离原点越远，极小值的尖锐度越高。

根据初始条件的不同，梯度下降迭代可能会表现出稳定性边缘现象。如果局部最小值过于尖锐，迭代点就会沿垂直轴来回跳动，使尖锐度降低，直至收敛到尖锐度约为 $2/\eta$的局部最小值。

在平方根损失上进行测试时，三种流呈现出不同的行为。梯度流会收敛到最近的局部最小值，无论其尖锐度如何。Rod Flow 能精确追踪梯度下降迭代：$\bar{w}_{\mathrm{RF}}$ 沿着 y-轴快速收敛，而包络线 $\bar{w}_{\mathrm{RF}} \pm \delta_{\mathrm{RF}}$ 与实际梯度下降迭代点的位置高度吻合。中心流沿着锐度流形运动，导致$\bar{w}_{\mathrm{CF}}$ 位于实际迭代点与迭代点平均值之间。

7.2 神经网络

为了进一步比较不同的流，我们在多种神经网络架构上进行了实验：三层多层感知机、三层卷积神经网络和两层Transformer。

由于稳定性边缘开始时存在湍流，我们运行梯度下降，直到迭代点安全地处于稳态稳定性边缘区域（第三阶段；见第3节）。随后，我们在此预热期之后，于梯度下降迭代所在位置初始化全部三种流（梯度流、杆流和中心流）。结果如图6所示。

Rod Flow 和中心流都明显优于梯度流。虽然梯度流的尖锐度不断增大，但 Rod Flow 和中心流都表现出自稳定，其各自的尖锐度保持在临界值 $2/\eta$。此外，中心流和杆流都更接近实际的梯度下降轨迹，衡量标准是它们各自的 $\bar{w}$ 与迭代点真实平均值之间的距离。

Rod Flow 在卷积神经网络上的表现强于中心流，而在多层感知机和 Transformer 上的表现则稍差。这些性能差异可能归因于海森矩阵谱中与架构相关的差异。更多实验细节见 E节。

8 结论

我们提出了 Rod Flow，这是一种用于梯度下降的连续时间模型，能够准确捕捉稳定性边缘处的动力学。通过将梯度下降建模为一个扩展对象——一根中心位于 $\bar{w}$ 、半长为 δ 的杆——我们得到一个常微分方程系统，用于追踪迭代点的平均位置和半差。

我们通过精确差分方程和后向误差分析进行推导，为该模型提供了严谨的理论基础。从理论上看，Rod Flow 能正确预测 $2/\eta$ 稳定性阈值，并在四次势函数中展现出自稳定性。从实证结果看，Rod Flow 比梯度流更能捕捉稳定性边缘现象，其表现与中心流相当，同时实现起来更简单。

局限性。 与所有梯度下降的常微分方程近似一样，Rod Flow 假设动力学足够平滑。在大型神经网络的稳定性边缘处，梯度范数偶尔会激增数个数量级，导致常微分方程无法捕捉的离散跃变。明确 Rod Flow 在何种损失函数正则性条件下仍能保持准确，是一个尚待解决的问题。此外，尽管 Rod Flow 的计算效率高于其他连续时间模型，但与梯度下降相比仍然代价高昂，因此不适合作为实际训练的替代方案。最后，渐进式尖锐化背后的机制——尖锐度为何会首先增加到 $2/\eta$ ——仍未得到解释。

未来展望。 我们希望，使用扩展对象对离散时间算法进行建模这一思路，能够为建立非凸优化的一般理论提供启发——超越深度学习这一特定场景。

致谢

我们感谢中心流论文的作者公开其代码。我们的实现基于 https://github.com/alex-damian/EOS 和 https://github. com/centralflows/centralflows 中的代码。

参考文献

A. Agarwala、F. Pedregosa 与 J. Pennington。二阶回归模型表现出向稳定性边缘的渐进式尖锐化。载于 第*40*届国际机器学习会议论文集，ICML'23。JMLR，2023年。K. Ahn、J. Zhang 与 S. Sra。理解梯度下降的不稳定收敛。载于国际机器学习会议，第247–257页。PMLR，2022。K. Ahn、S. Bubeck、S. Chewi、Y. T. Lee、F. Suarez 与 Y. Zhang。通过稳定性边缘学习阈值神经元。神经信息处理系统进展，36:19540–19569，2023年。S. Arora、Z. Li 与 A. Panigrahi。理解深度学习中处于稳定性边缘的梯度下降。载于 国际机器学习会议，第948–1024页。PMLR，2022。D. Barrett 与 B. Dherin。隐式梯度正则化。载于 国际学习表征会议，2021年。L. Chen 与 J. Bruna。借助两步梯度更新超越稳定性边缘。载于 国际机器学习会议，第4330–4391页。PMLR，2023年。X. Chen、K. Balasubramanian、P. Ghosal 与 B. K. Agrawalla。从稳定性到混沌：分析二次回归中的梯度下降动力学。机器学习研究汇刊，2024年。J. Cohen、S. Kaur、Y. Li、J. Z. Kolter 与 A. Talwalkar。神经网络上的梯度下降通常发生在稳定性边缘。载于国际学习表征会议，2021年。J. Cohen、A. Damian、A. Talwalkar、J. Z. Kolter 与 J. D. Lee。借助中心流理解深度学习中的优化。载于第十三届国际学习表征会议，2025年。A. Damian、E. Nichani 与 J. D. Lee。自稳定：梯度下降在稳定性边缘的隐式偏差。载于第十一届国际学习表征会议，2023年。S. Di Giovacchino、D. J. Higham 与 K. C. Zygalakis。后向误差分析与随机优化算法的定性行为：应用于随机坐标下降。计算动力学杂志，2024年。E. Hairer、C. Lubich 与 G. Wanner。几何数值积分：常微分方程的保结构算法。施普林格，2006年。K. Jiang、J. Cohen 与 Y. Li。理解神经正切核在稳定性边缘的演化。载于第三十九届神经信息处理系统年会，2025年。A. Lewkowycz、Y. Bahri、E. Dyer、J. Sohl-Dickstein 与 G. Gur-Ari。深度学习的大学习率阶段：弹射机制。*arXiv*预印本 *arXiv:2003.02218*，2020年。Q. Li、C. Tai 与 W. E.。随机修正方程与自适应随机梯度算法。载于国际机器学习会议，第2101–2110页。PMLR，2017年。

C. Ma、D. Kunin、L. Wu 和 L. Ying。《超越二次近似：神经网络损失景观的多尺度结构》。*arXiv* 预印本 *2204.11326*，2022年。M. Rosca、Y. Wu、C. Qin 和 B. Dherin。《关于梯度下降动力学及深度学习不稳定性的连续时间模型》。机器学习研究汇刊，2023年。B. Shi、S. S. Du、W. Su 和 M. I. Jordan。《动量方法的连续时间分析》。机器学习研究期刊，22(1):5915–5957，2021年。S. L. Smith、B. Dherin、D. Barrett 和 S. De。《随机梯度下降中隐式正则化的起源》。载于国际学习表征会议，2021年。M. Song 和 C. Yun。《轨迹对齐：通过分岔理论理解稳定性边缘现象》。载于神经信息处理系统进展，第36卷，2023年。Y. Wang、Z. Xu、T. Zhao 和 M. Tao。《良好的正则性会产生较大的学习率隐式偏差：稳定性边缘、平衡与弹射》。*arXiv* 预印本 *2310.17087*，2023年。J. H. Wilkinson。《style id='13'>代数过程中的舍入误差。Prentice-Hall，1963年。J. H. Wilkinson。《style id='15'>代数特征值问题。克拉伦登出版社，牛津，1965年。X. Zhu、Z. Wang、X. Wang、M. Zhou 和 R. Ge。《通过一个极简示例理解稳定性边缘训练动力学》。载于国际学习表征会议，2023年。

A 后向误差分析

A.1 历史背景

我们简要介绍后向误差分析，这是一种数值分析技术，可让我们通过连续动力学的视角理解离散时间算法。

“后向误差分析”这一术语起源于数值线性代数，主要源于Wilkinson在20世纪60年代的工作 [Wilkinson, 1963, 1965]。经典设定考虑的是计算问题 $y = f(x)$的近似解 $\tilde{y}$ 。前向误差分析问的是：计算出的答案与真实答案相差多远？这涉及对量 $\|\tilde{y} - f(x)\|$给出界。后向误差分析则将问题反过来：我们能否找到对输入的一个扰动，使得计算出的答案恰好正确？这涉及寻找 Δx ，使得 $\tilde{y} = f(x + \Delta x)$。后向误差分析视角的力量在于：如果后向误差 $\|\Delta x\|$ 很小，且问题是良态的，我们就可以断定前向误差也必然很小。

