arXiv:2602.05065v1 [cs.LG] 4 Feb 2026

随机梯度下降追求平坦性还是尖锐度？一个可精确求解的模型

Yizhou Xu[1], Pierfrancesco Beneventano[2], Isaac Chuang[2], Liu Ziyin[2,3]

[1]洛桑联邦理工学院[2]麻省理工学院[3]*NTT Research*

2026年2月6日

摘要

大量理论研究和实证工作都假设，神经网络在训练期间的损失景观平坦性与其性能之间存在关联。然而，关于随机梯度下降在训练期间何时偏好更平坦或更尖锐的解，已有观点截然相反的证据。本文通过识别并精确求解一个可解析求解模型，部分地从因果层面阐明了随机梯度下降追求平坦性的行为；该模型在训练期间同时表现出平坦化和锐化行为。在此模型中，随机梯度下降训练并没有先验的平坦性偏好，而只是偏好最小化梯度波动。由此可知，至少在此模型中，收敛时的尖锐度完全由数据分布决定；当且仅当标签噪声在所有输出维度上均为各向同性时，模型才偏好平坦极小值。当标签噪声具有各向异性时，模型则转而偏好尖锐度，并且可能收敛到任意尖锐的解，具体取决于标签噪声谱的不平衡程度。我们在不同模型架构（如多层感知机、循环神经网络和 Transformer）的受控设置中，复现了这一关键发现。

1 引言

损失景观的平坦性和尖锐度，已成为理解神经网络优化与泛化的重要研究对象 [HS97, KMN+16]，更不用说基于这两种机制的各种算法了 [FKMN20]。例如，平坦性是解释神经网络为何能够泛化的最常见假设之一：即使许多参数设置都能完美拟合训练数据，人们通常认为，位于损失景观较平坦区域的解会更加稳定，也能更好地泛化。这种（颇值得商榷的）直觉推动了大量研究，将随机梯度下降的步长和随机性与其对更平坦极小值的隐式偏好联系起来，而这种偏好通常通过基于海森矩阵的尖锐度度量来表征 [KMN+16, WM+18, XSS20, WS23]。与此同时，现代训练中一种看似矛盾的现象日益凸显：尖锐度往往会随着优化过程而增加，最终达到稳定性边缘（EoS）；此时曲率很大，训练也仅处于勉强稳定的状态 [CKL+21, AB24]。“随机梯度下降追求平坦性”和“随机梯度下降追求尖锐度”这两种倾向并存，正是我们所说的尖锐度“悖论”——当然，这并非字面意义上的悖论，而是反映了我们尚不了解随机梯度下降究竟天生倾向于学习更平坦还是更尖锐的解。要解决这一“悖论”相当困难，尤其是因为存在许多混杂变量：尖锐度的变化轨迹会受到初始化的混杂影响（从更平坦的起点开始会导致锐化，反之亦然），也会受到运行间变异的影响。

在这项工作中，我们在所谓的“最小波动”约束下，给出了一个数学模型中尖锐度的精确解，这使我们能够剔除这些混杂因素，并识别该模型中平坦化与锐化的精确原因。在该模型中，全局极小值构成一个高度退化流形，其中包含尖锐度取值范围很广的解——包括任意尖锐的解——然而随机梯度下降的学习动力学会选择一个尖锐度唯一的解。这使我们能够回答

隐式偏差问题的最简洁形式：当存在许多同样好的极小值时，随机梯度下降会选择哪种曲率，以及哪些因素决定了这一选择？我们的主要贡献是：

1. 在随机梯度下降使波动最小化的情形下，我们识别出一个关于尖锐度/平坦性的可精确求解模型；

2. We精确 ly identify the原因o f fl平坦性-see king beh行为o 随机梯度下降 in th是否 del
：（a）当标签不确定性是各向同性时，训练总是偏好所有解中最平坦的解；（b）否则，标签不确定性谱 Σ_ϵ 决定了收敛锐度，它与 Σ_ϵ的条件数成反比。

我们在非线性模型和现实的数据分布中检验了这些预测，结果表明，这些结论并不局限于我们求解的解析模型。我们的理论起点是近期的一项理论进展：该进展表明，随机梯度下降训练本身倾向于选择梯度波动较小的解 [Rob21、SDBD21、Ben23、ZXC25], ；而且，我们在解释真实神经网络方面取得的实证成功，进一步支持了最小波动理论。在最小波动理论中，关键所在——也是悖论的根源——是波动与尖锐度并非同一概念。只有在特殊条件下二者才一致，因此，最小化波动通常并不意味着最小化尖锐度。我们求解的精确模型是一个以带噪声标签训练的线性教师为对象的深度线性网络。在该模型中，随机梯度下降所选出的尖锐度可以用输入协方差、教师映射、深度以及各输出维度上的标签噪声[1] 协方差显式表示。

该理论的关键结论是从数据几何术语的角度重新审视尖锐度之争：相较于梯度下降，随机梯度下降看起来是在寻求平坦性，还是在使模型锐化，取决于具体情形；起决定作用的因素是标签上有效噪声的各向异性。

2 更尖锐还是更平坦?

尖锐度。深度学习理论界长期以来普遍认为，随机梯度下降（SGD）的隐式偏置会驱使模型趋向平坦极小值。这一思路 [HS97、KMN+16、FKMN20] 认为，平坦区域对参数扰动具有更强的鲁棒性，因此能够带来更好的泛化。大量理论研究从梯度噪声和动力学稳定性的角度，试图解释这种隐式偏置[ZWY+18，WM+18，WWS22，WS23，MM24]。然而，近期对神经网络训练离散动力学的研究表明，随机梯度下降并不总是停留在可能达到的最平坦区域。相反，它往往会收敛到明显“尖锐”的区域，这种现象称为渐进锐化 [XATB18、JKB+19、JSF+20、CKL+21、AB24]。因此，随机梯度下降对尖锐度的偏好仍是一个悖论。

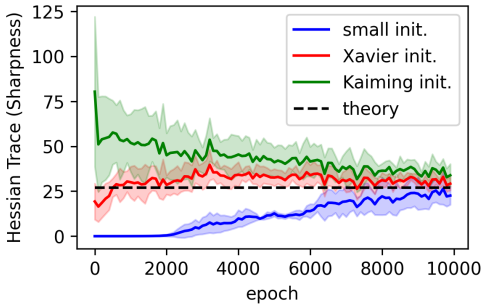

图1：我们复现了[ZWLW24], 中的实验：一个通过随机梯度下降训练的深度矩阵分解问题收敛到了相同的尖锐度，即使初始尖锐度强烈依赖于初始化尺度。理论曲线是定理 1 的预测。

初始化的任意性。然而，由于混杂变量数量众多，很难判断并理解何时出现的是更平坦的极小值，何时出现的是更尖锐的极小值。首要问题。第一个混杂变量是解的任意性。我们常常会测量训练步数增加过程中锐度的演变，并根据训练期间锐度的变化，断言模型偏好更尖锐或更平坦的解。然而，这种观点会受到所用初始化方案任意性的影响。设想一个损失景观，其全局最小值具有唯一且非零的锐度，但初始化的锐度可以任意取值。那么，以能够使模型到达全局最小值的训练轨迹为条件，较平坦的初始化意味着训练平均而言必须使损失景观变得更尖锐，而较尖锐的初始化则意味着训练平均而言必须使其变得更平坦。在这种情况下，

[1]为澄清这一点，在 我们的研究中，我们广义地使用“标签噪声”一词。它可能指动态注入的 d
噪声，也可能是 标签中固有的静态噪声，导致模型无法完美插值数据。

很难判断学习动力学究竟偏向平坦性还是尖锐度，因为这在很大程度上取决于初始化的任意性。第二个混杂变量是结果可能具有随机性。完全有可能的是，每种初始化（或采样过程）都会导向一个不同的局部最小值，其锐度也各不相同，因此难以跨设置进行比较。因此，理想情况下，理想的理论模型应当消除这种任意性。找到一种理论模型，使模型只收敛到唯一的锐度水平，将是进一步理解趋向平坦行为的重要一步。效果示例见图 1 。

内在尖锐度。 因此，可以设想如下情景最有助于理解学习算法寻求平坦性的行为：（a）构造一个具有退化的全局极小值流形的损失景观，其中包含尖锐度水平不同的解；（b）在学习轨迹到达全局最小值的条件下，研究学习动力学如何决定平均尖锐度水平。在这一情景中，所学解的尖锐度显然有两项贡献：（a）学习该问题所需的内在尖锐度：对于任意给定任务和任意给定架构，学习它所需的最低尖锐度水平应当非零；（b）由学习算法偏置的尖锐度：当存在尖锐度取值不同的解流形时，随机梯度下降会从中选择哪一个。概念上，我们可以写成（用 T 表示某种尖锐度度量）：

$$T = T_{\text{intrinsic}} + \Delta_{\text{learning}}. \tag{1}$$

第一项存在于任何能够到达全局最小值的学习算法中，因此任何仅改变第一项的效应都不能被论证为学习算法的“偏好”。第二项完全由学习算法的偏好及其超参数的选择决定。因此，影响第二项的因素可称为“隐式平坦性偏置”。当然，研究什么决定第一项同样重要，我们也将在我们的模型中给出答案。

随机梯度下降特有的锐化。 先前关于渐进锐化和随机稳定性边缘的研究记录了训练过程中曲率可能增大的现象，但其底层机制从根本上说是由损失景观驱动的：梯度下降和梯度流也会出现类似的锐化，而干预性研究将这种效应归因于指向更尖锐区域的损失景观斜率，而非随机梯度下降特有的随机性 [JSF+20、CKL+21、CGK+22、AB24]。我们的机制在性质上有所不同。相比之下，随机梯度下降确实会诱导一种选择规则：它近似最小化熵目标（12），因而会在极小值流形上选择梯度波动泛函的极小值点 $S(\theta)$（引理 2）。梯度流/梯度下降不受标签噪声协方差的影响 Σ_ϵ （它们依赖标签一阶矩，而非二阶矩），随机梯度下降则不然。在我们的可解模型中，小批量诱导项通过最小化梯度波动，在退化的全局极小值流形中进行选择；这种选择显式依赖于 Σ_ϵ的各向异性。因此，即使存在平坦得多的全局极小值，随机梯度下降仍可能沿着极小值流形移动到严格更尖锐、甚至可以任意尖锐的解。这并非“在收敛过程中逐渐锐化”，而是收敛时受噪声条件影响的隐式偏置，另见图 4。

在我们的模型中，我们将发现 Δ_{learning} 是由波动最小化所产生的随机梯度下降特有贡献（在各向同性噪声情形下消失，并随 $\kappa(\Sigma_\epsilon)$在我们的模型中按比例变化；推论1）。与这一解释一致，通过直接干预将随机梯度下降切换为全批量梯度下降，会遏制锐化漂移（图4）。总的来说，这至少对常见的启发式“随机梯度下降追求平坦性”的叙述提出了质疑，甚至可能予以推翻：在我们的模型中，随机梯度下降主要追求最小波动，而平坦性只会在波动与尖锐度一致这一人为构造的情形下涌现。

3 一个可精确求解的模型

在本节中，我们首先阐述问题设定的动机并对其进行形式化定义，然后给出主定理，其证明见附录。希望进一步了解直觉的读者，可以先阅读第 5 节概述的一般理论框架。下一主要章节中，我们将讨论该理论的各种含义。

记号： 我们用 θ 表示所有可训练参数的集合。不带下标时， ∇表示关于 θ的梯度。 $\lambda_{\max}(\Sigma)$表示 Σ的最大特征值。 $\nabla^2 z$ 表示标量函数 z的海森矩阵。在本文中，我们用 $\mathbb{E}$ 表示在训练集 $\mathrm{D}_{\text{train}}$上的平均值，其中训练集不一定是有限集合。 I_d 是 $\mathbb{R}^d$的恒等算子；当上下文明确时，通常省略 d 。其他记号将在相应上下文中引入。

3.1 模型与损失函数

我们研究一个 D 层深度线性网络。具体而言，该模型定义为

$$f_\theta(x) = W_D \cdots W_1 x. \tag{2}$$

其中 $\theta := \{W_k\}_{k=1}^D$ 表示所有权重矩阵的集合， x 是输入数据。该模型在回归任务上使用均方误差损失进行训练，其中逐数据的随机损失为

$$\ell(x, y, \theta) = \frac{1}{2}\|y - W_D \cdots W_1 x\|^2, \tag{3}$$

其中 y 是与 x 相关联的标签。在实践中，我们使用随机梯度下降训练该模型，但随机梯度下降在该模型上的训练动力学难以解析求解，因此我们采用另一种方法，可将其视为回答如下问题：“如果训练以随机梯度下降进行，那么该算法实际在最小化何种目标函数？”相比之下，这一替代问题可以在完全一般的情形下被精确求解。该替代目标函数将在第 3.3 节讨论。

