arXiv:2602.02986v1 [cs.LG] 3 Feb 2026

Why Some Models Resist Unlearning: A Linear Stability Perspective

Wei-Kai Chang普渡大学
chang986@purdue.edu

Rajiv Khanna普渡大学
rajivak@purdue.edu

Abstract

机器遗忘——无需从头重新训练即可消除特定训练样本影响的能力——对隐私、监管和效率至关重要。然而，机器遗忘领域的大多数进展都基于经验，对机器遗忘在何时以及为何有效，理论理解仍然有限。我们从渐近线性稳定性的视角来探讨这一问题，以刻画优化动力学与数据几何之间的相互作用。分析中的关键量是数据相干性，即最优点附近损失曲面方向的跨样本对齐。我们沿三个维度分解相干性：保留集内部、遗忘集内部，以及两者之间，并证明了精确的稳定性阈值，可用于区分收敛与发散。为进一步建立数据属性与可遗忘性之间的联系，我们研究了信号加噪声模型下的双层 ReLU 卷积神经网络，并表明更强的记忆会使遗忘更容易：当信噪比（SNR）较低时，跨样本对齐较弱，数据相干性随之降低，机器遗忘也更容易；相反，高信噪比且高度对齐的模型则难以遗忘。在实证验证方面，我们表明海森矩阵检验和 CNN 热力图与预测边界高度吻合，并根据批处理、混合以及数据/模型对齐情况，描绘了基于梯度的机器遗忘的稳定性前沿。我们的分析以随机矩阵理论工具为基础，首次对记忆、相干性与机器遗忘之间的权衡作出了有原则的阐释。

1 Introduction

机器遗忘——从模型中消除特定训练样本影响的能力——对于合规、隐私和模型维护至关重要。在实践中，移除一个样本后，从头开始重新训练一个含有N个样本的模型，成本高得令人难以承受，这推动了近似机器遗忘方法的蓬勃发展。（Shen 等人，2024b;a；Hatua 等人，2024；Bourtoule 等人，2020；Cao 和 Yang，2015；Golatkar 等人，2020；Ginart 等人，2019；Golatkar 等人，2021；Graves 等人，2020；Sekhari 等人，2021）。然而，尽管这一领域进展迅速，大多数机器遗忘研究仍停留在经验性和临时性层面。我们缺乏一个统一的理论框架，来预测何时以及为何能够高效地对特定模型进行机器遗忘。值得注意的是，即使是早期的理论研究（例如，从差分隐私的视角提供删除保证的研究（Sekhari 等人，2021；Chien 等人，2025）），也无法解释遗忘的动力学，或遗忘集与保留集之间的相互作用。这一空白促使我们开展本研究：我们旨在以模型损失景观的几何结构为基础，系统理解机器遗忘的优化动力学。

我们的方法。我们从渐近线性稳定性的角度研究机器遗忘，分析预训练极小值附近的扰动动力学，而不是采用标准的微调视角（Ding 等人，2025）。不同于从随机初始化开始的训练，机器遗忘始于局部最优点附近；我们刻画遗忘更新是保持受限（稳定，机器遗忘失败），还是引发发散（不稳定，机器遗忘成功）。在先前关于极小值附近随机梯度下降稳定性的分析基础上（Ma 和 Ying，2021；Dexter 等人，2024），该框架为收敛与灾难性遗忘之间的转变提供了一个易于处理的判据。我们的分析由数据相干性所支配，它衡量的是

最优点附近的跨样本梯度对齐。我们将相干性分解为三个组成部分：保留集内部、遗忘集内部，以及保留集与遗忘集之间的相干性，并根据这些量推导稳定性阈值。这揭示了数据几何与机器遗忘之间的直接联系：较高的保留–遗忘相干性会导致遗忘更新损害保留性能，而较低的集间相干性则允许存在稳定方向，在遗忘目标数据的同时不损害其余数据。综合来看，这些结果在相干空间中定义了一条稳定性前沿，将基于梯度的机器遗忘成功的区域与失败的区域区分开来。