“后向误差分析”的一种相关但不同的用法出现在常微分方程数值积分器的研究中，尤其是在几何积分文献中。此时问题变为：给定一个并不能精确求解 $\dot{x} = V(x)$的数值方法（例如欧拉法），是否存在一个修正的常微分方程 $\dot{x} = V(x) + \Delta t\, V_1(x) + \Delta t^2 V_2(x) + \cdots$，使得数值迭代点确实落在其上？其逻辑与Wilkinson的方法相平行：向量场是输入，流映射是输出。我们不是把离散化误差视为前向误差（迭代点偏离真实流），而是把迭代点视为精确地遵循一个受扰动的系统。我们从离散动力学“向后”回溯，以找出它们实际求解的连续系统。

这就是我们利用后向误差分析推导 Rod Flow 的方式。

A.2 二次情形的后向误差分析

自然会产生一个问题：为什么我们对变换变量 $\bar{w}$应用后向误差分析，而不是直接对原始迭代点 w 应用？答案在于修正向量场 V的˜微扰展开的收敛半径：恰恰是在稳定性边缘区域，该展开会失效。对于二次损失的情形，我们可以最清楚地看到这一点，因为我们能够精确地求出修正向量场在 η 中所有阶的表达式。

考虑二次损失 $L(w) = \frac{1}{2} S w^2$。梯度流常微分方程为：

$$\frac{dw}{dt} = -\eta S w,$$

其中，我们采用这样的约定：一个单位时间对应一个梯度下降步。梯度下降递推关系为：

$$w_{t+1} = w_t - \eta S w_t = (1 - \eta S) w_t. \tag{31}$$

˜我们要寻找一个常微分方程 $\dot{w} = V(w)$，使其流经过这些离散迭代点。

˜**简便的方法**。 无需借助微扰理论，我们也能优雅地求出 V。经过 t 个时间单位后，我们有：

$$w_t = (1 - \eta S)^t w_0 = e^{t \log(1 - \eta S)} w_0. \tag{32}$$

通过观察可知，这是常微分方程的解：

$$\dot{w} = \log(1 - \eta S)\, w. \tag{33}$$

将 ηS 中的速率泰勒展开可得：

$$\log(1-\eta S) = -\eta S - \frac{(\eta S)^2}{2} - \frac{(\eta S)^3}{3} - \cdots \tag{34}$$

首项 $-\eta S\, w = -\eta \nabla L(w)$ 恢复了梯度流，而 η 中包含高阶修正。

微扰方法。使用完整的后向误差分析工具推导修正向量场很有启发性，因为这种方法可以推广到二次函数之外。我们假设˜存在一个一阶常微分方程 $\dot{w} = V(w)$，其轨迹经过这些迭代点。通过泰勒展开，我们可以用 $w(t)$ 的时间导数来表示 $w(t+1)$：

$$w(t+1) - w(t) = \dot{w} + \frac{1}{2}\ddot{w} + \frac{1}{6}\dddot{w} + \cdots = \sum_{n=1}^{\infty} \frac{1}{n!}\frac{d^n w}{dt^n} \tag{35}$$

然后，利用梯度下降递推关系可得：

$$\sum_{n=1}^{\infty} \frac{1}{n!}\frac{d^n w}{dt^n} = -\eta S w \tag{36}$$

˜ ˜为了求出 V，我们需要用 V 表示 w 的时间导数。为此，我们将利用以下事实：沿着解，任意量 f 的全时间导数为：

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \nabla f \cdot \frac{dw}{dt}. \tag{37}$$

由于我们的系统是自治的（动力学没有显式的时间依赖），偏时间导数为零，因此有：

$$\frac{df}{dt} = \nabla f \cdot \tilde{V}. \tag{38}$$

˜这使我们能够用 V 表示 w 的时间导数：

$$\dot{w} = \tilde{V}, \tag{39}$$

$$\ddot{w} = \nabla\tilde{V} \cdot \tilde{V}, \tag{40}$$

$$\dddot{w} = \nabla(\nabla\tilde{V} \cdot \tilde{V}) \cdot \tilde{V} = (\nabla\tilde{V})^2\tilde{V} + \nabla^2\tilde{V}[\tilde{V}, \tilde{V}]. \tag{41}$$

现在我们采用一个试探形式：由于梯度流关于 w 是线性的，我们预期修正动力学仍然是线性的，因此假设 $\tilde{V}(w) = -Cw$ for some constant C。这意味着 $\nabla^2\tilde{V} = 0$，从而大大简化了高阶时间导数。事实上，我们可以证明：

$$\frac{d^n w}{dt^n} = (\nabla\tilde{V})^{n-1}\tilde{V} \tag{42}$$

对所有 $n \geq 1$ 均成立。

我们将使用归纳法。基础情形 $n = 1$ 显然成立。归纳步骤中，假设公式对 n成立。于是可得：

$$\begin{aligned}\frac{d^{n+1}w}{dt^{n+1}} &= \nabla\left[(\nabla\tilde{V})^{n-1}\tilde{V}\right] \cdot \tilde{V} \\ &= (\nabla\tilde{V})^n\tilde{V} + (n-1)\nabla^2\tilde{V}[(\nabla\tilde{V})^{n-2}\tilde{V}, \tilde{V}] \\ &= (\nabla\tilde{V})^n\tilde{V},\end{aligned} \tag{43}$$

其中，最后一个等号利用了 $\nabla^2\tilde{V}=0$。根据我们的线性拟设 $\tilde{V}(w)=-Cw$，可得 $-\nabla V=-C$。由此可见：

$$\frac{d^n w}{dt^n}=(-C)^n w. \tag{44}$$

可以证明：

$$\begin{aligned} w(t+1)-w(t) &= \sum_{n=1}^{\infty}\frac{1}{n!}\frac{d^n w}{dt^n} \\ &= \sum_{n=1}^{\infty}\frac{(-C)^n}{n!}w \\ &= (e^{-C}-1)w. \end{aligned} \tag{45}$$

令其等于方程 (36) 的右侧，我们可以解出 C：

$$(e^{-C}-1)w=-\eta S w \implies C=-\log(1-\eta S), \tag{46}$$

这与简单的方法一致。

收敛半径。 修正速率的表达式 $\log(1-\eta S)$ 揭示了微扰方法的一个根本局限：级数展开仅在 $\eta S<1$ 时收敛。这并非单纯的技术问题，而是反映了连续时间近似确实失效了。当 $\eta S>1$ 时，因子 $(1-\eta S)$ 为负。因此，迭代点的符号交替变化。不存在一阶常微分方程能够产生符号交替轨迹。

这就是为什么在研究稳定性边缘动力学时，我们不会直接对 w 应用后向误差分析。在 $\eta S=2$——即稳定性边界阈值处，迭代点以固定振幅振荡，没有光滑常微分方程能够对其进行插值。Rod Flow 旨在处理这种直接后向误差分析失效的情形。

A.3 $\bar{w}$ 方程的推导

现在我们应 用后向误差分析推导 $\bar{w}$ 动力学。从方程 (6):

$$\bar{w}_{t+1}-\bar{w}_t=-\frac{\eta}{2}\big[\nabla L(\bar{w}_t+\delta_t)+\nabla L(\bar{w}_t-\delta_t)\big]=:V_{\bar{w}}(\bar{w}_t,\delta_t). \tag{47}$$

应用后向误差分析公式（将 δ 视为 $\bar{w}$ 方程中的固定量），可得：

$$\frac{d\bar{w}}{dt}=V_{\bar{w}}-\frac{1}{2}\nabla_{\bar{w}}V_{\bar{w}}\cdot V_{\bar{w}}+O(\eta^3). \tag{48}$$

我们可以得到 $V_{\bar{w}}$ 的雅可比矩阵：

$$\nabla_{\bar{w}}V_{\bar{w}}=-\frac{\eta}{2}\big[\nabla^2 L(\bar{w}+\delta)+\nabla^2 L(\bar{w}-\delta)\big]. \tag{49}$$

由此得到修正向量场：

$$\begin{aligned} \tilde{V}_{\bar{w}} &= -\frac{1}{2}\nabla_{\bar{w}}V_{\bar{w}}\cdot V_{\bar{w}} \\ &= -\frac{1}{2}\cdot\Big(-\frac{\eta}{2}\Big)\big[\nabla^2 L(\bar{w}+\delta)+\nabla^2 L(\bar{w}-\delta)\big]\cdot\Big(-\frac{\eta}{2}\Big)\big[\nabla L(\bar{w}+\delta)+\nabla L(\bar{w}-\delta)\big] \\ &= -\frac{\eta^2}{8}\big[\nabla^2 L(\bar{w}+\delta)+\nabla^2 L(\bar{w}-\delta)\big]\big[\nabla L(\bar{w}+\delta)+\nabla L(\bar{w}-\delta)\big]. \end{aligned} \tag{50}$$