标签由线性教师 V 生成，并带有噪声 ϵ

$$y = Vx + \epsilon \in \mathbb{R}^{d_y}, \tag{4}$$

其中 $\mathbb{E}[x] = 0$ 且 $\mathbb{E}[\epsilon] = 0$。

我们定义协方差 $\Sigma_x := \mathbb{E}_x[xx^T]$ 和 $\Sigma_\epsilon := \mathbb{E}_\epsilon[\epsilon\epsilon^T]$。请注意，所有输入、标签、 V以及噪声都可以具有任意维度，只要它们的维度彼此一致。在该数据生成模型下，经验损失 L的全局极小值满足：

$$W_D \ldots W_1 x = Vx \tag{5}$$

对于训练集中的所有 x 。

作为分析的关键部分，请注意该模型具有许多对称性（即损失不变方向），我们称之为矩阵缩放对称性。对于任意可逆矩阵 A以及任意层索引 i，我们有

$$f(W_i, W_{i-1}) := W_D \cdots W_i W_{i-1} \cdots W_1 x \tag{6}$$

$$= f(W_i A, A^{-1} W_{i-1}). \tag{7}$$

这也意味着深度线性网络存在无穷多个全局极小值。正如我们将在下一节所示，这意味着存在 L 的一个全局最小值，其尖锐度可以任意高。

3.2 尖锐度

在本文的大部分内容中，我们使用以下尖锐度度量：

定义 1。 尖锐度 $T(\theta) := \text{迹}\mathbb{E}[\nabla^2 \ell(x, y, \theta)]$。

T 已在许多先前的研究中用作尖锐度的度量，并且通常被发现能更好地表征学习动力学的稳定性 [LWA21、WS23、ZWLW24], 这是研究损失景观尖锐度的主要原因之一。它还为海森矩阵的最大特征值 $\lambda_{\max}$ 提供了上界和下界：

$$T/d \le \lambda_{\max} \le T. \tag{8}$$

因此， $T(\theta)$ 可作为解的稳定性的可靠替代指标。在 $T \to \infty$,$\lambda_{\max}$ 趋于渐近极限时，也必须发散。

研究随机梯度下降的隐式偏置之所以必要，是因为损失景观中存在通往尖锐度任意高的极小值的路径。以下引理表明，对于称为指数对称性 [ZWLW24],的这类对称性，我们总能找到一个尖锐度为无穷大的全局最小值：

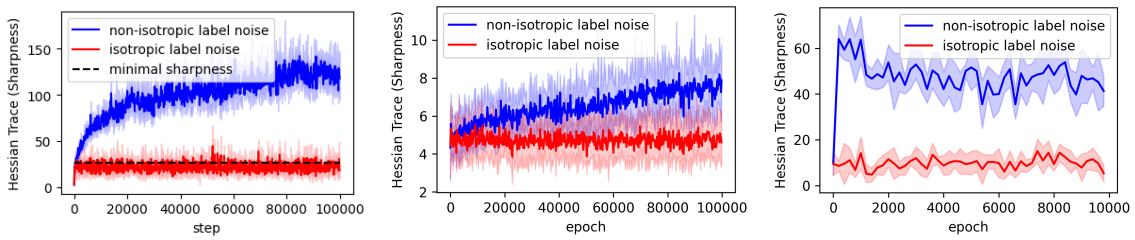

图 2：标签中的非各向同性噪声会导致渐进锐化。**左图**：线性网络，其最小锐度由（15）预测。**中图**：在师生设定下训练的两层 ReLU 网络。**右图**：在 MNIST 上训练的两层 ReLU 网络。每条曲线均为五次试验的平均值，并显示标准误差。实验细节及更多实验见附录B 。

引理 1。 假设 A 是一个对称矩阵，并且对任意 x, y, θ, $\ell(x, y, e^{\lambda A}\theta) = \ell(x, y, \theta)$。此外，假设 $A\mathbb{E}_{x,y}\nabla^2\ell(x, y, \theta) \neq 0$。那么，$\limsup_{|\lambda|\to+\infty} |T(e^{\lambda A}\theta)| = +\infty$。

矩阵缩放对称性也是一种指数对称性，并且满足引理1。这种对称性蕴含尖锐解存在，这一点最初是在 [DPBB17] 中针对简单缩放对称性的情形得到识别并证明的，而引理1 可视为这一结果最一般的形式。

作为理解其含义的一个例子，考虑将其应用于具有 $D = 2$ 的深度线性网络。对于$\ell(x, y, \theta) = \frac{1}{2}||y - W_2W_1x||^2$，我们有

$$T(W_2, W_1) = ||W_2||_F^2 \mathrm{Tr}\Sigma_x + d_y \mathrm{Tr}(W_1^T W_1 \Sigma_x), \tag{9}$$

在这种情况下，我们有 $\ell(x, y, W_2, W_1) = \ell(x, y, \lambda W_2, \lambda^{-1}W_1)$，但是

$$\lim_{|\lambda|\to+\infty} T(\lambda W_2, \lambda^{-1}W_1) = +\infty \tag{10}$$

对于任意 W_1, W_2。

正如下文的主定理 1 将要证明的，在全局极小值处，深度线性网络将（1）具有唯一的最小尖锐度值 $T_{\min}$，它仅取决于问题；并且（2）收敛到唯一的尖锐度值 $T \geq T_{\min}$，这意味着实际学习到的尖锐度水平可以分解为两个部分：

$$T = T_{\min} + \Delta_{\mathrm{SGD}}, \tag{11}$$

其中 $\Delta_{\mathrm{SGD}} \geq 0$, $T_{\min}$ 是拟合数据分布（给定架构）所需的最小尖锐度水平，任何能够达到全局最小值的学习算法都必须满足这一要求； Δ_{SGD} 是由随机梯度下降的隐式偏置导致的尖锐度水平（也就是说，源于我们将在下一节介绍的涨落最小化效应）。

更为标准的量 $\lambda_{\max}$ 同样值得关注，因为它直接关系到近期广受关注的稳定性边缘（EoS）问题。一般而言，尽管我们的理论允许对 $\lambda_{\max}$ 进行直接计算，但其函数形式过于复杂，因此我们不显式给出[2]。尽管如此，我们将在第4 节讨论若干情形，其中 $\lambda_{\max}$ 确实可以写成易于解释的形式。

3.3 最小波动拟设

一系列工作表明，在有限学习率下结合小批量采样进行随机梯度下降训练，等价于求解如下扰动优化问题，我们称之为熵损失[Rob21，SDBD21，Ben23，ZXC25]:

$$F_\eta(\theta) = \underbrace{\mathbb{E}_{x,y}\ell(x, y, \theta)}_{\text{learning, symmetry}} + \underbrace{\frac{\eta}{4}\mathbb{E}_{\mathcal{B}}||\mathbb{E}_{(x,y)\in\mathcal{B}}\nabla\ell(x, y, \theta)||^2}_{\text{due to discretization error, noise, } := S(\theta)}, \tag{12}$$

其中 B 表示小批量， η 为学习率。利用这一修正的熵损失，我们可以通过如下引理刻画随机梯度下降的隐式偏置。

[2]注意，这是 n　　这不是我们理论的问题，而是通常如何计算任意矩阵最大特征值的问题　　.

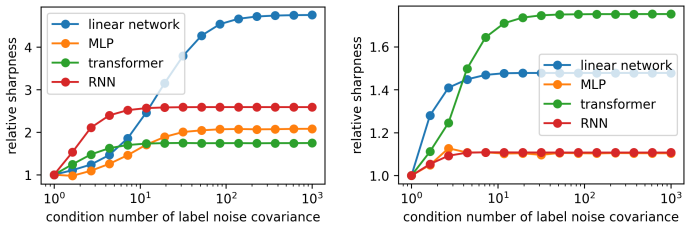

图3： 对于全部四种模型，以及回归和分类任务，我们观察到，随着标签噪声变得越来越非各向同性，学习过程会得到更尖锐的解。我们将相对尖锐度定义为非各向同性噪声下的海森矩阵迹与各向同性噪声下的海森矩阵迹之比。**左**：均方误差损失。**右**：交叉熵损失。

引理 2（最小波动引理，非正式表述）。 在极限 $\eta \to 0^+$ 下， $F_\eta(\theta)$ 的全局极小值为

$$\arg\min_{\theta \in \mathcal{M}} S(\theta), \tag{13}$$

其中 $M = \{\theta \mid \mathbb{E}_{x,y}\ell(x,y,\theta) = \inf_u \mathbb{E}_{x,y}\ell(x,y,u)\}$ 是全局极小值流形。

由于现代神经网络通常是过参数化的，损失景观 $\mathbb{E}_{x,y}\ell(x,y,\theta)$ 包含一个全局极小值流形。该引理表明，随机梯度下降并不会平等对待这些极小值，而是隐式地偏好波动最小的极小值 S。值得注意的是，对于全批量梯度下降（GD），这种偏差会消失，因为在任何全局最优点处，波动项都为零。

证明的一个关键机制是，在矩阵缩放对称性下，不同的对称方向实际上具有不同程度的波动（通常也对应不同程度的尖锐度，如引理 1 所示）——因此，学习算法对最小波动解的偏好，也会促使学习算法朝着某种尖锐度水平移动。

3.4 主要结果

设 $V' := \sqrt{\Sigma_\epsilon} V \sqrt{\Sigma_x}$ 其奇异值分解表示为 $V' = LSR$。我们假设网络宽度 $d \geq \text{rank}(V')$[3]。在随机梯度下降极限 $\eta \to 0^+$ 下，我们得到尖锐度的以下闭式结果：

定理 1。 对于 (12) 的所有全局最小值，海森矩阵的谱均相同（由 Σ_x、 Σ_ϵ和 V决定），且尖锐度为

$$T(\theta) = (D-1)\,(\text{Tr}S)^{(D-2)/D}\,(\text{Tr}\Sigma_\epsilon \text{Tr}\Sigma_x)^{1/D}\text{Tr}[\Sigma_\epsilon^{-1} L S L^T] + d_y\,(\text{Tr}S)^{2(D-1)/D}\,\frac{(\text{Tr}\Sigma_x)^{1/D}}{(\text{Tr}\Sigma_\epsilon)^{(D-1)/D}}. \tag{14}$$

此外， $\mathbb{E}_x\ell(x,\theta)$ 的全局最小值的最小锐度为

$$T_{\min} := \min_\theta T(\theta) = D(\text{Tr}\hat{S})^{2(D-1)/D}(d_y \text{Tr}\Sigma_x)^{1/D}, \tag{15}$$

其中， $\hat{S}$ 是 $V\sqrt{\Sigma_x}$ 的奇异值。

定理 1 揭示了一个引人注目的性质：尽管存在无穷多个尖锐度各异（甚至为无穷大）的全局极小值，随机梯度下降却始终选择一个具有**唯**一尖锐度水平的解。这表明，只要模型收敛到全局最优点，最终的尖锐度就与初始化无关。此外，式（14）的第一项表明， $D-1$ 层对尖锐度的贡献方式相同，而最后一层（式（14）的第二项）的贡献方式则不同。然而，在最小尖锐度解（15）处，所有层的贡献方式都相同。我们将在第 4 节详细分析该定理的含义。

[3] d 定义为 t W_i 对所有 $i \in [D]$，输出维度与输入维度中的较小维度。当 $d < \text{rank}(V')$ 时，i t 学习最大的 d 特征值 ，结果类似。

4 个见解

可精确求解模型的一大优点是，能够精确且因果地理解模型中的每一种效应和现象。因此，本节旨在阐释定理 1 中的关键观察与见解。随后，我们将展示针对多种模型架构开展的实验，以说明该预测在定性层面上的相关性。理论结果的证明也将在附录中给出。

噪声不平衡是尖锐度的来源。

推论 1. 假设 Σ_ϵ 和 $V\sqrt{\Sigma_x}$ 满秩。设 Σ_ϵ 的条件数为 $\kappa(\Sigma_\epsilon) := \lambda_{\max}(\Sigma_\epsilon)/\lambda_{\min}(\Sigma_\epsilon)$。那么，

$$c_1 D\sqrt{\kappa(\Sigma_\epsilon)} \le T(\theta) \le c_2 D\kappa(\Sigma_\epsilon) \tag{16}$$

在式(12)的全局最小值处，其中 c_1, c_2 是取决于 $V\sqrt{\Sigma_x}$ 和 d_x, d_y的常数。

推论 1 表明，尖锐度主要取决于标签噪声的不平衡，而与深度关系不大。作为一个说明性的特例，我们可以选择 $\Sigma_x = I, V = I$ ($d_x = d_y = d$)，得到 $V' = \sqrt{\Sigma_\epsilon} = LSL^T$，因此

$$T(\theta) = (D-1)(\mathrm{Tr}[\Sigma_\epsilon^{1/2}])^{(D-2)/D}(d\mathrm{Tr}[\Sigma_\epsilon])^{1/D}\mathrm{Tr}[\Sigma_\epsilon^{-1/2}] + d^{(D+1)/D}\mathrm{Tr}[\Sigma_\epsilon^{1/2}]^{2(D-1)/D}\mathrm{Tr}[\Sigma_\epsilon]^{-(D-1)/D}, \tag{17}$$

它可以任意大。另一方面，我们有 $\min T(\theta) = Dd^{(D+2)/D}$，它不依赖于 Σ_ϵ。因此，噪声谱的不平衡可能导致任意高的尖锐度。如果我们进一步选择$\Sigma_\epsilon = \mathrm{diag}(1, \sigma^2)$，则有

$$T(\theta) = D(1)2^{1/D}(1+\sigma)^{(D-2)/D}(1+\sigma^2)^{1/D}(1+1/\sigma) + 2^{(D+1)/D}(1+\sigma)^{2(D-1)/D}(1+\sigma^2)^{-(D-1)/D} \tag{18}$$

可以验证，它关于 $\sigma \to \sigma^{-1}$ 是对称的，并且在 $\sigma = 1$ 处有唯一的极小值。近似地，我们有 $T_{\min} \approx D$，并且

$$\frac{T(\theta)}{T_{\min}} \approx \max(\sigma, 1/\sigma) \tag{19}$$

对于较大的 D 以及标签噪声严重不平衡的情况。因此，随机梯度下降的尖锐度与最小锐度之间的差距近似正比于标签噪声的不平衡程度，且与深度无关，正如推论 1 所示。

这或许能为 Transformer 中观察到的锐化现象提供一种理论解释。在词元预测中，“停用词”（例如“the”）的条件熵很低，而实词（名词/动词）的熵很高。词元级噪声的这种巨大不平衡会自然地推动学习趋向更尖锐的解。出于同样的原因，当类别不平衡时，标签噪声也往往是不平衡的，从而使解变得更尖锐。幂律协方差的示例见附录A.10 。