由此得到方程（4）。

A.4 为什么不对 Σ进行后向误差修正

Σ 方程包含衰减项 -2Σ，因此其特征时间尺度为 $O(1)$。相比之下，$\bar{w}$ 以速率 $O(\eta)$演化。当 $\eta \ll 1$时，这种时间尺度分离意味着 Σ 达到平衡的速度远快于 $\bar{w}$ 发生显著变化的速度。在准平衡状态下， Σ满足：

$$\Sigma \approx \frac{\eta^2}{8}\big[\nabla L(\bar{w}+\delta)\otimes\nabla L(\bar{w}+\delta)+\nabla L(\bar{w}-\delta)\otimes\nabla L(\bar{w}-\delta)\big]. \tag{51}$$

我们出于两个原因省略了对 Σ 的后向误差修正。首先，由于 $\bar{w}$ 在一个时间步内变化很小， Σ 的准平衡值（ Σ会迅速趋向该值）也变化很小。其次，对 Σ 的任何修正都将与其时间导数成正比，而当 Σ接近准平衡时，该导数接近于零。综合这些观察可知，对 Σ 进行后向误差修正的影响可以忽略不计。

B 不动点分析

B.1 四次势

考虑一维中的偶四次势：

$$L(w)=\frac{S}{2}w^2+\frac{Q}{4}w^4\,, \tag{52}$$

现在我们将对四次势进行更为深入的分析。与第 5 节一样，我们要求二次项系数 S 大于零。但现在我们将同时考虑正的和负的四次项系数 Q。我们将考虑四种情形： $S<2/\eta$ 或 $S>2/\eta$，分别与 $Q<0$ 或 $Q>0$组合。

B.1.1 推导 Σ 动力学

梯度表示为 $\nabla L(w)=Sw+Qw^3$。在 $\bar{w}=0$ 且满足 $\delta=\sqrt{\Sigma}$ 时，杆两端点处的梯度为：

$$\nabla L(\pm\delta)=\pm\big(S\sqrt{\Sigma}+Q\Sigma^{3/2}\big)\,. \tag{53}$$

由于势能具有对称性， $\frac{d\bar{w}}{dt}=0$：中心保持在原点。支配Σ 的常微分方程为：

$$\frac{d\Sigma}{dt}=\Big(\frac{\eta^2S^2}{2}-2\Big)\Sigma+\eta^2SQ\,\Sigma^2+\frac{\eta^2Q^2}{2}\Sigma^3\,. \tag{54}$$

为方便起见，将右侧定义为 $f(\Sigma)$：

$$f(\Sigma)=\Big(\frac{\eta^2S^2}{2}-2\Big)\Sigma+\eta^2SQ\,\Sigma^2+\frac{\eta^2Q^2}{2}\Sigma^3\,. \tag{55}$$

B.1.2 不动点

令 $f(\Sigma)=0$，我们发现 $\Sigma^*=0$ 始终是不动点。提出因子 Σ 后，非零不动点满足：

$$\frac{\eta^2Q^2}{2}(\Sigma^*)^2+\eta^2SQ\,\Sigma^*+\Big(\frac{\eta^2S^2}{2}-2\Big)=0\,. \tag{56}$$

应用求根公式，可得：

$$\Sigma^* = \frac{-\eta^2 SQ \pm \sqrt{\eta^4 S^2 Q^2 - 4 \cdot \frac{\eta^2 Q^2}{2}\left(\frac{\eta^2 S^2}{2} - 2\right)}}{\eta^2 Q^2}. \tag{57}$$

令 Δ 表示判别式。判别式可简化为：

$$\begin{aligned}\Delta &= \eta^4 S^2 Q^2 - \eta^2 Q^2\left(\eta^2 S^2 - 4\right) \\ &= \eta^4 S^2 Q^2 - \eta^4 S^2 Q^2 + 4\eta^2 Q^2 \\ &= 4\eta^2 Q^2 \,. \end{aligned} \tag{58}$$

因此 $\sqrt{\Delta} = 2\eta|Q|$，不动点为：

$$\Sigma^* = \frac{-\eta^2 SQ \pm 2\eta|Q|}{\eta^2 Q^2} = -\frac{1}{Q}\left(S \pm \frac{2}{\eta}\right). \tag{59}$$

注意，由于 Σ 表示展度，只有非负值才具有物理意义 .

B.1.3 稳定性分析

要确定不动点的稳定性，我们需要计算 f 在该点处的导数。当 $f'(\Sigma^*) < 0$ 时，不动点稳定；当 $f'(\Sigma^*) > 0$ 时，不动点不稳定。我们有：

$$f'(\Sigma) = \frac{\partial}{\partial \Sigma}\frac{d\Sigma}{dt} = \left(\frac{\eta^2 S^2}{2} - 2\right) + 2\eta^2 SQ\,\Sigma + \frac{3\eta^2 Q^2}{2}\Sigma^2 \,. \tag{60}$$

当 $\Sigma^* = 0$ 时，我们有：

$$f'(0) = \frac{\eta^2 S^2}{2} - 2 \,. \tag{61}$$

因此， $\Sigma^* = 0$ 不动点在以下条件下稳定：

$$f'(0) < 0 \implies S < \frac{2}{\eta}. \tag{62}$$

在非零不动点处，我们可以利用如下事实简化 $f'(\Sigma^*)$ 的表达式：根据定义，每个非零不动点 Σ^* 都必须满足方程（56）。代入不动点条件：

$$f'(\Sigma^*) = \left(\frac{\eta^2 S^2}{2} - 2\right) + 2\eta^2 SQ\,\Sigma^* + \frac{3\eta^2 Q^2}{2}(\Sigma^*)^2 \tag{63}$$

$$= \left(-\eta^2 SQ\,\Sigma^* - \frac{\eta^2 Q^2}{2}(\Sigma^*)^2\right) + 2\eta^2 SQ\,\Sigma^* + \frac{3\eta^2 Q^2}{2}(\Sigma^*)^2 \tag{64}$$

$$= \eta^2 SQ\,\Sigma^* + \eta^2 Q^2(\Sigma^*)^2 \tag{65}$$

$$= \eta^2 Q^2 \Sigma^* \left[\frac{S}{Q} + \Sigma^*\right]. \tag{66}$$

定义：

$$\Sigma^*_{\pm} = -\frac{1}{Q}\left(S \pm \frac{2}{\eta}\right)$$

我们希望求解 $\Sigma_{\pm}^*$ 在何时具有物理意义（大于零），以及它们在何时稳定。

回想一下 $S > 0$，我们有：

$$\Sigma_+^* > 0 \implies Q < 0. \tag{67}$$

并且：

$$\Sigma_-^* > 0 \implies S > \frac{2}{\eta} \text{ and } Q < 0 \quad \text{OR} \quad S < \frac{2}{\eta} \text{ and } Q > 0. \tag{68}$$

关于它们的稳定性，我们有：

$$\begin{aligned} f'(\Sigma_{\pm}^*) &= \eta^2 Q^2 \Sigma_{\pm}^* \left[\frac{S}{Q} + \Sigma_{\pm}^*\right] \\ &= \eta^2 Q^2 \left[-\frac{1}{Q}\left(S \pm \frac{2}{\eta}\right)\right]\left[\frac{S}{Q} - \frac{1}{Q}\left(S \pm \frac{2}{\eta}\right)\right] \\ &= \eta^2 Q^2 \left[-\frac{1}{Q}\left(S \pm \frac{2}{\eta}\right)\right]\left[\mp\frac{2}{\eta Q}\right] \\ &= \pm 2\eta \left(S \pm \frac{2}{\eta}\right) \\ &= 2\eta \left(\frac{2}{\eta} \pm S\right) \end{aligned} \tag{69}$$

要使不动点稳定，我们要求 $f'(\Sigma^*) < 0$。

对于 Σ_+^*，导数为 $2\eta S + 4$。由于 $S,\ \eta > 0$，因此它始终为正。所以 Σ_+^* 始终不稳定。

对于 Σ_-^*，导数为 $-2\eta S + 4$。稳定性要求：

$$f'(\Sigma_-^*) = -2\eta S + 4 < 0 \implies S > 2/\eta$$

结合 $\Sigma_-^* > 0$ 的存在条件，我们发现，只有当 $S > 2/\eta$ 和 $Q < 0$ 时，才会出现稳定的非零不动点。

B.1.4 情况分析

只有当 $S > 2/\eta$ 和 $Q < 0$ 时，才会存在稳定的非零不动点。我们可以通过考察全部四种情况及其物理意义来理解这一点。

情况 1： $S < 2/\eta$ 且 $Q > 0$。

原点处的尖锐度低于临界阈值 $2/\eta$，因此原点是稳定的，振荡会衰减至零。正四次项意味着势能离开原点后会变得更加陡峭。不稳定不动点 Σ_-^* 充当分界轨道：如果 Σ 的初始值高于此阈值，它就会无限增长；如果初始值低于此阈值，它就会衰减至零。

情况 2： $S > 2/\eta$ 且 $Q > 0$。

原点处的尖锐度超过 $2/\eta$，因此原点是不稳定不动点。正四次项意味着势能在远离原点时会变得更加陡峭。在这个模型中，振荡会无限增长。实际上，在振幅较大时，高阶项最终会带来稳定化。

情况 3： $S < 2/\eta$ 和 $Q < 0$。

原点是稳定的，因为 $S < 2/\eta$。负四次项意味着势能在远离原点时趋于展平，但由于原点本身已经具有吸引性，这不会改变定性行为。不稳定不动点 Σ_+^* 标志着吸引域的边界：从其外部初始化的轨迹会进入势能向下无界的失控区域。在现实的损失景观中，正的高阶项会阻止这种失控发散。

情况 4： $S > 2/\eta$ 且 $Q < 0$。

这是<strong? self-stabilization 情况。原点处的尖锐度超过 $2/\eta$，因此小幅振荡会被放大。然而，负四次项意味着有效曲率在更大振幅下会降低：杆会不断伸展，直到其端点采样到势能中更平坦的区域，并在 Σ^*处达到稳定平衡。较大的不动点 Σ_+^* 是不稳定的，并标志着吸引域的边界。

Case	条件	不动点	稳定性
1	$S < 2/\eta, \quad Q > 0$	$0, \quad \Sigma_-^*$	稳定，不稳定
2	$S > 2/\eta, \quad Q > 0$	0	不稳定
3	$S < 2/\eta, \quad Q < 0$	$0, \Sigma_+^*$	稳定、不稳定
4	$S > 2/\eta, Q < 0$	$0, \Sigma_-^*, \Sigma_+^*$	不稳定、稳定、不稳定

表 1：四次势的不动点与稳定性摘要。

B.1.5 与神经网络训练的联系

在神经网络优化的背景下，情况 4 最为相关。

条件 $S > 2/\eta$ 对应稳定性边缘区域。条件 $Q < 0$ 通常出现在局部极小值附近：损失曲面通常具有正曲率，且随着与最小值距离增大而减小。虽然 $Q < 0$ 会使四次势无下界，但实际的损失曲面具有正的高阶项，可以防止失控发散。四次分析抓住了自稳定机制的本质，同时仍保持解析上的可处理性。

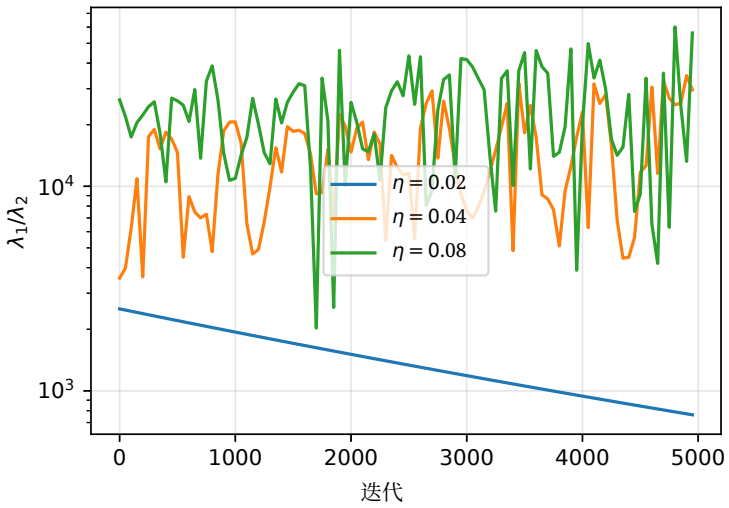
C Rod Flow 的计算实现

在规模符合实际的神经网络上实现 Rod Flow 面临两项挑战，两者都涉及大型矩阵。首先，展度 $\Sigma \in \mathbb{R}^{p\times p}$ 有 p^2 个条目，其中 p 是参数的数量。对于典型的神经网络——其参数通常有数十万甚至数百万个——显式存储 Σ 需要数 TB 内存。其次， $\bar{w}$ 动力学涉及海森矩阵 $\nabla^2 L \in \mathbb{R}^{p\times p}$，这也带来了类似的存储与计算挑战。

本附录介绍我们如何应对这两项挑战：为 Σ构造一种低秩表示，以利用其经验观察到的谱结构；以及采用自动微分技术计算海森向量积，而无需显式构造海森矩阵。

C.1 Σ的低秩表示

我们可以通过考察其常微分方程的结构来理解为何 Σ 应当被预期为低秩。 Σ 弛豫至平衡的时间尺度远快于 $\bar{w}$所处的时间尺度



η	Mean	Min
0.02	1442.6	763.6
0.04	14711.3	3555.0
0.08	24205.7	2027.2

图7：**的特征值比** Σ。不同学习率下 Σ的最大特征值与次大特征值之比，证实 Σ 仍近似为秩-1。**左图：**训练过程中的演变。**右图：**汇总统计。

变化显著 y。这意味着我们可以在任意给定时刻用其稳态值来近似 Σ ：

$$\Sigma \approx \frac{\eta^2}{8}\left(\nabla L_+ \otimes \nabla L_+ + \nabla L_- \otimes \nabla L_-\right) \tag{70}$$

稳态的秩至多为 2。然而，实际上梯度方向在单个时间步内变化不大，因此 Σ 实际上是秩-1（这与其离散时间对应项$\delta_t \otimes \delta_t$一致）。

由于 Σ 是低秩的，我们无需存储整个矩阵，只需存储其前几个特征向量和特征值。具体而言，我们以因子分解形式维护 Σ:

$$\Sigma = V \Lambda V^\top, \tag{71}$$

其中， $V \in \mathbb{R}^{p\times r}$ 是一个矩阵，其正交归一列表示 Σ的前 r 个特征向量；$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^{r\times r}$ 是由相应特征值构成的对角矩阵。这将存储量从 $O(p^2)$ 降至 $O(pr)$。我们在所有实验中都使用 $r = 3$ 。

半差 δ 定义为 Σ 的主特征向量乘以相应特征值的平方根。由于 Λ 是对角矩阵，最大特征值 λ_1 就是最大的对角元素，而相应的特征向量则是 V 基中的向量。因此:

$$\delta = \sqrt{\lambda_1} \cdot v_1, \tag{72}$$

其中 v_1 是 V的第一列。

仅存储 Σ是不够的。我们还必须在每个时间步更新它。这带来了一个挑战，因为更新不一定局限于与 Σ 本身相同的低维子空间。我们的更新分为五个步骤。

步骤 1：衰减。应用常微分方程中 -2Σ 项所产生的指数衰减 ：

$$\Lambda \leftarrow (1 - 2\,dt) \cdot \Lambda. \tag{73}$$

由于 Λ 是对角的，这只是逐元素乘法。

步骤 2：投影梯度。 将每个梯度分解为平行于当前特征子空间的分量和垂直于当前特征子空间的分量：

$$\nabla L_+^{\parallel} = V^\top \nabla L_+ , \qquad \nabla L_+^{\perp} = \nabla L_+ - V \nabla L_+^{\parallel} , \tag{74}$$

$$\nabla L_-^{\parallel} = V^\top \nabla L_- , \qquad \nabla L_-^{\perp} = \nabla L_- - V \nabla L_-^{\parallel} . \tag{75}$$

平行分量是 r 维向量。垂直分量是 p 维的，但与当前基底正交。

步骤 3：扩充基底。如果垂直分量的范数显著（高于阈值 ϵ），则将它们作为新的基方向加入：

$$V_{\text{aug}} = \left[V \quad \frac{\nabla L_+^{\perp}}{\|\nabla L_+^{\perp}\|} \quad \frac{\nabla L_-^{\perp}}{\|\nabla L_-^{\perp}\|} \right] \in \mathbb{R}^{p \times (r+k)} , \tag{76}$$

其中 $k \in \{0, 1, 2\}$ 取决于超出阈值的垂直分量有多少。我们通过格拉姆-施密特方法将这些新方向彼此正交化。

步骤 4：在扩充后的基底中更新。通过为新方向补零，将 Λ 扩展到扩充后的基底，然后加上外积项：

$$\Lambda_{\text{aug}} \leftarrow \Lambda_{\text{aug}} + \frac{\eta^2}{4} dt \left[(V_{\text{aug}}^\top \nabla L_+)(V_{\text{aug}}^\top \nabla L_+)^\top + (V_{\text{aug}}^\top \nabla L_-)(V_{\text{aug}}^\top \nabla L_-)^\top \right] . \tag{77}$$

这会得到一个不再是对角矩阵的对称矩阵 $\Lambda_{\text{aug}} \in \mathbb{R}^{(r+k)\times(r+k)}$ 。

步骤 5：截断。计算特征分解 $\Lambda_{\text{aug}} = U \tilde{\Lambda} U^\top$ U，其中各列为特征向量，Λ 为按特征值降序排列的对角矩阵。仅保留前 r 个特征对，并通过旋转到新的特征向量坐标系来更新基底：

$$V \leftarrow V_{\text{aug}} U_{:,:r} , \tag{78}$$

wh其中 $U_{:,:r}$ 表示 U的前 r 列。新的特征值矩阵为

$$\Lambda \leftarrow \text{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r) , \tag{79}$$