各向同性噪声下最小锐度的恢复。

推论 2. 当且仅当 $\Sigma_\epsilon = I$ 时，全局最小值最平坦。

推论 2 背后的原因是，对于具有 $\Sigma_\epsilon = I$ 的深度线性网络，尖锐度与波动恰好成正比（线性网络的二阶导数(24) 中的二阶导数项消失），因此，最小化波动等价于最小化尖锐度。

最大特征值具有类似的性质

推论3. 假设 VV^T 与 Σ_ϵ 对易，且 V^TV 与 Σ_x 对易。那么，

$$c_1 D\sqrt{\kappa(\Sigma_\epsilon)} \le \lambda_{\max}(\mathbb{E}[H]) \le c_2 D\kappa(\Sigma_\epsilon), \tag{20}$$

其中 H 是 *Hess* 在全局最小值处的海森矩阵(12) 和 c_1，c_2 是取决于 V，Σ_x 和 d_y的常数。

推论 3 表明，在所有矩阵都对齐这一特殊情况下，(12) 的全局最小值处海森矩阵的最大特征值主要由噪声不平衡决定，而非由深度决定。我们预期这一结论通常也成立。作为一个示例，我们可以取 $V = \Sigma_x = I$，$\Sigma_\epsilon = \mathrm{diag}(1,\ \sigma^2)$。于是，我们可以得到

$$\begin{aligned}\lambda_{\max}(\mathbb{E}[H]) = (D-2)2^{1/D}(1+\sigma)^{-2/D}(1+\sigma^2)^{1/D}\max(1,\sigma^{-1})\max(1,\sigma)\\ + 2^{1/D}(1+\sigma)^{(D-2)/D}(1+\sigma^2)^{-(D-1)/D}\max(1,\sigma) + 2^{-(D-1)/D}(1+\sigma)^{(D-2)/D}(1+\sigma^2)^{1/D}\max(1,\sigma^{-1}).\end{aligned} \tag{21}$$

根据定理3。可以验证，它关于 $\sigma \to \sigma^{-1}$ 是对称的，并在 $\sigma = 1$处达到最小值，而该值是最小锐度解的最大特征值。我们还有

$$\lambda_{\max}(\mathbb{E}[H]) \approx D\max(\sigma, \sigma^{-1}) \tag{22}$$

当 D 和 $\sigma \gg 1$ 较大或 $\sigma^{-1} \gg 1$时。因此，最大特征值与海森矩阵的迹具有相似的尺度，并且会随着噪声变得不均衡而增大。

渐进锐化或平坦化。 定理 1表明，随机梯度下降会收敛到一个与初始化无关的尖锐度水平。因此，如果初始化足够小，就会出现渐进锐化（至少在整个轨迹上平均而言如此）；而如果初始化较大，则会出现渐进平坦化。在命题 1中，我们刻画了初始尖锐度，并观察到，对于标准初始化方案，它与标签无关。将命题 1 与定理 1 进行比较，可以发现渐进锐化存在多种机制。注意到初始尖锐度与标签或噪声分布无关，而最终尖锐度取决于标签和噪声，因此可以得到一种简单直接的平坦化与锐化机制：对于固定的初始化，将标签重新缩放 y 得更小会导致平坦化，而将 y 重新缩放得更大则可能导致锐化。这再次展现了核心观点：随机梯度下降本身可能并不偏好尖锐度，真正改变（如果不是决定）尖锐度的是数据分布。

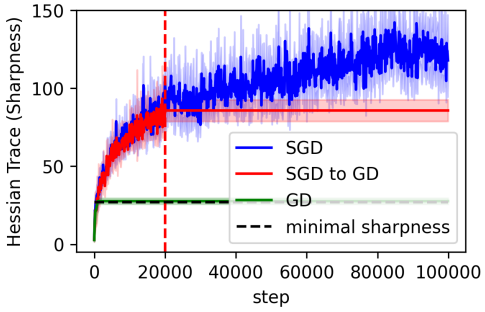

图4： 标签中的非各向同性噪声会导致只有随机梯度下降才具有独特的渐进锐化现象。红色虚线表示我们切换到梯度下降的时刻。

这是一种新的渐进锐化机制。如第2节所述，以往已知的机制涉及梯度流和全批量梯度下降。我们的理论发现的渐进锐化机制则不然。事实上，它与噪声协方差有关，在确定性训练流程中无法观察到，例如 [XATB18、JKB+19、JSF+20、CKL+21、CGK+22、AB24]中的情况。事实上，对于梯度下降，式（12）的第二项在任何局部最小值处始终为零，因此不会导致渐进锐化。图4展示了这一点。此外，我们的渐进锐化机制通常与随机稳定性边界及不稳定性无关；即使学习率趋于零，也能观察到这一机制。

实验。为了对我们的理论进行实证验证，我们在深度线性网络和若干非线性架构（多层感知机、循环神经网络和 Transformer）上，针对回归和分类任务进行了实验。图2 和3所示的结果表明，观察到的现象并非线性模型所独有；具体而言，不平衡标签噪声会导致更尖锐的解，这与推论1的预测一致。此外，图2左侧验证了推论2：在线性网络中，面对各向同性标签噪声时，模型会收敛到最小尖锐度解。

在图 3中，我们观察到，标签噪声的非各向同性越强，尖锐度就越高。高条件数下的平台期是有限时域效应所致：不平衡噪声极强时，需要很长的收敛时间才能达到最尖锐的极小值。为单独考察随机性的作用，图 4 展示了一项直接干预：在线性网络设置中切换到梯度下降后，渐进锐化现象便消失了。这表明该机制是随机梯度下降所特有的，与（12）一致，其中第二项在梯度下降中会消失

在局部极小值处，不会产生任何隐式偏置。关于特征值、条件数和 Transformer 的完整细节及更多实验，请参见附录 B。

5 一般理论

如前文所述，尽管许多研究表明随机梯度下降促进平坦性，但近期对 EoS 行为的观察显示，随机梯度下降也可能收敛到非常“尖锐”的解。我们的理论认为，这种倾向上的差异源于尖锐度与波动之间一个细微的区别。在本节中，我们将从一般视角审视这一问题，并强调我们的结果如何有助于深化对此的理解。

考虑均方误差（MSE）损失[4]

$$\ell(x,\theta) := \frac{1}{2}\|f(x,\theta) - y(x)\|^2. \tag{23}$$

海森矩阵为

$$H(\theta) := \mathbb{E}_x \nabla^2 \ell(x,\theta) = \mathbb{E}_x \left(\nabla f(\theta,x) \nabla f(\theta,x)^T + \sum_{j=1}^{d_y} [e(x,\theta)]_j \nabla_\theta^2 [f(x,\theta)]_j \right), \tag{24}$$

其中 $e(x,\theta) := y(x) - f(\theta,x)$ 为残差误差。在全局最小值附近，$e(x,\theta)$ 变得可以忽略，因此我们可以将尖锐度（用海森矩阵的迹来量化）近似为

$$\mathrm{Tr} H(\theta) \approx \mathbb{E}_x \|\nabla f(\theta,x)\|^2, \tag{25}$$

这种近似在现有文献中已广为人知 [WS23，MVL23]。相比之下，近期研究表明，随机梯度下降最小化的并非尖锐度，而是以下梯度波动：

$$\mathbb{E}_x \|\nabla \ell(x,\theta)\|^2 = \mathbb{E}_x \|\nabla f(\theta,x)^T e(x,\theta)\|^2, \tag{26}$$

假设模型训练良好，残差 $e(x,\theta)$ 主要反映标签噪声。如果我们将该噪声视为与 x 独立、协方差为 Σ_ϵ 的随机变量，则波动变为：

$$\mathbb{E}_x \|\nabla \ell(x,\theta)\|^2 = \mathbb{E}_x [\nabla f(x,\theta)^T \Sigma_\epsilon \nabla f(x,\theta)], \tag{27}$$

它与尖锐度相关，但并不相同。

定理 2。 假设残差 $e(x,\theta) := y(x) - f(\theta,x)$ 是一个零均值、方差为 Σ_ϵ 且与数据集独立的随机变量。那么，

$$\lambda_{\min}(\Sigma_\epsilon) S \le T \le \lambda_{\max}(\Sigma_\epsilon) S, \tag{28}$$

其中，T 是尖锐度，S 是梯度波动。

证明。 根据（24）我们有

$$T := \mathrm{Tr} H(\theta) = \mathbb{E}_x \|\nabla f(\theta,x)\|^2, \tag{29}$$

其中我们使用 假设 $e(x,\theta)$ 与 x 相互独立且具有零均值。根据（27），我们有 e

$$S := \mathbb{E}_x \|\nabla \ell(x,\theta)\|^2 = \mathrm{Tr}[\Sigma_\epsilon \mathbb{E}_x [\nabla f(x,\theta) \nabla f(x,\theta)^T]]. \tag{30}$$

将它们结合起来即可完成证明。 □

这个不等式明确指出，在什么情况下，最小化波动与最小化平坦性并不相同。关键在于，只有当标签噪声是各向同性时，波动才是尖锐度的良好近似。当 $\lambda_{\max}(\Sigma_\epsilon)$ $\lambda_{\min}(\Sigma_\epsilon)$ 较大时，这个界会变得非常松，最小化波动也就不再能导向平坦性的最小化。

[4]类似地，我们可以 通过考察标签的谱不平衡，将讨论扩展到一般损失函数 l
海森矩阵：$\nabla_y^2 \ell$.

6 相关工作

关于尖锐度和稳定性边缘的现有极简模型。在随机梯度下降和梯度下降中首次通过实证发现渐进锐化和稳定性边缘（EoS）效应后 [XATB18, JKB+19, JSF+20, CKL+21, AB24, CDT+24]，人们投入了大量精力，借助极简模型分析这些现象。 [APP22] 从解析上证明，二次模型中的渐进锐化是 EoS 的直接结果。 [LWL22]展示了采用 tanh 激活函数的浅层网络中的渐进锐化。[DNL22] 将这些效应归因于损失函数泰勒展开中的三次项的影响。我们的工作不同于这些研究：我们证明，这些现象可能源于随机梯度下降内在的随机性，而不只是步长的函数。值得注意的是， [ABC+23] 证明了两层 ReLU 网络在 EoS 状态下存在振荡，并强调了振荡对泛化的益处；而 [BW25, GKW+25] 则认为，线性模型中的振荡会隐式地使梯度下降偏向更平坦的解。近期研究考察了 EoS 状态下的收敛及其他隐式正则化方面，例如参见[ALP22, WCZT22, AZS22, ZWW+23, LLA23]。

深度线性网络。我们研究的核心对象是深度线性网络模型（也称为深度矩阵分解）。深度线性网络是理解复杂架构训练动力学的重要代理。然而，现有文献大多关注它们在梯度下降或梯度流下的行为 [BH89, SMG13, GWB+17, ACGH18, Eft20, BRTW22, DH19, GBLJ19, TFHV21, XMT+23, NRT24], ，其中一些研究表明，在这些模型中，梯度下降在参数空间中具有偏向更平坦的极小值的隐式偏置 [MC24, BW25, GKW+25]。尽管一些研究也刻画了它们的损失景观——表明所有局部极小值都是全局极小值 [Kaw16, LB18]——噪声的作用仍较少被探索。尽管 [ZLM22] 分析了带权重衰减的深度线性网络的全局极小值，但我们的分析专门针对随机梯度下降的随机性所诱导的隐式偏置。其他密切相关的工作有[PPVF21, EPGF23, BPP24], ，它们分析了对角神经网络（具有一维标签协方差的多重动力学），并表明随机性是有益的。

标签噪声的影响。 关于标签噪声的隐式偏置已有大量研究。这些研究表明，标签噪声具有隐式正则化效应，会使模型偏向噪声较弱的极小值 [BGVV20、LWA21、HWLM21] 或更平坦的极小值 [DML21]。这些研究还表明，这种正则化效应有助于提升泛化能力 [PVRF22、HHS+25]。重要的是，所有这些工作都只考虑一维输出。一维矩阵的条件数始终为 1。相比之下，我们的工作表明，噪声标注也可能使模型偏向更尖锐的解，具体取决于条件数。

7 结论

在这项工作中，我们深入研究了随机梯度下降训练在何种情况下会表现出寻求平坦性或尖锐度的行为。我们针对满足最小梯度波动约束的深度线性网络，给出了其动力学的精确解析解。在该模型中，我们证明收敛锐度是唯一的，且与初始化无关。一个关键发现是，标签不平衡噪声是锐化现象的主要驱动因素；通过在不同模型类型和真实数据集上进行实验，我们进一步验证了这一结论。这是一种新颖的渐进锐化形式，其性质不同于文献中先前讨论的形式。

其中每一项都值得在未来的工作中加以解决。首先，我们的理论本质上是静态的，无法直接揭示训练的动力学方面。因此，一个有价值的开放问题是研究如何动态地达到更尖锐和更平坦的解。其次，我们的理论只研究了训练期间 η 的主导阶效应，而这种效应倾向于抑制波动。这本质上对应于训练中的小学习率状态（但学习率还没有小到梯度流所需的程度），而熵损失中的高阶项可能发挥什么作用尚不清楚。因此，较大学习率下的平坦化和锐化效应仍有待研究。例如，学习率非常大时，寻锐行为很可能会再次受到强烈抑制。出于类似的原因，在我们的框架内只能解释渐进锐化，无法解释 EoS。这是因为要推导熵损失（12），基本上需要假设学习

在某种程度上是稳定的 [SDBD21]。解决这些问题无疑将推进我们对深度学习优化的理解。

致谢

作者感谢 Tomaso Poggio 教授的讨论。ILC 感谢人工智能与基本相互作用研究所（IAIFI）通过美国国家科学基金会资助编号 PHY-2019786 提供的部分支持。

参考文献

[AB24] Arseniy Andreyev 和 Pierfrancesco Beneventano。随机稳定性边界：重新审视随机梯度下降的稳定性边缘。arXiv 预印本 arXiv:2412.20553，2024年。[ABC+23] Kwangjun Ahn、Sébastien Bubeck、Sinho Chewi、Yin Tat Lee、Felipe Suarez 和 Yi 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。[ACGH18] Sanjeev Arora、Nadav Cohen、Noah Golowich 和 Wei Hu。深度线性神经网络梯度下降的收敛性分析。arXiv 预印本 arXiv:1810.02281，2018年。[ALP22] Sanjeev Arora、Zhiyuan Li 和 Abhishek Panigrahi。理解深度学习中稳定性边缘上的梯度下降，2022年10月。arXiv:2205.09745 [cs]。[APP22] Atish Agarwala、Fabian Pedregosa 和 Jeffrey Pennington。二阶回归模型表现出渐进锐化，直至达到稳定性边缘。arXiv 预印本 arXiv:2210.04860，2022年。[AZS22] Kwangjun Ahn、Jingzhao Zhang 和 Suvrit Sra。理解梯度下降的不稳定收敛。发表于第39届国际机器学习会议论文集，2022年6月。[Ben23] Pierfrancesco Beneventano。关于随机梯度下降不放回时的轨迹。arXiv 预印本 arXiv:2312.16143，2023年。[BGVV20] Guy Blanc、Neha Gupta、Gregory Valiant 和 Paul Valiant。由类似奥恩斯坦-乌伦贝克过程驱动的深度神经网络隐式正则化。发表于学习理论会议，页码483–513。PMLR，2020年。[BH89] Pierre Baldi 和 Kurt Hornik。神经网络与主成分分析：无需局部极小值的示例学习。《神经网络》，2(1)：53–58，1989年1月。[BPP24] Pierfrancesco Beneventano、Andrea Pinto 和 Tomaso Poggio。神经网络如何学习支撑集是随机梯度下降的一种隐式正则化效应。arXiv:2406.*11110* [cs, math, stat]，2024年6月。arXiv:2406.11110 [cs, math, stat]。[BRTW22] Bubacarr Bah、Holger Rauhut、Ulrich Terstiege 和 Michael Westdickenberg。深度线性神经网络学习：黎曼梯度流与收敛至全局极小值点。《IMA信息与推断期刊》，11(1)：307–353，2022年。[BW25] Pierfrancesco Beneventano 和 Blake Woodworth。在浅层线性网络中，梯度下降线性收敛至比梯度流更平坦的极小值。arXiv 预印本 arXiv:2501.09137，2025年。[CDT+24] Jeremy M. Cohen、Alex Damian、Ameet Talwalkar、Zico Kolter 和 Jason D. Lee。通过中心流理解深度学习中的优化，2024年10月。arXiv:2410.24206。[CGK+22] Jeremy M Cohen、Behrooz Ghorbani、Shankar Krishnan、Naman Agarwal、Sourabh Medapati、Michal Badura、Daniel Suo、David Cardoze、Zachary Nado、George E Dahl 等。稳定性边缘处的自适应梯度方法。arXiv 预印本 arXiv:2207.14484，2022年。

[CKL+21]Jeremy M 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性边缘。arXiv 预印本 arXiv:2103.00065，2021年。[DH19] Simon Du 和 Wei Hu。宽度在深度线性神经网络优化中具有可证明的重要性。发表于国际机器学习会议，页码1655–1664。PMLR，2019年。[DML21] Alex Damian、Tengyu Ma 和 Jason D Lee。带标签噪声的随机梯度下降可证明会偏好平坦的全局极小值。神经信息处理系统进展，34：27449–27461，2021年。[DNL22] Alex Damian、Eshaan Nichani 和 Jason D Lee。自稳定：稳定性边缘处梯度下降的隐式偏置。arXiv 预印本 arXiv:2209.15594，2022年。[DPBB17] L. Dinh、R. Pascanu、S. Bengio 和 Y. Bengio。尖锐极小值能够在深度网络中实现泛化。ArXiv 电子预印本，2017年3月。[Eft20] Armin Eftekhari。线性神经网络训练：非局部收敛与复杂度结果。发表于国际机器学习会议，页码2836–2847。PMLR，2020年。[EPGF23] Mathieu Even、Scott Pesme、Suriya Gunasekar 和 Nicolas Flammarion。（s）gd 在对角线性网络上的表现：隐式偏置、大步长与稳定性边缘。神经信息处理系统进展，36：29406–29448，2023年。[FKMN20] Pierre Foret、Ariel Kleiner、Hossein Mobahi 和 Behnam Neyshabur。用于高效提升泛化能力的尖锐度感知最小化。arXiv 预印本 arXiv:2010.01412，2020年。[GBLJ19] Gauthier Gidel、Francis Bach 和 Simon Lacoste-Julien。线性神经网络中离散梯度动力学的隐式正则化。发表于神经信息处理系统进展，第32卷。Curran Associates, Inc.，2019年。[GKW+25] Avrajit Ghosh、Soo Min Kwon、Rongrong Wang、Saiprasad Ravishankar 和 Qing Qu。超越稳定性边缘的深度线性网络学习动力学，2025年。[GWB+17] Suriya Gunasekar、Blake E Woodworth、Srinadh Bhojanapalli、Behnam Neyshabur 和 Nati Srebro。矩阵分解中的隐式正则化。发表于神经信息处理系统进展，页码6151–6159，2017年。[HHS+25] Wei Huang、Andi Han、Yujin Song、Yilan Chen、Denny Wu、Difan Zou 和 Taiji Suzuki。低信噪比区间内，标签噪声梯度下降如何提升泛化能力？arXiv 预印本 arXiv:2510.17526，2025年。[HS97] Sepp Hochreiter 和 Jürgen Schmidhuber。平坦极小值。《神经计算》，9(1)：1–42，1997年。[HWLM21] Jeff Z HaoChen、Colin Wei、Jason Lee 和 Tengyu Ma。形状很重要：理解噪声协方差的隐式偏置。发表于学习理论会议，页码2315–2357。PMLR，2021年。[JKB+19] Stanisław Jastrzębski、Zachary Kenton、Nicolas Ballas、Asja Fischer、Yoshua Bengio 和 Amos Storkey。深度神经网络损失的最尖锐方向与随机梯度下降步长之间的关系，2019年12月。arXiv:1807.05031 [stat]。[JSF+20]Stanisław Jastrzębski、Maciej Szymczak、Stanislav Fort、Devansh Arpit、Jacek Tabor、Kyunghyun Cho 和 Krzysztof Geras。深度神经网络优化轨迹上的盈亏平衡点。arXiv:2002.*09572* [cs, stat],2020年2月。arXiv:2002.09572。[Kaw16] Kenji Kawaguchi。没有较差局部极小值的深度学习。神经信息处理系统进展，29，2016年。[KMN+16]Nitish Shirish Keskar、Dheevatsa Mudigere、Jorge Nocedal、Mikhail Smelyanskiy 和 Ping Tak Peter Tang。关于深度学习的大批量训练：泛化差距与尖锐极小值。arXiv 预印本 arXiv:1609.04836，2016年。

[KP25] Anil Kamber与Rahul Parhi. 深度矩阵分解中极小值的尖锐度：精确表达式. arXiv预印本 arXiv:2509.25783, 2025年.[LB18] Thomas Laurent与James Brecht. 具有任意损失的深度线性网络：所有局部极小值都是全局极小值. 载于国际机器学习会议, 第2902–2907页. PMLR, 2018年.[LLA23] Kaifeng Lyu、Zhiyuan Li与Sanjeev Arora. 理解归一化层的泛化收益：尖锐度降低, 2023年1月. arXiv:2206.07085 [cs].[LWA21] Zhiyuan Li、Tianhao Wang与Sanjeev Arora. 随机梯度下降达到零损失后会发生什么？——一个数学框架. 载于国际学习表征会议, 2021年.[LWL22] Zhouzi Li、Zixuan Wang与Jian Li. 沿梯度下降轨迹分析尖锐度：渐进锐化与稳定性边缘, 2022年11月. arXiv:2207.12678.[MC24] Pierre Marion与Lénaïc Chizat. 用于回归的深度线性网络被隐式正则化趋向平坦极小值. 神经信息处理系统进展, 37:76848–76900, 2024年.[MM24] Rotem Mulayoff与Tomer Michaeli. 随机梯度下降的精确均方线性稳定性分析, 2024年.[MVL23] Lachlan Ewen MacDonald、Jack Valmadre与Simon Lucey. 论渐进锐化、平坦极小值与泛化. arXiv预印本 arXiv:2305.14683, 2023年.[NRT24]Gabin Maxime Nguegnang、Holger Rauhut与Ulrich Terstiege. 学习线性神经网络时梯度下降的收敛. 连续与离散模型进展, 2024(1):23, 2024年.[PPVF21] Scott Pesme、Loucas Pillaud-Vivien与Nicolas Flammarion. 随机梯度下降对对角线性网络的隐式偏置：随机性的可证明收益. 神经信息处理系统进展, 34:29218–29230, 2021年.[PVRF22] Loucas Pillaud-Vivien、Julien Reygner与Nicolas Flammarion. 标签噪声（随机）梯度下降对二次参数化隐式求解套索回归. 载于学习理论会议, 第2127–2159页. PMLR, 2022年.[Rob21] Daniel A. Roberts. 随机梯度下降隐式正则化泛化误差. arXiv:2104.*04874* [cs, stat],2021年4月. arXiv:2104.04874.[SDBD21] Samuel L Smith、Benoit Dherin、David GT Barrett与Soham De. 论随机梯度下降中隐式正则化的起源. arXiv预印本 arXiv:2101.12176, 2021年.[SMG13] Andrew M Saxe、James L McClelland与Surya Ganguli. 深度线性神经网络中学习非线性动力学的精确解. arXiv预印本 arXiv:1312.6120, 2013年.[TFHV21] Salma Tarmoun、Guilherme Franca、Benjamin D Haeffele与Rene Vidal. 理解过参数化线性模型中梯度流的动力学. 载于国际机器学习会议, 第10153–10161页. PMLR, 2021年.[WCZT22] Yuqing Wang、Minshuo Chen、Tuo Zhao与Molei Tao. 大学习率驯服齐次性：收敛与平衡效应. 载于国际学习表征会议, 2022年.[WM+18] Lei Wu、Chao Ma等. 随机梯度下降如何在过参数化学习中选择全局极小值：动力学稳定性视角. 神经信息处理系统进展, 31, 2018年.[WS23] Lei Wu与Weijie J Su. 随机梯度下降中动力学稳定性的隐式正则化. 载于国际机器学习会议, 第37656–37684页. PMLR, 2023年.[WWS22] Lei Wu、Mingze Wang与Weijie Su. 随机梯度下降噪声的对齐性质及其如何帮助选择平坦极小值：一项稳定性分析. 神经信息处理系统进展 , 35:4680–4693, 2022 年 .

[XATB18] Chen Xing、Devansh Arpit、Christos Tsirigotis与Yoshua Bengio. 与随机梯度下降同行, 2018年5月. arXiv:1802.08770 [cs, stat].[XMT+23] Ziqing Xu、Hancheng Min、Salma Tarmoun、Enrique Mallada与René Vidal. 具有一般初始化的有限宽度过参数化线性网络上梯度下降的线性收敛. 载于国际人工智能与统计大会, 第2262–2284页. PMLR, 2023年.[XSS20] Zeke Xie、Issei Sato与Masashi Sugiyama. 深度学习动力学的扩散理论：随机梯度下降以指数方式偏好平坦极小值. *arXiv*预印本 *arXiv:2002.03495*, 2020年.[ZLM22] Liu Ziyin、Botao Li与Xiangming Meng. 深度线性网络的精确解. 神经信息处理系统进展, 35:24446–24458, 2022年.[ZWLW24] Liu Ziyin、Mingze Wang、Hongchao Li与Lei Wu. 随机梯度下降的参数对称性与噪声平衡. 载于第三十八届神经信息处理系统年会, 2024年.[ZWW+23] Xingyu Zhu、Zixuan Wang、Xiang Wang、Mo Zhou与Rong Ge. 用极简示例理解稳定性边缘训练动力学, 2023年.[ZWY+18] Zhanxing Zhu、Jingfeng Wu、Bing Yu、Lei Wu与Jinwen Ma. 随机梯度下降中的各向异性噪声：其逃离尖锐极小值的行为与正则化效应. *arXiv*预印本 *arXiv:1803.00195*, 2018年.[ZXC25] Liu Ziyin、Yizhou Xu与Isaac Chuang. 神经热力学：深度与通用表征学习中的熵力. 神经信息处理系统进展, 2025年.