按构造，该矩阵是对角矩阵。我们通过 QR 分解对 V 重新正交化，以纠正数值漂移。

每次更新都需要 $O(pr)$来进行梯度投影，需要 $O(r^3)$来进行特征分解，还需要$O(pr)$来更新 V。由于 $r \ll p$，每次更新的主要开销为 $O(pr)$，与单次梯度计算相当；这与演化完整 $p \times p$ 矩阵所需的 $O(p^3)$开销形成鲜明对比。

C.2 通过自动微分计算海森向量积

在 Rod Flow 中，$\bar{w}$ 动力学包含一项形式为 $\nabla^2 L(w) \cdot v$的项，其中 $\nabla^2 L \in \mathbb{R}^{p\times p}$ 是损失的海森矩阵，$v \in \mathbb{R}^p$ 是一个向量（具体而言，$v = \nabla L(\bar{w} + \delta) + \nabla L(\bar{w} - \delta)$）。

计算并存储完整的海森矩阵的成本高得难以承受：构造它需要 $O(p^2)$次运算，而存储它需要 $O(p^2)$个条目（对于规模合理的网络而言，这相当于数 TB 的存储空间）。幸运的是，我们从不需要完整的海森矩阵，只需要它作用于特定向量时的结果。这些结果称为海森向量积（HVP）。海森向量积 $\nabla^2 L(w) \cdot v$是梯度沿 v方向的方向导数：

$$\nabla^2 L(w) \cdot v = \lim_{\epsilon \to 0} \frac{\nabla L(w + \epsilon v) - \nabla L(w)}{\epsilon} = \frac{d}{d\epsilon}\bigg|_{\epsilon=0} \nabla L(w + \epsilon v) . \tag{80}$$

为了计算导数，现代深度学习框架（如 PyTorch 和 JAX）实现了自动微分（autodiff），这是一种用于计算由计算机程序定义的函数的导数的技术。自动微分通过将计算分解为基本运算，并系统地应用链式法则来实现。有两种模式：

前向模式 通过在计算过程中同步传播切向量来计算方向导数 $\nabla f(x) \cdot v$。给定输入扰动 v，前向模式会追踪该扰动如何影响每个中间值，最终确定其对输出的影响。当前向模式的输出数量超过输入数量时效率较高，因为它每次传播只计算相对于一个输入方向的导数。

反向模式 通过先执行一次前向传播 来计算梯度 $\nabla f(x)$，该前向传播在求值函数的同时记录计算图，然后再执行一次反向传播，将伴随量（灵敏度）从输出反向传播回输入。这就是用于训练神经网络的熟悉的"反向传播"。当输入数量超过输出数量时，反向模式效率较高——这正是神经网络训练的情形：我们有数百万个参数，但损失是标量。

通过结合自动微分的两种模式，可以高效计算海森向量积。海森矩阵 $\nabla^2 L$ 是梯度 ∇L的雅可比矩阵。由于 $\nabla L\colon \mathbb{R}^p \to \mathbb{R}^p$ 将多个输入映射到多个输出，无论采用哪种模式，计算其完整雅可比矩阵的代价都很高。然而，我们只需要海森矩阵对向量 v的作用，这正是梯度的方向导数——也正是前向模式能够高效计算的内容。

这表明可以采用前向套反向 方法：对通过反向模式计算梯度的过程应用前向模式微分。具体而言：

1. 将梯度定义为函数 $g(w) = \nabla L(w)$，通过反向模式计算。
2. 沿方向 v 对 g 应用前向模式微分，得到 $\frac{d}{d\epsilon}\big|_{\epsilon=0} g(w + \epsilon v) = \nabla^2 L(w) \cdot v$。

在 PyTorch 中，这通过 torch.func 实现：

```
def hvp(loss_fn, w,v): grad_fn =torch.func.jacrev(loss_fn) # 反向模式
return torch.func.jvp(grad_fn, (w,), (v,))[1]  # 前向模式
```

这里，jacrev 构造了一个通过反向模式计算梯度的函数，而 jvp 则使用前向模式沿方向 v 对该函数求导。此方法需要 $O(p)$ 内存，计算成本约为单次梯度计算的两倍，同时避免显式构造 $p \times p$ 海森矩阵。

操作	Time	内存
梯度 $\nabla L(w)$	$O(p)$	$O(p)$
低秩 Σ 更新	$O(pr + r^3)$	$O(pr)$
海森向量积 $\nabla^2 L(w) \cdot v$	$O(p)$	$O(p)$

表 2：Rod Flow 操作的计算成本。我们实现中的所有操作都随 p 线性增长（将 r 视为常数），避免了显式矩阵表示带来的二次成本。

总之，每个 Rod Flow 时间步需要：

- 两次梯度评估：$\nabla L(\bar{w} + \delta)$ 和 $\nabla L(\bar{w} - \delta)$
- 用于 $\bar{w}$ 更新的两个海森向量积

- 一次低秩 Σ 更新

每个子步的总成本为 $O(p)$，这使得 Rod Flow 对于拥有数百万参数的网络切实可行。这与标准梯度下降的成本相当，仅有常数因子的额外开销（由于额外的梯度与 HVP 计算，约为 4–6×）。

D 中心流

本附录对中心流 [Cohen 等人， 2025] 提供了自成体系的介绍。我们阐述中心流的核心思想、数学表述与计算实现，然后说明它与 Rod Flow 的关系，并讨论这两种方法之间的权衡。

D.1 概述与动机

中心流背后的核心思想是：尽管梯度下降精确的振荡动力学难以分析，但其时间平均轨迹要容易处理得多。中心流为这条平滑轨迹推导出一个常微分方程。

中心流跟踪两个量：

- $\bar{w}$：振荡的中心（时间平均迭代点）
- Σ：表示振荡围绕中心分布范围的协方差矩阵 r

原论文有意对“时间平均”的精确定义留有一定灵活性。

Damian 等人 [2023], 提出了一个关键见解：由于损失中的非线性（对损失进行三阶泰勒展开即可看出——这比传统优化中常见的二次展开高一阶），沿最大海森特征值对应的特征向量发生的振荡会触发最大海森特征值降低。这解释了为什么梯度下降会在稳定性边缘自行稳定，而不是发散。

为了解释这一点，假设梯度下降围绕中心 $\bar{w}$，沿海森矩阵 u的最大特征值对应的特征向量发生振荡，那么当前迭代点就是 $w = \bar{w} + xu$，其中位移为 x。对梯度进行泰勒展开：

$$\nabla L(\bar{w} + xu) = \nabla L(\bar{w}) + \nabla^2 L(\bar{w})(xu) + \frac{1}{2}\nabla^3 L(\bar{w})[xu, xu] + O(x^3) \tag{81}$$

$$= \nabla L(\bar{w}) + xS(\bar{w})u + \tfrac{1}{2}x^2 \nabla S(\bar{w}) + O(x^3) \tag{82}$$

其中， S（$\bar{w}$）是 $\bar{w}$处的尖锐度。第二项（xSu）会导致经典理论预测的来回振荡。第三项（$\frac{1}{2}x^2\nabla S$）是关键：梯度步会隐式地沿 $-\nabla S$ 方向迈出一步，从而降低尖锐度。这种尖锐度降低的幅度与 x^2成正比，即振荡的方差。

其结果是自稳定。当尖锐度超过 $2/\eta$时，振荡会增大。增大的振荡随后会降低尖锐度。由于尖锐度降低，振荡开始减小。振荡减小使迭代点能够接近最小值，从而导致尖锐度再次上升。结果是，尖锐度在 $2/\eta$附近徘徊。

D.2 中心流方程

D.2.1 中心的动力学

中心 $\bar{w}$ 的演化遵循:

$$\frac{d\bar{w}}{dt} = -\eta \nabla L(\bar{w}) - \frac{\eta}{2} \operatorname{Tr}(\Sigma \nabla^3 L(\bar{w})), \tag{83}$$

第一项是标准梯度流。第二项体现了振荡降低尖锐度的效应。

当振荡集中在最大特征值对应的特征向量方向上时，可简化为:

$$\frac{d\bar{w}}{dt} \approx -\eta \nabla L(\bar{w}) - \frac{\eta}{2} \operatorname{Tr}(\Sigma) \nabla S(\bar{w}). \tag{84}$$

中心沿损失梯度以及尖锐度梯度向下移动，后者的权重为 Tr(Σ)——总振荡方差。

D.2.2 协方差矩阵

如何确定协方差矩阵 Σ? 在中心流模型中，有两个性质约束着振荡：（1）振荡只会在稳定性边缘发生；（2）振荡仅限于海森矩阵的尖锐方向。两者共同意味着， Σ 的秩至多为 k，其中 k 是尖锐特征值的数量。

中心流将注意力限制在临界子空间内——该子空间由特征值接近失稳阈值 $2/\eta$的海森特征向量张成。令 $U \in \mathbb{R}^{p \times k}$ 表示一个矩阵，其列为 $\nabla^2 L(\bar{w})$的前 k 个特征向量，对应的特征值为 λ_1、.....、 λ_k。