理论

A.1 引理证明 1

证明。 根据 A-对称性的定义，损失函数对于任意λ均满足以下不变性：

$$\ell(x, y, e^{\lambda A}\theta) = \ell(x, y, \theta). \tag{31}$$

为分析尖锐度，我们对等式两边关于 θ 求二阶导数。应用链式法则，可得：

$$e^{\lambda A}\nabla^2_{e^{\lambda A}\theta}\ell(x, y, e^{\lambda A}\theta)e^{\lambda A} = \nabla^2\ell(x, y, \theta), \tag{32}$$

其中我们利用了 A 是对称的这一事实。

利用迹的线性性和循环性质，我们有：

$$T(e^{\lambda A}\theta) := \mathrm{Tr}[\mathbb{E}_{x,y}\nabla^2_{e^{\lambda A}\theta}\ell(x, y, e^{\lambda A}\theta)] = \mathrm{Tr}[e^{-2\lambda A}\mathbb{E}_{x,y}\nabla^2\ell(x, y, \theta)]. \tag{33}$$

令 $A := \sum_i \mu_i n_i n_i^T$。于是，尖锐度可以表示为加权和：

$$T(e^{\lambda A}\theta) = \sum_i e^{-2\lambda\mu_i}(n_i^T\mathbb{E}_{x,y}\nabla^2\ell(x, y, \theta)n_i). \tag{34}$$

根据假设 $A\mathbb{E}x,\ y\nabla^2\ell(x,\ y,\ \theta) \neq 0$，必定存在至少一个指标 i，使得 $\mu_i \neq 0$且$n_i^T\mathbb{E}\nabla^2\ell(x,\ \theta)n_i \neq 0$。因此，当 λ 趋于无穷大且符号适当时，项 $e^{-2\lambda\mu_i}$ 将占主导地位。具体而言，若存在 $\mu_i < 0$，则有 $\lim_{\lambda\to+\infty}|T(e^{\lambda A}\theta)| = +\infty$；若存在 $\mu_i > 0$，则有 $\lim_{\lambda\to-\infty}|T(e^{\lambda A}\theta)| = +\infty$。 □

A.2 引理 2 的证明

我们首先正式陈述引理 2。

引理 3。 设 f、S为连续函数，且 f为强制的（即，$\lim_{|\theta|\to\infty} f(\theta) = +\infty$）。定义扰动目标函数

$$F_\eta(\theta) = f(\theta) + \eta S(\theta). \tag{35}$$

令 $\{\eta_k\}_{k=1}^\infty$ 为一个正数列，使得 $\eta_k \to 0$ 当 $k \to \infty$时成立。假设

$$\theta_k \in \arg\min_\theta f(\theta) + \eta_k S(\theta). \tag{36}$$

并且 $\theta_k \to \bar{\theta}$[5] 当 $k \to \infty$时成立。则 $\bar{\theta}$ 在 f的极小值点集合上使 S 达到最小，即

$$\bar{\theta} \in \arg\min_{\theta\in M} S(\theta), \quad where\ M := \arg\min_\theta f(\theta). \tag{37}$$

证明。 由于 f 是强制的且连续的，极小值点集合 M 非空且紧。令 $y \in M$为 f的任意一个全局极小化点。根据 θ_k关于 F_{η_k}的最优性，我们有：

$$f(\theta_k) + \eta_k S(\theta_k) \le f(y) + \eta_k S(y). \tag{38}$$

重新整理各项，得到：

$$f(\theta_k) - f(y) \le \eta_k(S(y) - S(\theta_k)) \tag{39}$$

由于 $y \in M$，我们有 $f(\theta_k) \ge f(y)$，这意味着 $f(\theta_k) - f(y) \ge 0$。因此，由（39）以及 $\eta_k > 0$这一事实可得：

$$S(\theta_k) \le S(y). \tag{40}$$

[5]由于 f 是强制的， M 是紧集，因此存在 $\{\theta_k\}$ 的一个子序列收敛。为简单起见，我们假设 e $\{\theta_k\}$ 收敛。

在（39）中令 $k \to \infty$，并注意到 S 是连续的（因此 $S(\theta_k)$ 在 $\theta_k \to \theta$时仍有界），右端趋于零：

$$0 \le \lim_{k\to\infty} (f(\theta_k) - f(y)) \le \lim_{k\to\infty} \eta_k (S(y) - S(\theta_k)) = 0. \tag{41}$$

根据 f 的连续性，我们有 $f(\theta) = f(y)$，这意味着 $\theta \in M$。

最后，我们考虑不等式（40）。对两边取极限 $k \to \infty$ 并利用 S的连续性，可得：

$$S(\bar{\theta}) \le S(y) \quad \text{for all } y \in M. \tag{42}$$

该不等式表明，θ 在 M 上使 S 取得最小值，即 $\theta \in \arg\min_{\theta\in M} S(\theta)$，至此证明完成。 □

A.3 定理1的证明

证明。 步骤 1：计算海森矩阵的迹。

我们首先计算关于权重矩阵 $\{W_k\}_{k=1}^D$ 的海森矩阵的迹。对于每一层 k，令 $A_k := W_D \cdots W_{k+1}$ 和 $B_k := W_{k-1} \cdots W_1$，并约定 $A_D = I_{d_y}$ 和 $B_1 = I_{d_x}$。单个样本的损失函数可写为 $\ell(x, y, \theta) = \frac{1}{2}\|y - A_k W_k B_k x\|^2$。关于 W_k 的二阶偏导数给出以下恒等式：

$$\mathrm{Tr}\frac{\partial^2 \ell}{\partial W_k^2} = \mathrm{Tr} A_k^T A_k \mathrm{Tr} B_k x x^T B_k^T. \tag{43}$$

对数据分布取期望，并对所有层 $k = 1$、……、D 求和，可得总尖锐度 $T(\theta)$：

$$T(\theta) = \sum_{k=1}^{D} \mathbb{E}_{x,y} \mathrm{Tr}\frac{\partial^2 \ell}{\partial W_k^2} = \sum_{k=1}^{D} \mathrm{Tr}[A_k^T A_k] \mathrm{Tr}[B_k \Sigma_x B_k^T]. \tag{44}$$

这就确立了尖锐度的结构形式。根据（ $T(\theta)$ （15）的最小化，直接对（4 44）中的求和应用引理 即可得出。

步骤 2：分析（12）的最小值处的尖锐度。

对于引理5中的解，乘积矩阵可分解为

$$A_k = \sqrt{\Sigma_\epsilon^{-1}} L \Sigma_D \cdots \Sigma_{k+1} U_k^T, \quad B_k = U_{k-1} \Sigma_{k-1} \cdots \Sigma_1 R \Sigma_x^{-1/2}. \tag{45}$$

对于中间层 $k \in \{2, \ldots, D-1\}$，迹乘积可简化为：

$$\begin{aligned} \mathrm{Tr}[A_k^T A_k] \mathrm{Tr}[B_k \Sigma_x B_k^T] &= \mathrm{Tr}[\Sigma_\epsilon^{-1} L \Sigma_D^2 \cdots \Sigma_{k+1}^2 L^T] \mathrm{Tr}[\Sigma_1^2 \cdots \Sigma_{k-1}^2] \\ &= (\mathrm{Tr} S)^{-2/D} (\mathrm{Tr}[\Sigma_\epsilon] \mathrm{Tr}[\Sigma_x])^{1/D} \mathrm{Tr}[\Sigma_\epsilon^{-1} L S L^T] \mathrm{Tr}[S] \end{aligned} \tag{46}$$

边界层需要单独处理。对于 $k = 1$ 和 $k = D$，有：

$$\mathrm{Tr}[A_1^T A_1] \mathrm{Tr}[B_1 \Sigma_x B_1^T] = \mathrm{Tr}[\Sigma_\epsilon^{-1} L \Sigma_D^2 \cdots \Sigma_2^2 L^T] \mathrm{Tr}[\Sigma_x] = (\mathrm{Tr} S)^{(D-2)/D} \frac{\mathrm{Tr}[\Sigma_\epsilon]^{1/D}}{\mathrm{Tr}[\Sigma_x]^{(D-1)/D}} \mathrm{Tr}[\Sigma_\epsilon^{-1} L S L^T] \mathrm{Tr}[\Sigma_x] \tag{47}$$

and

$$\mathrm{Tr}[A_D^T A_D] \mathrm{Tr}[B_D \Sigma_x B_D^T] = \mathrm{Tr}[I_{d_y}] \mathrm{Tr}[\Sigma_1^2 \cdots \Sigma_{D-1}^2] = d_y (\mathrm{Tr} S)^{(D-2)/D} \frac{\mathrm{Tr}[\Sigma_x]^{1/D}}{\mathrm{Tr}[\Sigma_\epsilon]^{(D-1)/D}} \mathrm{Tr}[S]. \tag{48}$$

将这些贡献从 $k = 1, \ldots, D$ 相加，得到 $T(\theta)$ 的最终表达式：

$$\begin{aligned} T(\theta) &= \sum_{k=1}^{D} \mathrm{Tr}[A_k^T A_k] \mathrm{Tr}[B_k \Sigma_x B_k^T] \\ &= (D-1) (\mathrm{Tr} S)^{(D-2)/D} (\mathrm{Tr}[\Sigma_\epsilon] \mathrm{Tr}[\Sigma_x])^{1/D} \mathrm{Tr}[\Sigma_\epsilon^{-1} L S L^T] + d_y (\mathrm{Tr} S)^{2(D-1)/D} \frac{\mathrm{Tr}[\Sigma_x]^{1/D}}{\mathrm{Tr}[\Sigma_\epsilon]^{(D-1)/D}}, \end{aligned} \tag{49}$$

这就证明了（14）。

步骤 3：海森矩阵谱的不变性。

最后，我们验证海森矩阵 $H(\theta)$在引理 5 所定义的解流形上处处保持常数。考虑两组权重矩阵 $\theta = \{W_k\}$ 和 $\theta' = \{W'_k\}$。根据该引理，任意两个这样的解都可以通过一系列正交变换 $R_k \in O(d_k)$ 相互关联，满足 $W'_k = R_k W_k R_{k-1}^T$。该变换可表示为 $\theta' = \mathrm{M}\theta$，其中 M 是块对角正交矩阵。由于损失函数在这些内部旋转下保持不变，海森矩阵变换为 $H(\theta') = \mathrm{M}H(\theta)\mathrm{M}^T$。由于 M 是正交矩阵，这属于相似变换，因此海森矩阵的特征值（以及其谱）保持不变。

□

以下是定理 1的技术引理。

引理 4。考虑

$$\begin{aligned} T(W_1,\cdots,W_D) = d_y \mathrm{Tr}[W_{D-1}\cdots W_1 \Sigma W_1^T \cdots W_{D-1}^T] + \mathrm{Tr}[W_D^T W_D]\mathrm{Tr}[W_{D-2}\cdots W_1 \Sigma W_1^T \cdots W_{D-2}^T] \\ + \cdots + \mathrm{Tr}[W_2^T \cdots W_D^T W_D \cdots W_2]\mathrm{Tr}[\Sigma], \end{aligned} \tag{50}$$

其中 d_y 是输出维度。然后，在约束 $W_D \cdots W_1 = V$ 下，我们有 ve

$$\min T(W_1,\cdots,W_D) = D(\mathrm{Tr}\hat{S})^{2(D-1)/D}(d_y \mathrm{Tr}\Sigma)^{1/D}, \tag{51}$$

其中 $V\sqrt{\Sigma} = \hat{}L\hat{}S\hat{}R$ 是奇异值分解。最小值在以下条件下达到

$$\begin{aligned} W_D &= (\mathrm{Tr}S)^{-\frac{D-2}{2D}} d_y^{\frac{D-1}{2D}} (\mathrm{Tr}\Sigma)^{-\frac{1}{2D}} \hat{L}\hat{S}^{\frac{1}{2}} U_{D-1}^T, \\ W_k &= (\mathrm{Tr}S)^{\frac{1}{D}} (d_y \mathrm{Tr}\Sigma)^{-\frac{1}{2D}} U_k^T U_{k-1}, \\ W_1\sqrt{\Sigma} &= (\mathrm{Tr}S)^{-\frac{D-2}{2D}} d_y^{-\frac{1}{2D}} (\mathrm{Tr}\Sigma)^{\frac{D-1}{2D}} U_1 \hat{S}^{\frac{1}{2}} \hat{R}, \end{aligned} \tag{52}$$

对于 $k = 2$、$\cdots$、$D-1$，其中 $\{U_i\}_{i=1}^{D-1}$是满足 $U_i^T U_i = I_{d\times d}$ 的任意矩阵。

证明。步骤 1：下界。

我们希望最小化（ D 50）中的各项之和。为简洁起见，令 $W'_1 = W_1\Sigma^{1/2}$。根据算术-几何平均不等式，尖锐度 T 的下界为其各分量的几何平均数：

$$\begin{aligned} T(W_1,\cdots,W_D)^D/D^D &\geq d_y \|W_{D-1}\cdots W'_1\|_F^2 \|W_D\|_F^2 \|W_{D-2}\cdots W'_1\|_F^2 \cdots \|W_D\cdots W_2\|_F^2 \mathrm{Tr}[\Sigma] \\ &= d_y \mathrm{Tr}[\Sigma] \Pi_{k=1}^{D-1} \|W_D\cdots W_{k+1}\|_F^2 \|W_k\cdots W'_1\|_F^2 \end{aligned} \tag{53}$$

应用矩阵不等式

$$\|A\|_F \|B\|_F \geq \|AB\|_* \tag{54}$$

其中，$\|\cdot\|_*$ 是核范数（奇异值之和），对于乘积中的每一对，我们有：

$$\|W_D \ldots W_{k+1}\|_F \|W_k \ldots W'_1\|_F \geq \|W_D \ldots W'_1\| = \|V\Sigma^{1/2}\|_* = \mathrm{Tr}\hat{S}. \tag{55}$$

将其代入算术-几何平均结果，我们得到最小锐度：

$$T(W_1,\cdots,W_D)^D/D^D \geq d_y \mathrm{Tr}[\Sigma](\mathrm{Tr}\hat{S})^{2(D-1)}. \tag{56}$$

步骤 2：等号成立条件。

当且仅当 $A_k^T A_k = cB_k B_k^T$ 对于某个 $c \geq 0$ 且式（50）中的每一项都相等时，才达到最小值，这要求满足以下关系：

$$W_{k+1}^\top \frac{W_{k+2}^\top \cdots W_D^\top W_D \cdots W_{k+2}}{\mathrm{Tr}\left[W_{k+2}^\top \cdots W_D^\top W_D \cdots W_{k+2}\right]} W_{k+1} = W_k \frac{W_{k-1}\cdots W_1 \Sigma W_1^\top \cdots W_{k-1}^\top}{\mathrm{Tr}\left[W_{k-1}\cdots W_1 \Sigma W_1^\top \cdots W_{k-1}^\top\right]} W_k^\top \tag{57}$$

对于 $k = 2,3,\cdots,D-2$ 和

$$\begin{aligned} W_2^\top \frac{W_3^\top \cdots W_D^\top W_D \cdots W_3}{\mathrm{Tr}\left[W_3^\top \cdots W_D^\top W_D \cdots W_3\right]} W_2 &= \frac{W_1 \Sigma W_1^\top}{\mathrm{Tr}\left[\Sigma\right]}, \\ \frac{W_D^\top W_D}{d_y} &= W_{D-1} \frac{W_{D-2}\cdots W_1 \Sigma W_1^\top \cdots W_{D-2}^\top}{\mathrm{Tr}\left[W_{D-2}\cdots W_1 \Sigma W_1^\top \cdots W_{D-2}^\top\right]} W_{D-1}^\top. \end{aligned} \tag{58}$$

步骤 3：确定权重。

根据 [引理 225],，我们可以将矩阵 W_1'、W_2、$\cdots$、W_{D-1}、W_D 分解为

$$W_D = U_D \Sigma_D U_{D-1}^T,\ W_{D-1} = U_{D-1}\Sigma_{D-1}U_{D-2}^T, \cdots,\ W_1' = U_1\Sigma_1 U_0, \tag{59}$$

其中 $W_1' := W_1\sqrt{\Sigma_\circ}$ 由 $W_D \cdots W_1 = V$ 可得 $U_D = \hat{L}$、$U_0 = \hat{R}$，以及

$$\Sigma_D \Sigma_{D-1} \cdots \Sigma_1 = \hat{S}. \tag{60}$$

不失一般性，我们可以假设 Σ_D、$\cdots$、$\Sigma_1 \in \mathbb{R}^{d\times d}$，因为根据式（57）、（58），它们的秩相同。于是，由式（57）可得

$$\frac{\Sigma_{i+1}^2\Sigma_{i+2}^2\cdots\Sigma_D^2}{\mathrm{Tr}[\Sigma_{i+2}^2\cdots\Sigma_D^2]} = \frac{\Sigma_1^2\cdots\Sigma_{i-1}^2\Sigma_i^2}{\mathrm{Tr}[\Sigma_1^2\cdots\Sigma_{i-1}^2]}, \tag{61}$$

因此，对于 $i = 2,\ 3,\ \cdots,\ D-2$ 以及 $\Sigma_i = cI_d$

$$\frac{\Sigma_1^2}{\mathrm{Tr}[\Sigma_1^2]} = \frac{\Sigma_D^2}{\mathrm{Tr}[\Sigma_D^2]}. \tag{62}$$

式（58）给出

$$\frac{\Sigma_2^2\Sigma_3^2\cdots\Sigma_D^2}{\mathrm{Tr}[\Sigma_3^2\cdots\Sigma_D^2]} = \frac{\Sigma_1^2}{\mathrm{Tr}[\Sigma]},\ \frac{\Sigma_D^2}{d_y} = \frac{\Sigma_1^2\cdots\Sigma_{D-1}^2}{\mathrm{Tr}[\Sigma_1^2\cdots\Sigma_{D-2}^2]}, \tag{63}$$

因此

$$c^2\frac{\Sigma_D^2}{\mathrm{Tr}[\Sigma_D^2]} = \frac{\Sigma_1^2}{\mathrm{Tr}[\Sigma]},\ \frac{\Sigma_D^2}{d_y} = c^2\frac{\Sigma_1^2}{\mathrm{Tr}[\Sigma_1^2]}. \tag{64}$$

结合（60）、（62）和（64），证明得证。 □

下面的引理是 [ZXC25]中的定理 12 的重述。

引理 5。考虑一个 D–层线性网络，每层都具有超过 d 个隐藏单元。那么在（12）的全局最小值处，有

$$\sqrt{\Sigma_\epsilon}W_D = L\Sigma_D U_{D-1}^T,\ W_i = U_i\Sigma_i U_{i-1}^T,\ W_1\sqrt{\Sigma_x} = U_1\Sigma_1 R, \tag{65}$$

对于 $i = 2$、$\cdots$、$D-1$，其中 $\{U_i\}_{i=1}^{D-1}$是满足 $U_i^T U_i = I_{d\times d}$ 的任意矩阵，且 $\Sigma_x = \mathbb{E}[xx^T]$、$\Sigma_\epsilon = \mathbb{E}[\epsilon\epsilon^T]$。此外，

$$\begin{aligned}
\Sigma_1 &= (\mathrm{Tr}S')^{-\frac{D-2}{2D}}\frac{\mathrm{Tr}[\Sigma_x]^{\frac{D-1}{2D}}}{\mathrm{Tr}[\Sigma_\epsilon]^{\frac{1}{2D}}}\sqrt{S'},\\
\Sigma_D &= (\mathrm{Tr}S')^{-\frac{D-2}{2D}}\frac{\mathrm{Tr}[\Sigma_\epsilon]^{\frac{D-1}{2D}}}{\mathrm{Tr}[\Sigma_x]^{\frac{1}{2D}}}\sqrt{S'},\\
\Sigma_i &= (\mathrm{Tr}S')^{1/D}(\mathrm{Tr}[\Sigma_\epsilon]\mathrm{Tr}[\Sigma_x])^{-\frac{1}{2D}}I_d.
\end{aligned} \tag{66}$$

A.4 初始尖锐度

为了理解训练开始时的优化损失景观，我们分析标准随机初始化方案下的期望尖锐度。

命题 1. 考虑一个 D 层线性网络，其中每个权重矩阵 $W_k \in \mathbb{R}^{d_{k+1}\times d_k}$ 都使用独立同分布条目进行初始化 $w_i^{(k)}{}_{,\ j} \sim \mathcal{N}(0,\ \sigma_k^2)$，其中 $d_1 = d_x$ 和 $d_{D+1} = d_y$。在初始化时，期望尖锐度为：

$$\mathbb{E}[T(\theta)] = d_y\Pi_{k=2}^D d_k \mathrm{Tr}[\Sigma_x]\sum_{k=1}^D \Pi_{j\neq k}\sigma_j^2. \tag{67}$$

如果我们进一步假设对所有 k 都有 $\sigma_k^2 = \sigma^2$ 和 $d_k = d$，则可简化为：

$$\mathbb{E}[T(\theta)] = D\cdot d_y\mathrm{Tr}[\Sigma_x](d\sigma^2)^{D-1}. \tag{68}$$

证明。回顾一下，$T(\theta) = \sum_{k=1}^{D} \mathrm{Tr}[A_k^T A_k]\mathrm{Tr}[B_k \Sigma_x B_k^T]$。由于权重矩阵 $\{W_k\}$ 是独立初始化的，迹乘积的期望可分解为各期望的乘积。对于 A_k，我们有

$$\mathbb{E}[\mathrm{Tr}[A_k^T A_k]] = \mathbb{E}[\mathrm{Tr}[W_{k+1}^T \cdots W_D W_D^T \cdots W_{k+1}]] = d_y \Pi_{j=k+1}^{D} d_j \sigma_j^2. \tag{69}$$

同理，对于 B_k:

$$\mathbb{E}[\mathrm{Tr}[B_k \Sigma_x B_k^T]] = \mathbb{E}[\mathrm{Tr}[W_{k-1} \cdots W_1 \Sigma_x W_1^T \cdots W_{k-1}]] = \mathrm{Tr}[\Sigma_x] \Pi_{j=1}^{k-1} d_{j+1} \sigma_j^2. \tag{70}$$

将这些结果代入总尖锐度表达式（44）可得（67）。 □

对于 Xavier 初始化，我们有 $\sigma_k^2 = \frac{2}{d_k + d_{k+1}}$。令所有 k 均取 $d_k = d$，则有 $\mathbb{E}[T(\theta)] = D d_y \mathrm{Tr}[\Sigma_x]$。对于采用线性激活的 Kaiming 初始化，我们有 $\sigma^2 = \frac{1}{d_k}$。那么，对于任意 k，只要 $d_k = d$，尖锐度就相同。

A.5 最大特征值

以下定理刻画了最优解处海森矩阵的谱半径。

定理 3. 假设 $\Sigma_x = \hat{R}^{-1} \Lambda_x \hat{R}$ 是某个对角矩阵 Λ_x，其中 $\hat{R}$ 是 $V\sqrt{\Sigma_x}$的右奇异矩阵。则，解 (52) 对应的海森矩阵的最大特征值为

$$\begin{aligned}\lambda_{\max}(\mathbb{E}[H(\theta)]) = (D-2)(\mathrm{Tr}\hat{S})^{-\frac{2}{D}} (d_y \mathrm{Tr}\Sigma_x)^{\frac{1}{D}} \lambda_{\max}(\hat{S})^2 + (\mathrm{Tr}\hat{S})^{\frac{D-2}{D}} d_y^{-\frac{D-1}{D}} (\mathrm{Tr}\Sigma_x)^{\frac{1}{D}} \lambda_{\max}(\hat{S}) \\ + (\mathrm{Tr}\hat{S})^{\frac{D-2}{D}} d_y^{\frac{1}{D}} (\mathrm{Tr}\Sigma_x)^{-\frac{D-1}{D}} \lambda_{\max}(\hat{S}\Lambda_x^{-1}) \lambda_{\max}(\Sigma_x)\end{aligned} \tag{71}$$

此外，假设 $\Sigma_x = R^{-1} \Lambda_x R$ 且 $\Sigma_\epsilon = L \Lambda_\epsilon L^{-1}$ 对于某些对角矩阵 Λ_x，Λ_ϵ其中 L，R是 $\sqrt{\Sigma_\epsilon} V \sqrt{\Sigma_x}$的左、右奇异矩阵。那么，解(66)对应的海森矩阵的最大特征值为

$$\begin{aligned}\lambda_{\max}(\mathbb{E}[H(\theta)]) = (D-2)(\mathrm{Tr}S)^{-\frac{2}{D}} (\mathrm{Tr}\Sigma_\epsilon \mathrm{Tr}\Sigma_x)^{\frac{1}{D}} \lambda_{\max}(\Lambda_\epsilon^{-1} S) \lambda_{\max}(S) + (\mathrm{Tr}S)^{\frac{D-2}{D}} (\mathrm{Tr}\Sigma_\epsilon)^{-\frac{D-1}{D}} (\mathrm{Tr}\Sigma_x)^{\frac{1}{D}} \lambda_{\max}(S) \\ + (\mathrm{Tr}S)^{\frac{D-2}{D}} (\mathrm{Tr}\Sigma_\epsilon)^{\frac{1}{D}} (\mathrm{Tr}\Sigma_x)^{-\frac{D-1}{D}} \lambda_{\max}(\Lambda_\epsilon^{-1} S \Lambda_x^{-1}) \lambda_{\max}(\Sigma_x)\end{aligned} \tag{72}$$

证明。步骤 1：海森矩阵的二次型。

根据（24），对于深度线性网络 $f(\theta, x) = W_D \cdots W_1 x$，海森矩阵为

$$H(\theta) = \mathbb{E}_x \nabla f(\theta, x) \nabla f(\theta, x)^T. \tag{73}$$

令 $\Delta = \mathrm{vec}(\Delta_D, \cdots, \Delta_1)$ 为扰动，其中 Δ_i 与 W_i维度相同，海森矩阵的二次型为

$$\Delta^T H(\theta) \Delta = \mathbb{E}_x \left[\left\| \sum_{k=1}^{D} A_k \Delta_k B_k x \right\|_2^2 \right] = \left\| \sum_{k=1}^{D} A_k \Delta_k B_k \sqrt{\Sigma_x} \right\|_F^2, \tag{74}$$

其中 $A_k := W_D \cdots W_{k+1}$ 和 $B_k := W_{k-1} \cdots W_1$，特别是 $A_D = I_{d_y}$、$B_1 = I$。

步骤 2：上界。

利用三角不等式和谱范数的次乘性，我们可以对扰动之和的大小进行界定：

$$\sqrt{\Delta^T H(\theta) \Delta} \le \sum_{k=1}^{D} \|A_k\|_2 \|B_k \sqrt{\Sigma_x}\|_2 \|\Delta_k\|_F. \tag{75}$$

对求和应用柯西-施瓦茨不等式，我们得到：

$$\frac{\Delta^T H(\theta) \Delta}{\Delta^T \Delta} \le \frac{\left(\sum_{k=1}^{D} \|A_k\|_2 \|B_k \sqrt{\Sigma_x}\|_2 \|\Delta_k\|_F \right)^2}{\sum_{k=1}^{D} \|\Delta_k\|_F^2} \le \sum_{k=1}^{D} \|A_k\|_2^2 \|B_k \sqrt{\Sigma_x}\|_2^2. \tag{76}$$