协方差在该子空间内参数化为 $\Sigma = U \Omega U^\top$，其中 $\Omega \in \mathbb{R}^{k \times k}$ 为半正定矩阵。这里，Ω 表示以临界特征向量为基所表达的振荡协方差。以下条件均以 Ω 而非 Σ的形式给出，强调中心流作用于这一降维 $k \times k$ 表示，而非完整的 $p \times p$ 协方差。这将问题从跟踪 $O(p^2)$ 个参数简化为跟踪 $O(k^2)$ 个参数，其中通常有 $k \ll p$。

D.2.3 半定互补问题

中心流通过求解以下半定互补问题 (SDCP)来确定每个时刻的协方差 Ω:

$$\boxed{\Omega \succeq 0, \qquad \Lambda + \mathcal{S}[\Omega] \succeq 0, \qquad \langle \Omega, \Lambda + \mathcal{S}[\Omega] \rangle = 0} \tag{85}$$

Λ 是稳定裕度。它衡量临界子空间中的每个特征值距离失稳阈值 $2/\eta$有多远。 Λ的特征值越大，相应的特征方向就离稳定性边缘越远。裕度定义如下:

$$\Lambda = U^\top \left(\frac{2}{\eta} I - \nabla^2 L \right) U \in \mathbb{R}^{k \times k}. \tag{86}$$

如果 $\nabla^2 L$ 的特征向量构成 U 的各列，对应的特征值为 $\lambda_1, \ldots, \lambda_k$，那么:

$$\Lambda = \operatorname{diag}\left(\frac{2}{\eta} - \lambda_1, \ldots, \frac{2}{\eta} - \lambda_k \right). \tag{87}$$

当且仅当海森矩阵的每个跟踪特征值都满足 $\lambda_i \le 2/\eta$ 时，矩阵 Λ 才是半正定的。因此，$\Lambda \succeq 0$ 精确地表达了所有特征值都保持在稳定性阈值以下（含阈值）的要求。

$\mathcal{S}$ 是尖锐度降低算子。线性算子 $\mathcal{S}: \mathbb{R}^{k\times k} \to \mathbb{R}^{k\times k}$ 刻画了振荡如何降低尖锐度（从而增大稳定裕度）。

算子 $\mathcal{S}$ 与三阶导数 $\nabla^3 L$密切相关。二者都以振荡信息为输入，并以尖锐度变化为输出。差异在于它们的定义域：$\nabla^3 L$ 是作用于整个参数空间中振荡的张量，而 $\mathcal{S}$ 是该作用在临界子空间上的限制。

只要存在不等式约束，锥互补问题就会自然出现。最简单的例子是球在地板上弹跳。设 z 表示球的高度，N 表示地板施加的法向力，且两个变量均以向上为正方向。每个变量各自被约束为非负：球必须位于地面之上（$z \ge 0$），法向力只能推而不能拉（$N \ge 0$）。此外，这些量满足互补条件：$zN = 0$。要么球在地面上（$z = 0$）且地板可以施加法向力（$N \ge 0$），要么球在空中（$z > 0$）且法向力必须为零（$N = 0$）。球在空中时存在非零法向力是不可能的。三个条件 $z \ge 0$、$N \ge 0$和 $zN = 0$ 共同构成一个SDCP。

对于中心流的 SDCP：我们不再要求 $z \ge 0$，而是要求 $\Lambda + \mathcal{S}[\Omega] \succeq 0$（计入振荡导致的尖锐度降低后，裕度必须为半正定）。我们不再要求$N \ge 0$，而是要求 $\Omega \succeq 0$（振荡协方差必须为半正定）。此外，我们不再使用标量积 $zN = 0$，而是要求矩阵内积 $\langle \Omega\ \Lambda + \mathcal{S}[\Omega]\rangle = 0$。互补条件确保振荡和正裕度不能在同一方向上共存——这就像弹跳球的高度和法向力不能同时为正一样。振荡只会出现在海森矩阵的尖锐方向上。

中心流的 SDCP 有唯一解吗？有。当 $\mathcal{S}$ 为半正定时（实践中确实如此），SDCP 等价于一个凸二次规划。为说明这一点，考虑约束 $\Lambda + \mathcal{S}[\Omega] \succeq 0$ 的拉格朗日松弛。通过惩罚项引入这一约束，并构造如下规划问题：

$$\min_{\Omega \succeq 0} \frac{1}{2}\langle \Omega, \mathcal{S}[\Omega]\rangle + \langle \Lambda, \Omega\rangle. \tag{88}$$

由于公式（88）是在凸域上的凸二次规划，因此当 $\mathcal{S} \succ 0$ 时，它具有唯一解。

D.3 投影梯度流解释

中心流可以优雅地解释为投影梯度流。考虑在满足关于尖锐度的约束的情况下最小化损失：

$$\min_{\bar{w}} L(\bar{w}) \quad \text{subject to} \quad S(\bar{w}) \le 2/\eta, \tag{89}$$

可行域 $\mathcal{F} = \{\bar{w}: \ S(\bar{w}) \le 2/\eta\}$ 以稳定性边缘为其边界。

标准梯度流 $d\bar{w}/dt = -\eta \nabla L(\bar{w})$ 在沿梯度方向移动会使尖锐度超过 $2/\eta$ 时，可能会离开可行域。投影梯度流则会在保持可行的同时沿梯度方向移动：

$$\frac{d\bar{w}}{dt} = -\eta \nabla L(\bar{w}) - \mu \nabla S(\bar{w}), \tag{90}$$

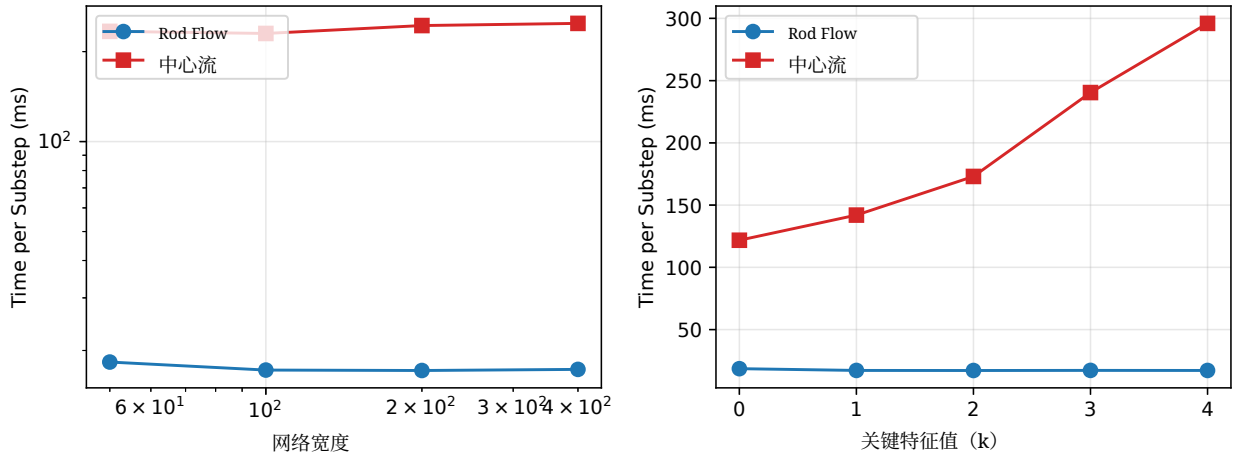


图 8：**Rod Flow 与中心流每次迭代的耗时。左：**计算时间随多层感知机宽度的变化。**右：**计算时间随关键特征值数量的变化 k。Rod Flow 在这两种情形下都实现了显著的加速比。

其中 $\mu \geq 0$ 是拉格朗日乘子。当 $S(\bar{w}) < 2/\eta$ 时，我们有 $\mu = 0$，动力学便简化为标准梯度流。当 $S(\bar{w}) = 2/\eta$ 时，乘子 μ 的取值会使轨迹在保持位于边界上的同时沿梯度方向移动。

与公式（83）比较，我们识别出 $\mu = \frac{\eta}{2}\mathrm{Tr}(\Sigma)$。协方差 Σ 起到拉格朗日乘子的作用：SDCP 精确确定维持可行性所需的投影量。

D.4 计算权衡

	Rod Flow	**中心流**
Σ 确定	ODE 积分	SDCP 求解
每步成本	$O(pr + r^3)$	$O(pk + k^6)$
对 k 的依赖关系	None	$O(k^6)$
特征值计算	不需要	需要（LOBPCG）

表 3：计算比较。其中，p 是参数数量，r 是 Rod Flow 的近似秩（通常为 3），k 是关键特征值的数量。

k	Rod Flow（毫秒）	中心流（毫秒）	加速比
0	18.6	121.8	6.5×
1	17.2	142.0	8.3×
2	17.1	173.0	10.1×
3	17.2	240.4	14.0×
4	17.2	296.0	17.2×