这给出

$$\begin{aligned}\lambda_{\max}(\mathbb{E}[H(\theta)]) &\le \sum_{k=1}^{D} \|W_D \dots W_{k+1}\|_2^2 \cdot \|W_{k-1} \dots W_1 \sqrt{\Sigma_z}\|_2^2 \\ &= \sum_{k=1}^{D} \|\hat{L}\Sigma_D \cdots \Sigma_{k+1} U_{k+1}\|_2^2 \|U_{k-1}\Sigma_{k-1} \cdots \Sigma_1 \hat{R}\|_2^2 \\ &= (D-2)(\mathrm{Tr}\hat{S})^{-2/D}(d_y \mathrm{Tr}\Sigma_x)^{\frac{1}{D}} \lambda_{\max}(\hat{S})^2 + (\mathrm{Tr}\hat{S})^{\frac{D-2}{D}} d_y^{-\frac{D-1}{D}} (\mathrm{Tr}\Sigma_x)^{\frac{1}{D}} \lambda_{\max}(\hat{S}) \\ &\quad + (\mathrm{Tr}\hat{S})^{\frac{D-2}{D}} d_y^{\frac{1}{D}} (\mathrm{Tr}\Sigma_x)^{-\frac{D-1}{D}} \lambda_{\max}(\hat{S}\Lambda_x^{-1}) \lambda_{\max}(\Sigma_x).\end{aligned} \tag{77}$$

步骤 3：紧性与构造。

为证明上界可达，我们构造一个特定的扰动 Δ。令 u_D 和 u_0分别为全网络乘积的最大左、右奇异向量。对于解（52），我们选择秩一扰动：$\Delta_k = \alpha_k u_i u_{i-1}^T$，其中 u_k 是 U_k 中对应于 $\hat{S}$（或$\hat{S}\Lambda_x^{-1/2}$，对于 $k=0$）最大奇异向量的列，并特别地 $U_0 := \hat{R}$，$U_D := \hat{L}$。因此，我们有

$$A_k \Delta_k B_k \sqrt{\Sigma_x} = \alpha_k ||A_k||_2 ||B_k \sqrt{\Sigma_x}||_2 u_D u_0^T, \tag{78}$$

其中我们用到了这一事实：

$$W_1 = U_1 \Sigma_1 \hat{R} \Sigma_x^{-1/2} = U_1 \Sigma_1 \Lambda_x^{-1/2} \hat{R} \tag{79}$$

在假设 $\Sigma_x = \hat{R}^{-1}\Lambda_x \hat{R}$ 下。于是，我们有

$$\Delta^T H(\theta) \Delta = \left\| \sum_{k=1}^{D} \alpha_k ||A_k||_2 ||B_k \sqrt{\Sigma_x}||_2 u_D u_0^T \right\|_F^2 = \left(\sum_{k=1}^{D} \alpha_k ||A_k||_2 ||B_k \sqrt{\Sigma_x}||_2 \right)^2, \tag{80}$$

通过选择 $\alpha_k = ||A_k||_2 ||B_k \sqrt{\Sigma_x}||_2$，可得

$$\frac{\Delta^T H(\theta) \Delta}{\Delta^T \Delta} = \sum_{k=1}^{D} ||A_k||_2^2 ||B_k \sqrt{\Sigma_x}||_2^2. \tag{81}$$

这证明了（71）。

对于解（66），我们有

$$W_D = \Sigma_\epsilon^{-1/2} L \Sigma_D U_{D-1}^T = L \Lambda_\epsilon^{-1/2} \Sigma_D U_{D-1}^T \tag{82}$$

在假设 $\Sigma_\epsilon = L \Lambda_\epsilon L^{-1}$ 的条件下。然后我们可以用同样的方法证明（72）。 □

注。 由 （74）可得

$$\Delta^T H(\theta) \Delta = \left\| \left[\sum_{k=1}^{D} (B_k \sqrt{\Sigma_x})^T \otimes A_k \right] vec(\Delta_k) \right\|_2^2, \tag{83}$$

这意味着

$$\lambda_{\max}(H(\theta)) = \sigma_{\max} \left(\sum_{k=1}^{D} (B_k \sqrt{\Sigma_x})^T B_k \sqrt{\Sigma_x} \otimes A_k A_k^T \right), \tag{84}$$

这恢复了同期工作的定理*3.1* *[KP25]*。然而，除非 $B_k \sqrt{\Sigma_x}$ 和 Σ_x 具有相同的奇异空间，(84) 在解 (52)下无法进一步简化。因此，我们在定理*3*中引入交换性假设。

A.6 推论1的证明

证明。我们首先运用冯·诺伊曼迹不等式，推导出三个有用的界：

$$\lambda_{\min}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_* \le \mathrm{Tr}S \le \lambda_{\max}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_*, \tag{85}$$

$$\mathrm{Tr}[\Sigma_\epsilon^{-1}LSL^T] \ge \frac{\sigma_{\min}(S)}{\lambda_{\min}(\Sigma_\epsilon)} \ge \frac{\sigma_{\min}(V\sqrt{\Sigma_x})}{\sqrt{\lambda_{\min}(\Sigma_\epsilon)}}, \tag{86}$$

and

$$\mathrm{Tr}[\Sigma_\epsilon^{-1}LSL^T] \le \frac{\|S\|_*}{\lambda_{\min}(\Sigma_\epsilon)} \le \frac{\lambda_{\max}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_*}{\lambda_{\min}(\Sigma_\epsilon)}, \tag{87}$$

其中， $\|\cdot\|_*$ 表示核范数。

为了求出 $T(\theta)$的下界，我们利用尖锐度表达式的第一项。应用（85）和（86），得到：

$$\begin{aligned}(\mathrm{Tr}S)^{(D-2)/D}(\mathrm{Tr}\Sigma_\epsilon)^{1/D}\mathrm{Tr}[\Sigma_\epsilon^{-1}LSL^T] &\ge \left(\lambda_{\min}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_*\right)^{(D-2)/D}(\mathrm{Tr}\Sigma_\epsilon)^{1/D}\frac{\sigma_{\min}(V\sqrt{\Sigma_x})}{\sqrt{\lambda_{\min}(\Sigma_\epsilon)}}\\ &\ge \max(\sigma_{\min}(V\sqrt{\Sigma_x})^2,1)\frac{\lambda_{\max}(\Sigma_\epsilon)^{\frac{D-2}{2D}+\frac{1}{D}}}{\sqrt{\lambda_{\min}(\Sigma_\epsilon)}}\\ &:= 2c_1\sqrt{\kappa(\Sigma_\epsilon)},\end{aligned} \tag{88}$$

其中， c_1 是一个常数，仅通过 $V\sqrt{\Sigma_x}$依赖于数据和模型。由（14）可知，总尖锐度的下界为条件数的平方根：

$$T(\theta) \ge 2(D-1)c_1\sqrt{\kappa(\Sigma_\epsilon)} \ge Dc_1\sqrt{\kappa(\Sigma_\epsilon)}. \tag{89}$$

对于上界，我们同时考虑尖锐度的两项。首先，利用（85）和（87），我们对主要项进行上界估计：

$$\begin{aligned}&(\mathrm{Tr}S)^{(D-2)/D}(\mathrm{Tr}\Sigma_\epsilon)^{1/D}\mathrm{Tr}[\Sigma_\epsilon^{-1}LSL^T]\\ &\le (\lambda_{\max}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_*)^{(D-2)/D}(d_y\lambda_{\max}(\sqrt{\Sigma_\epsilon}))^{1/D}\frac{\lambda_{\max}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_*}{\lambda_{\min}(\Sigma_\epsilon)}\\ &\le \max(d_y^{1/D}\|V\sqrt{\Sigma_x}\|_*^2,1)\frac{\lambda_{\max}(\Sigma_\epsilon)}{\lambda_{\min}(\Sigma_\epsilon)}.\end{aligned} \tag{90}$$

接下来，我们利用（ T（θ）中的（85）来界定

$$\begin{aligned}d_y(\mathrm{Tr}S)^{2(D-1)/D}\frac{(\mathrm{Tr}\Sigma_x)^{1/D}}{(\mathrm{Tr}\Sigma_\epsilon)^{(D-1)/D}} &\le d_y(\mathrm{Tr}\Sigma_x)^{1/D}\frac{(\lambda_{\max}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_*)^{2(D-1)/D}}{(d_y\lambda_{\min}(\Sigma_\epsilon))^{(D-1)/D}}\\ &\le d_yd_x\max(\|V\sqrt{\Sigma_x}\|_*^2,1)\frac{\lambda_{\max}(\Sigma_\epsilon)}{\lambda_{\min}(\Sigma_\epsilon)}.\end{aligned} \tag{91}$$

将这两个上界代入（14）后，我们得到最终的线性关系：

$$T(\theta) \le (D-1)\max(d_y\|V\sqrt{\Sigma_x}\|_*^2,1)\frac{\lambda_{\max}(\Sigma_\epsilon)}{\lambda_{\min}(\Sigma_\epsilon)} + d_yd_x\max(\|V\sqrt{\Sigma_x}\|_*^2,1)\frac{\lambda_{\max}(\Sigma_\epsilon)}{\lambda_{\min}(\Sigma_\epsilon)} \le Dc_2\kappa(\Sigma_\epsilon), \tag{92}$$

其中 c_2 为 由 $V\sqrt{\Sigma_x}$ 和维度 d_x、 d_y 确定的常数。证明完毕。 □

A.7 推论2的证明

证明。 该证明直接比较引理中推导出的最小锐度条件4 与引理中刻画的随机梯度下降解即可得出5。令 $\Sigma_\epsilon = I$ 代入（66），各层的矩阵平衡和奇异值分布便与最小锐度（52）的要求完全吻合。

□

作为合理性检验 于是，当我们将 $\Sigma_\epsilon = I$ 代入（14），得到 $V\prime = V\sqrt{\Sigma_x} = LSL^T$, $S = \hat{}S$，因此 s

$$T(\theta) = D(\mathrm{Tr}S)^{2(D-1)/D}(d_y \mathrm{Tr}\Sigma_x)^{1/D}, \tag{93}$$

这与最小锐度相同。

A.8 推论3 的证明

证明。 在交换假设下，协方差矩阵 Σ_ϵ， Σ_x 的特征向量与模型 V 的奇异向量对齐。这使得谱分析得以简化：

$$\Sigma_\epsilon = L\Lambda_\epsilon L^{-1},\ \Sigma_x = R^{-1}\Lambda_x R,\ \sqrt{\Sigma_\epsilon}V\sqrt{\Sigma_x} = LSR. \tag{94}$$

因此 ，我们可以界定乘积矩阵 S 及其缩放变体的谱半径 s:

$$\lambda_{\max}(\sqrt{\Sigma_\epsilon})\sigma_{\min}(V\sqrt{\Sigma_x}) \le \lambda_{\max}(S) \le \lambda_{\max}(\sqrt{\Sigma_\epsilon})\sigma_{\max}(V\sqrt{\Sigma_x}) \tag{95}$$

and

$$\frac{\lambda_{\min}(V\sqrt{\Sigma_x})}{\lambda_{\min}(\sqrt{\Sigma_\epsilon})} \le \lambda_{\max}(\Lambda_\epsilon^{-1}S) \le \frac{\lambda_{\max}(V\sqrt{\Sigma_x})}{\lambda_{\min}(\sqrt{\Sigma_\epsilon})}. \tag{96}$$

对于上界，我们使用迹不等式（72）分别计算（85）中的三项。对于第一项：

$$\begin{aligned}
&(\mathrm{Tr}S)^{-\frac{2}{D}}(\mathrm{Tr}\Sigma_\epsilon \mathrm{Tr}\Sigma_x)^{\frac{1}{D}}\lambda_{\max}(\Lambda_\epsilon^{-1}S)\lambda_{\max}(S) \\
&\le (\lambda_{\min}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_*)^{-\frac{2}{D}}(d_y\lambda_{\max}(\Sigma_\epsilon)\mathrm{Tr}\Sigma_x)^{\frac{1}{D}}\frac{\sigma_{\min}(V\sqrt{\Sigma_x})}{\lambda_{\min}(\sqrt{\Sigma_\epsilon})}\lambda_{\max}(\sqrt{\Sigma_\epsilon})\sigma_{\min}(V\sqrt{\Sigma_x}) \\
&\le \max(\sigma_{\min}(V\sqrt{\Sigma_x})^2, 1)\max(d_y\mathrm{Tr}\Sigma_x, 1)\kappa(\Sigma_\epsilon).
\end{aligned} \tag{97}$$

对于第二项：

$$\begin{aligned}
&(\mathrm{Tr}S)^{\frac{D-2}{D}}(\mathrm{Tr}\Sigma_\epsilon)^{-\frac{D-1}{D}}(\mathrm{Tr}\Sigma_x)^{\frac{1}{D}}\lambda_{\max}(S) \\
&\le (\lambda_{\max}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_*)^{\frac{D-2}{D}}\lambda_{\min}(\Sigma_\epsilon)^{-\frac{D-1}{D}}(\mathrm{Tr}\Sigma_x)^{\frac{1}{D}}\lambda_{\max}(\sqrt{\Sigma_\epsilon})\sigma_{\max}(V\sqrt{\Sigma_x}) \\
&\le \max(\|V\sqrt{\Sigma_x}\|_*^2, 1)\max(\mathrm{Tr}\Sigma_x, 1)\kappa(\Sigma_\epsilon).
\end{aligned} \tag{98}$$

对于第三项：

$$\begin{aligned}
&(\mathrm{Tr}S)^{\frac{D-2}{D}}(\mathrm{Tr}\Sigma_\epsilon)^{\frac{1}{D}}(\mathrm{Tr}\Sigma_x)^{-\frac{D-1}{D}}\lambda_{\max}(\Lambda_\epsilon^{-1}S\Lambda_x^{-1})\lambda_{\max}(\Sigma_x) \\
&\le (\lambda_{\max}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_*)^{\frac{D-2}{D}}\lambda_{\max}(\Sigma_\epsilon)^{\frac{1}{D}}(\mathrm{Tr}\Sigma_x)^{-\frac{D-1}{D}}\frac{\sigma_{\max}(V\sqrt{\Sigma_x})}{\lambda_{\min}(\sqrt{\Sigma_\epsilon})\lambda_{\min}(\Sigma_x)}\lambda_{\max}(\Sigma_x) \\
&\le \max(\|V\Sigma_x\|_*^2, 1)\max(\lambda_{\max}(\Sigma_x)^{-1}, 1)\kappa(\Sigma_x)\kappa(\Sigma_\epsilon).
\end{aligned} \tag{99}$$