表 4：按关键特征值数量进行的计时比较 k。

E 附加实验细节

本附录补充介绍了使用三层多层感知机、三层卷积神经网络和两层Transformer进行的实验细节。每项实验均遵循两阶段协议。

第一阶段（预热）： 运行梯度下降，直到优化器达到稳定态稳定性边缘阶段。

第二阶段（流比较）： 从预热结束时的梯度下降状态初始化所有流（梯度流、Rod Flow 和中心流）。随后，在剩余迭代次数内，让它们与梯度下降锁步运行。这使我们能够直接比较每种流跟踪梯度下降迭代点的效果。我们跟踪了以下指标：

- 中心（$\bar{w}$）和边（$\bar{w} \pm \delta$）处的损失；
- 中心和边处的尖锐度；
- 振荡幅度（$\|\delta\|$）；
- 中心与梯度下降的偏差（$\|\bar{w}_{\text{梯度下降}} - \bar{w}_{\text{流}}\|$）；
- Delta 对齐（$|\cos(\delta_{\text{梯度下降}}, \delta_{\text{流}})|$）。

由于权重向量规模庞大，我们必须谨慎进行内存管理。所有实验均采用“锁步”执行策略，让梯度下降、梯度流、Rod Flow 和中心流同时运行。这样可以避免存储权重历史记录（长时间运行时达 $\sim$10 GB），代价是无法从任意位置重新启动这些流。我们定期执行显式垃圾回收（gc.collect()）和 CUDA缓存清理（MLP 每 20 次迭代一次，卷积神经网络和 Transformer 每 500 次迭代一次），以防止内存累积。

结果见图10（多层感知机）、图11（卷积神经网络）和图 12（Transformer）。

CIFAR-10 上的多层感知机 我们将多层感知机的隐藏维度从先前工作中使用的标准值 200 降至 50。我们观察到，对于较小的网络，Rod流跟踪梯度下降的效果更好，这可能是因为振荡动力学的维度较低，更容易由低秩 Σ 捕捉

CIFAR-10 上的卷积神经网络滤波器数量从标准的 32 个减少到 8 个，主要是为了提高计算效率。通过卷积层计算海森向量积，需要对卷积运算进行两次反向传播（一次用于计算梯度，一次用于计算海森向量积）；主导项的计算成本随滤波器数量呈二次增长。使用 8 个滤波器时，实验能在合理时间内完成，同时仍呈现出明显的稳定性边缘行为；较少的参数数量（5,938 个，而 32 个滤波器时为 $\sim$80,000 个）也提供了更合适的设置。计算特征值时，我们使用 SciPy 的稀疏 Lanczos求解器，而不是 LOBPCG，因为我们发现，对于卷积神经网络海森结构，LOBPCG 偶尔无法可靠收敛。

（a）多层感知机配置

参数	值
结构	展平 $\rightarrow$ 线性层(3072, 50) $\rightarrow$ SiLU $\rightarrow$ 线性层(50, 50) $\rightarrow$ SiLU $\rightarrow$ 线性层(50, 10)
参数总量	156,710
数据集	CIFAR-10（前 1,000 个样本），均方误差损失
学习率 η	0.025（稳定性阈值 $2/\eta = 80$）
迭代次数	总计 26,000 次，其中预热 22,000 次
特征值计算	热启动 LOBPCG，取前 6 个特征值，每 10 次迭代计算一次

(b) CNN 配置

参数	值
结构	二维卷积(3, 8, kernel=3, padding=1) $\rightarrow$ SiLU $\rightarrow$二维平均池化(2) 二维卷积(8, 8, kernel=3, padding=1) $\rightarrow$ SiLU $\rightarrow$二维平均池化(2) 展平 $\rightarrow$ 线性层($8 \times 8 \times 8$, 10)
参数总量	5,938
数据集	CIFAR-10（前 1,000 个样本），均方误差损失
学习率 η	0.05（稳定性阈值 $2/\eta = 40$）
迭代次数	总计 24,000 次，其中预热 20,000 次
特征值计算	SciPy Lanczos，前 3 个，每 500 次迭代计算一次（共 18,000 次）

(c) Transformer 配置

参数	值
嵌入	词元：嵌入层(30522, 16)，位置：嵌入层(32, 16)
Transformer 层	2 层：2 头注意力机制（注意力头维度=8）、多层感知机（16$\rightarrow$16$\rightarrow$16）、层归一化
输出	均值池化 $\rightarrow$线性层(16, 1)
参数总量	492,273
数据集	SST-2（前 1,000 个样本），32 个词元，均方误差损失
学习率 η	0.01（稳定性阈值 $2/\eta = 200$）
迭代次数	总计 31,000 次，预热 27,000 次
特征值计算	SciPy Lanczos，前 3 个，从 26,000 次迭代开始，每 500 次迭代计算一次

图 9：神经网络实验的实验配置。

SST-2 上的 Transformer隐藏维度从常见的 64 减少到 16，序列长度则限制为 32 个词元。这两项调整主要是为了提高计算效率：Transformer 参数量随隐藏维度 d 呈 $O(d^2)$关系，因此从 64 减少到 16 可使 HVP 计算加速约$16\times$ 倍。自注意力的复杂度与序列长度 n呈 $O(n^2)$关系，会影响前向传播和 HVP 计算。我们发现，32 个词元足以处理 SST-2 中的短句。与多层感知机和卷积神经网络实验相比，较低的学习率对于数值稳定性是必要的，因为参数数量较大，且注意力机制十分敏感。通过注意力层计算 HVP 的成本尤其高昂，因为在前向传播和反向传播期间都必须重新计算 softmax 注意力权重，而序列长度的二次依赖性同时体现在注意力计算及其导数中。尽管成本高昂，Transformer 实验仍表明，Rod Flow 不仅适用于简单架构，也能推广到现代深度学习基础模块。

表 5：神经网络实验配置汇总。

	MLP	**CNN**	**Transformer**
参数	156,710	5,938	492,273
数据集	CIFAR-10	CIFAR-10	SST-2
样本	1,000	1,000	1,000
学习率 η	0.025	0.05	0.01
$2/\eta$	80	40	200
总迭代次数	26,000	24,000	31,000
预热迭代次数	22,000	20,000	27,000
特征值求解器	LOBPCG	SciPy	SciPy