将这些贡献相加，我们得出海森矩阵的谱半径满足上界

$$\lambda_{\max}(\mathbb{E}[H(\theta)]) \le c_2 D\kappa(\Sigma_\epsilon) \tag{100}$$

其中 c_2 是由问题几何结构决定的常数（V， Σ_x， d_y）。

我们以相同的方式推导下界。对于第一项：

$$\begin{aligned}
&(\mathrm{Tr}S)^{-\frac{2}{D}}(\mathrm{Tr}\Sigma_\epsilon \mathrm{Tr}\Sigma_x)^{\frac{1}{D}}\lambda_{\max}(\Lambda_\epsilon^{-1}S)\lambda_{\max}(S) \\
&\ge (\lambda_{\max}(\sqrt{\Sigma_\epsilon})\|V\sqrt{\Sigma_x}\|_*)^{-\frac{2}{D}}(\lambda_{\max}(\Sigma_\epsilon)\mathrm{Tr}\Sigma_x)^{\frac{1}{D}}\frac{\sigma_{\min}(V\sqrt{\Sigma_x})}{\lambda_{\min}(\sqrt{\Sigma_\epsilon})}\lambda_{\max}(\sqrt{\Sigma_\epsilon})\sigma_{\min}(V\sqrt{\Sigma_x}) \\
&\ge \min(\sigma_{\min}(V\sqrt{\Sigma_x})^2, 1)\min(\mathrm{Tr}\Sigma_x, 1)\sqrt{\kappa(\Sigma_\epsilon)}.
\end{aligned} \tag{101}$$

对于第二项：

$$\begin{aligned}
&(\mathrm{Tr}S)^{\frac{D-2}{D}}(\mathrm{Tr}\Sigma_\epsilon)^{-\frac{D-1}{D}}(\mathrm{Tr}\Sigma_x)^{\frac{1}{D}}\lambda_{\max}(S)\\
&\geq(\lambda_{\max}(\sqrt{\Sigma_\epsilon})\sigma_{\min}(V\sqrt{\Sigma_x}))^{\frac{D-2}{D}}\lambda_{\max}(\Sigma_\epsilon)^{-\frac{D-1}{D}}(\mathrm{Tr}\Sigma_x)^{\frac{1}{D}}\lambda_{\max}(\sqrt{\Sigma_\epsilon})\sigma_{\min}(V\sqrt{\Sigma_x})\\
&\geq\min(\sigma_{\min}(V\sqrt{\Sigma_x})^2,1)\min(\mathrm{Tr}\Sigma_x,1)\sqrt{\kappa(\Sigma_\epsilon)}.
\end{aligned}\tag{102}$$

最后，对于第三项：

$$\begin{aligned}
&(\mathrm{Tr}S)^{\frac{D-2}{D}}(\mathrm{Tr}\Sigma_\epsilon)^{\frac{1}{D}}(\mathrm{Tr}\Sigma_x)^{-\frac{D-1}{D}}\lambda_{\max}(\Lambda_\epsilon^{-1}S\Lambda_x^{-1})\lambda_{\max}(\Sigma_x)\\
&\geq(\lambda_{\max}(\sqrt{\Sigma_\epsilon})\sigma_{\min}(V\sqrt{\Sigma_x}))^{\frac{D-2}{D}}\lambda_{\max}(\Sigma_\epsilon)^{\frac{1}{D}}(\mathrm{Tr}\Sigma_x)^{-\frac{D-1}{D}}\frac{\sigma_{\min}(V\sqrt{\Sigma_x})}{\lambda_{\min}(\sqrt{\Sigma_\epsilon})\lambda_{\min}(\Sigma_x)}\lambda_{\max}(\Sigma_x)\\
&\geq\min(\sigma_{\min}(V\Sigma_x)^2,1)\min(d_y^{-1}\lambda_{\max}(\Sigma_x)^{-1},1)\kappa(\Sigma_x)\sqrt{\kappa(\Sigma_\epsilon)}.
\end{aligned}\tag{103}$$

将这些下界代入（72）可得：

$$\lambda_{\max}(\mathbb{E}[H(\theta)])\geq c_1D\sqrt{\kappa(\Sigma_\epsilon)}\tag{104}$$

其中某个常数 c_1 取决于 V、Σ_x 和 d_y。证明至此完成。 □

A.9 推论4的证明

证明。将$\mathrm{Tr}\Sigma_\epsilon$、$\mathrm{Tr}\Sigma_x$、$\mathrm{Tr}S=\Theta(d^{\max(0,1-a)})$ $\Theta(d^{\max(0,1-b)})$ $\Theta(d^{\max(0,1-a/2-b/2-c)})$ 代入（14）即可证明。 □

A.10 示例：幂律谱

推论4。假设 Σ_ϵ、Σ_x、V具有幂律谱 $\{i^{-a}\}_{i=1}^d$、$\{i^{-b}\}_{i=1}^d$、$\{i^{-c}\}_{i=1}^d$，对于 a、b、$c>0$成立，并且它们彼此对易。那么对于足够大的 d，我们有

$$T(\theta)=\Theta\left((D-1)d^{\kappa_1}+d^{\kappa_2}\right)\tag{105}$$

在 (12) 的全局最小值处，其中

$$\begin{aligned}
\kappa_1:=&\frac{D-2}{D}\max\left(0,1-\frac{a}{2}-\frac{b}{2}-c\right)+\frac{1}{D}\max(0,1-a)\\
&+\frac{1}{D}\max(0,1-b)+\max\left(0,1+\frac{a}{2}-\frac{b}{2}-c\right)
\end{aligned}\tag{106}$$

and

$$\begin{aligned}
\kappa_2:=&1+\frac{2(D-1)}{D}\max(0,1-\frac{a}{2}-\frac{b}{2}-c)\\
&+\frac{1}{D}\max(0,1-b)-\frac{D-1}{D}\max(0,1-b).
\end{aligned}\tag{107}$$

推论4 进一步证实，噪声不平衡是导致锐化的主要原因，因为只有当 a 增大时，尖锐度才会增加；当 b、c 增大时则不会。

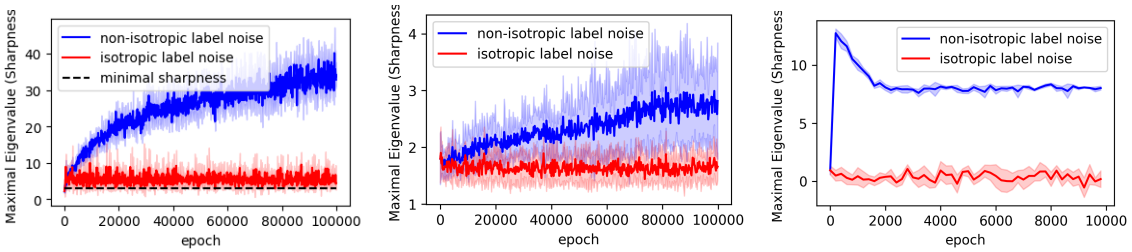

图 5： 海森矩阵的最大特征值演变趋势与迹相似。**左图**：线性网络，其最小锐度由式（71）预测。**中图**：在师生设定下训练的两层 ReLU 网络。**右图**：在 MNIST 上训练的两层 ReLU 网络。每条曲线均为五次试验的平均值，并显示标准误差。

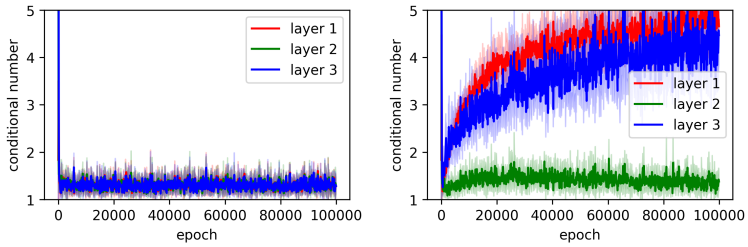

图 6： 三层线性网络的条件数演变。**左图**：标签中的平衡噪声。**右图**：标签中的不平衡噪声。

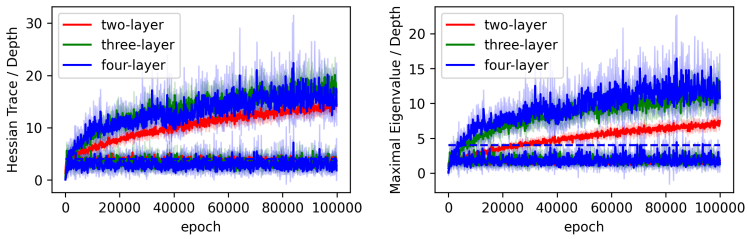

图 7： 不同深度的线性网络的锐度演变。**左图**：海森矩阵的迹。**右图**：海森矩阵的最大特征值。

B 实验

线性网络实验 对于线性网络设置，我们令 $D = 3$、 $d_x = d_y = 2$和$V = I$，数据取自标准高斯分布（$\Sigma_x = I$）。非各向同性情形下，噪声尺度设为$\sqrt{\Sigma_\epsilon}$ =diag(10, 0.1)；各向同性情形下，则设为 $\sqrt{\Sigma_\epsilon}$ =diag(10, 10)。使用在线SGD训练 10^5 个训练轮次（学习率为0.01，批大小为10）。通过200个随机样本估计海森矩阵，其中迹精确计算，最大特征值则使用幂法计算。对于图1和图4中的实验，我们保持这些超参数不变，但调整初始化强度和学习率（设为0.001），或改用单批次梯度下降（GD）设置（批大小为1000）。

为补充图 2中的观察结果，我们评估了海森矩阵的最大特征值，如图 5所示。所得曲线呈现出一致的趋势，进一步佐证了推论3。

图6 展示了三层线性网络各层的条件数。与我们的解析解（66）一致，我们观察到，标签中的平衡噪声会使所有层的条件数都较小。相反，不平衡噪声主要增大第一层和最后一层的条件数，而中间层保持稳定。此外，图7表明，深度仅作为尖锐度的乘性因子，正如推论1所示。

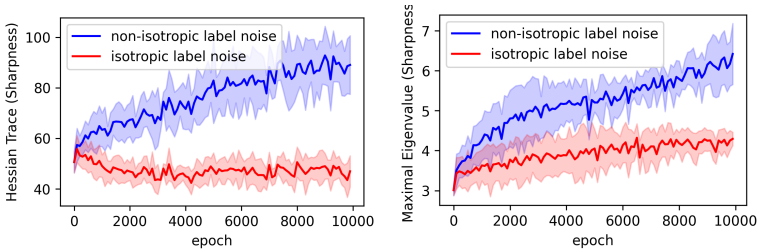

图 8： 非各向同性噪声也会导致 Transformer 渐进锐化。**左**：海森矩阵迹的演化。**右**：最大特征值的演化。

非线性设定下的实验在ReLU师生设定中（见图2 和5），我们采用具有 $d_x = d_y = 2$的单隐藏层师生框架。教师网络和学生网络分别包含100 个和4 个隐藏单元，均使用Kaiming方法初始化。我们使用均方误差损失和在线SGD（学习率0.1，批大小100）进行 10^5 个训练轮次。噪声尺度设置为diag(3,0.03)和diag(3,3)。Hessian统计量使用200 个样本计算，其中迹由Hutchinson估计器近似。

对于 MNIST 数据集（见图 2 和 5），使用交叉熵损失训练两层ReLU网络（128 个隐藏单元）。通过向独热标签添加高斯噪声，再执行 softmax 操作，引入标签噪声。非各向同性噪声协方差 $\sqrt{\Sigma_\epsilon}$ 设置为 diag$(10,\ldots,10,1,\ldots,1)$（每个数值各出现五次），而各向同性情形使用 $\sqrt{\Sigma_\epsilon} = 10I$。我们使用梯度下降训练 10^4 个训练轮次（学习率为 0。01，批大小为 128），并在批大小为 1024的测试批次上评估海森矩阵。

为验证我们的发现是否超越线性网络和多层感知机具有普适性，我们在教师-学生框架下，将实验扩展到 Transformer 和循环神经网络。在图 3 和 8中，我们采用了一个简化的 Transformer，其由一个双头自注意力层和一个带残差连接的多层感知机组成。模型输入维度为 8，序列长度为 4，输出维度与输入相同。教师模型的多层感知机隐藏维度设为 100 ，学生模型设为 4 ，二者均使用 Kaiming方法初始化。

使用在线SGD（学习率0。03，批大小为300）进行训练，共 10^4 个训练轮次，使用标准高斯数据。海森矩阵在大小为1000的测试集上计算。我们将噪声尺度配置为$\sqrt{\Sigma_\epsilon} = \mathrm{diag}(4,0.04,\ldots,0.04)$，用于非各向同性情形；并配置为 $\sqrt{\Sigma_\epsilon} = \mathrm{diag}(4,\ldots,4)$，用于各向同性情形。图8中的结果证实，标签中的非各向同性噪声会在 Transformer 中引发与较简单模型中观察到的相同的渐进锐化现象。

此外，我们在图3中研究噪声协方差条件数对尖锐度的具体影响。在保持实验设置不变的情况下，我们系统地改变Σ_ϵ 中较小的特征值，同时固定其他超参数。我们还将此分析扩展到循环神经网络，其中教师和学生的隐藏单元分别设为100 和4。对于涉及交叉熵损失的实验，我们在网络输出端额外应用一个 softmax 层。在所有这些多样化的设置中——包括不同的架构和损失函数——观察到的趋势都高度一致，进一步验证了理论预测的稳健性。