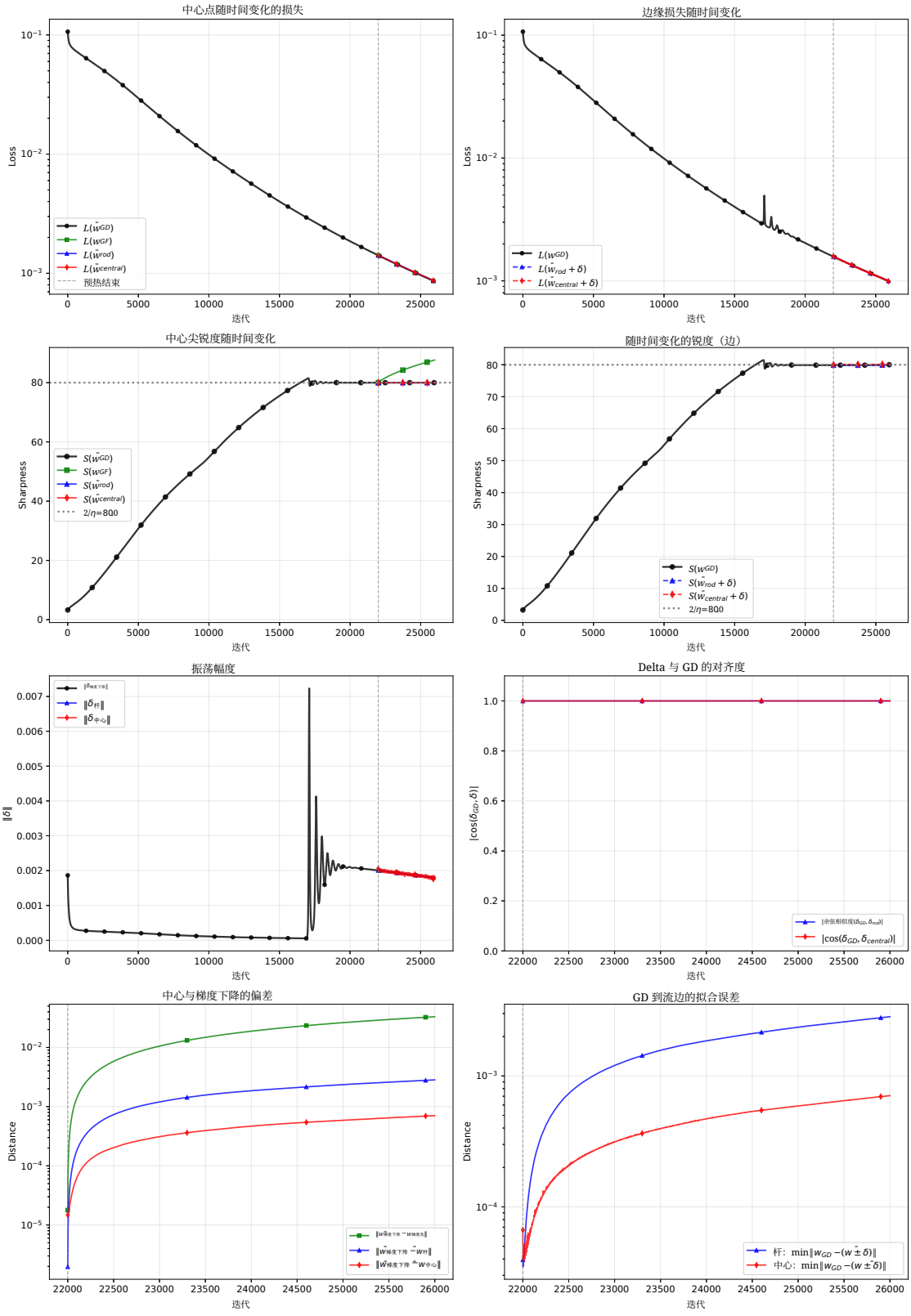


图10：流比较（三层多层感知机）。

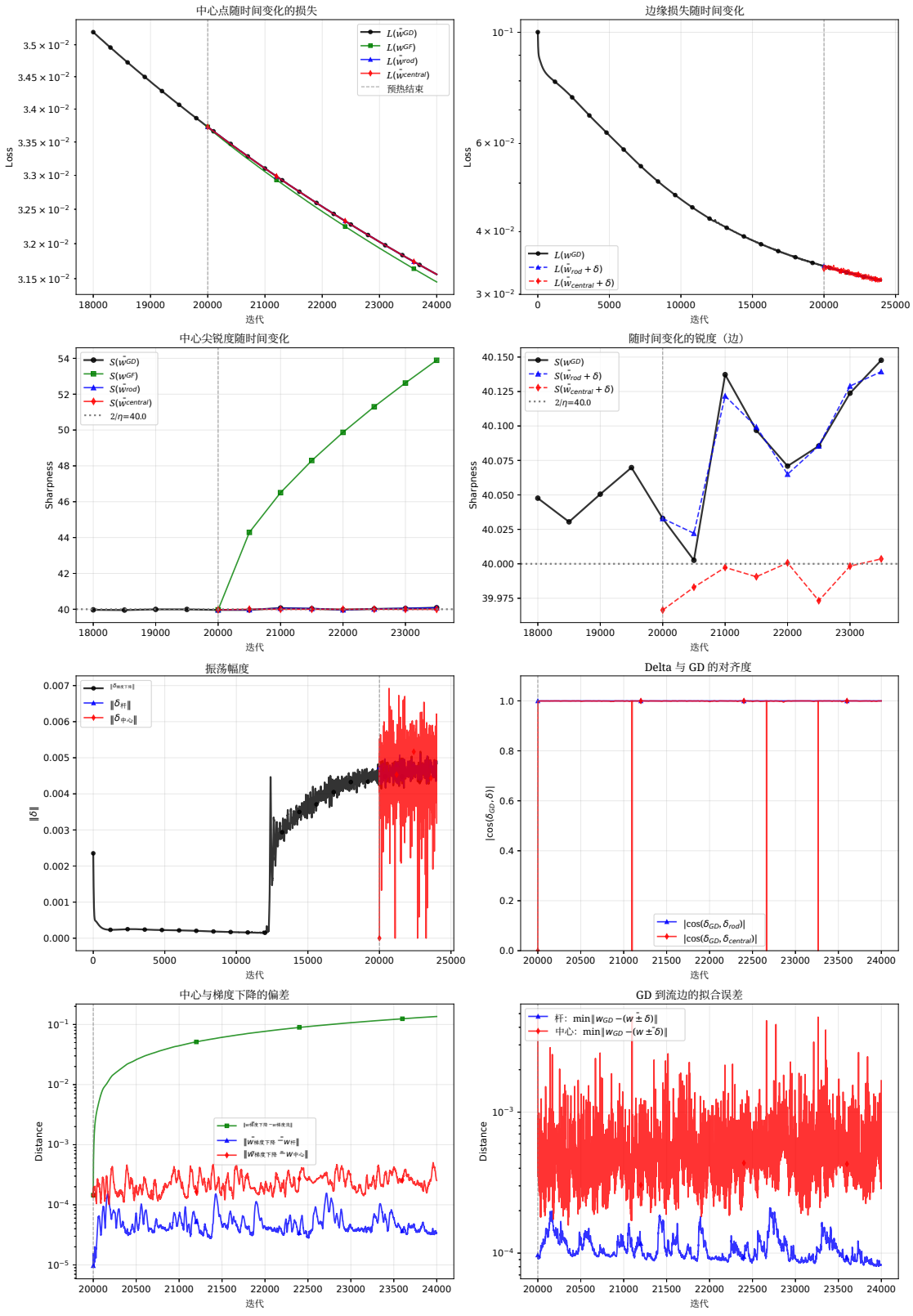


图11： 三层卷积神经网络上的流比较。

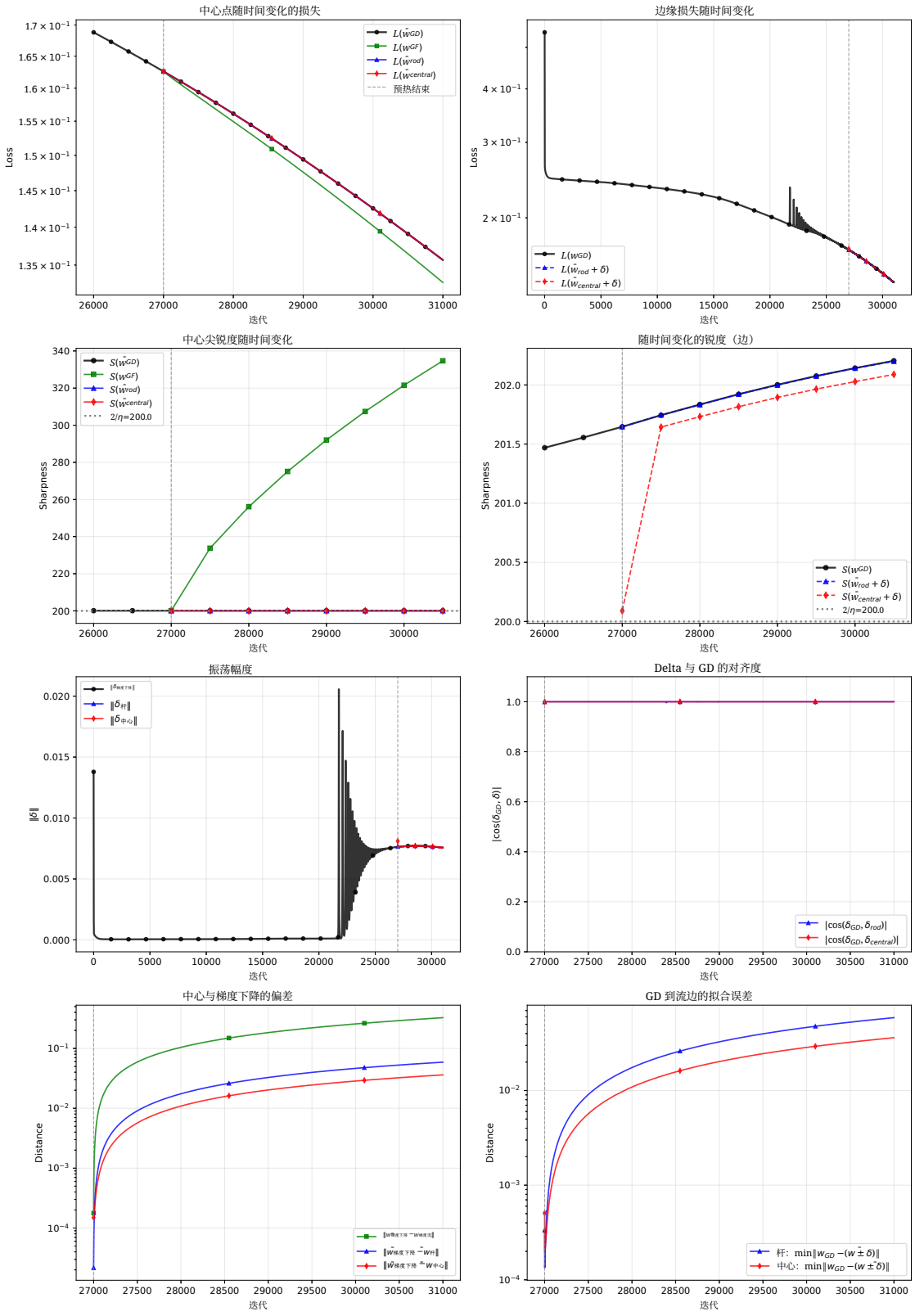


图 12：**Transformer 上的流比较。**