我们的框架还对训练记忆与后续可遗忘性之间的关系给出了一个耐人寻味的洞见。有趣且或许出人意料的是，我们发现更强的记忆反而会使遗忘更容易。在我们的框架中，记忆对应于复杂数据的一种情形，此时模型拟合了特异的细节。我们通过改编先前研究中关于良性过拟合的双层 ReLU 卷积神经网络信号加噪声模型（Kou等人（2023））来形式化这一点。利用随机矩阵理论工具，我们证明：当数据中的信噪比（SNR）较低时（即模型必须记忆更多虚假噪声），梯度的跨样本对齐较弱——从而降低相干性项，使这些样本能够被有效遗忘。相比之下，在高信噪比数据（具有强对齐的主导特征）上训练的模型具有高度相干的梯度，将其推向稳定性前沿的边缘，从而使其抗拒机器遗忘——任何试图遗忘某一个样本影响的尝试都会强烈扰动许多其他样本。我们解析地识别出这一相干控制的稳定性边界，实验结果也证实了这一趋势：例如，关于遗忘与保留影响的Hessian特征值检验和CNN热力图与预测边界一致，并刻画了批大小、遗忘/保留数据的混合或网络对齐的变化如何影响机器遗忘的结果。

贡献 概括而言，我们的贡献如下：（1）我们建立了首个基于线性稳定性分析的机器遗忘理论框架，用以回答何种局部优化动力学支配机器遗忘？。我们推导了标准的基于梯度的机器遗忘在何种精确条件（以Hessian谱和数据相干性刻画）下收敛或发散。（2）我们探讨保留集与遗忘集如何定量地相互作用，从而决定稳定性？ 为此，我们引入新颖的相干性指标来量化保留–遗忘交互，并证明每个相干性分量（保留-保留、遗忘-遗忘、保留-遗忘）如何影响遗忘过程的稳定性。这些结果形式化地刻画了数据几何与数据分布在遗忘动力学中的共同作用。（3）为回答模型的记忆倾向如何与其遗忘能力相互作用？，我们在记忆与可遗忘性之间建立了出人意料的联系：利用具有可控噪声的两层卷积神经网络，我们证明记忆增强（更低信噪比）会扩大稳定遗忘的范围（使遗忘更容易），而高信噪比（更少过拟合）则会缩小该范围（更抗拒遗忘）。据我们所知，这是首个将模型的泛化/记忆性质与其遗忘行为严格联系起来的结果。（4）我们的实证测试在多种条件下测量稳定性指标（例如经由Hessian特征值得到的锐度）与遗忘性能，并显示其与理论稳定性前沿高度吻合。综合而言，我们的结果首次对现代机器学习模型中记忆、数据相干性与机器遗忘之间的权衡给出了有原则的阐释。

2 Related Work

线性稳定性 先前研究利用线性稳定性，通过随机梯度下降噪声与损失景观之间的对齐来刻画收敛—发散边界（Wu 等人，2018年；2022年；Wu与Su，2023）。这些对齐特性与随机梯度下降中的简单性偏差和隐式正则化有关，相关研究还将其扩展至高阶噪声矩（Ma 和 Ying，2021）。近期，Dexter 等人（2024）提出了数据相干性，该指标直接量化跨样本梯度对齐，并细致呈现样本间的相互作用。

与梯度流或动力系统近似相比，线性稳定性明确揭示了模型架构、数据分布与优化动力学之间的联系。然而，机器遗忘需要分析相互耦合的保留–遗忘交互，而这类交互

在经典稳定性设定中并不存在。我们通过将相干性推广到混合保留—遗忘情形，弥补了这一空白，从而为机器遗忘提出了新的稳定性判据，并首次在线性稳定性框架内建立了记忆与遗忘之间的形式化联系。

机器遗忘理论研究。许多奠基性研究通过考察优化轨迹及其相对于原始训练动力学的偏离，分析机器遗忘。例如，Golatkar 等人（2020）在二次损失下对机器遗忘进行建模，并通过比较无限时间极限下的原始权重与遗忘后的权重，刻画优化轨迹的漂移。Ding 等人（2025）通过权重空间距离研究线性模型中的近似机器遗忘，并将微调模型与遗忘后模型之间的损失差作为核心指标。近期研究（Mavrothalassitis 等人，2025）分析了线性逻辑回归，并以原始最优解与遗忘后的解之间权重差的闭式表达来描述遗忘。Thudi 等人（2022）展开随机梯度下降递推关系，通过线性化梯度流近似研究机器遗忘动力学，并直接在权重空间中定义机器遗忘，将其与成员推断脆弱性联系起来。更多讨论请参见附录 A.4 和附录 A.1。我们的工作建立在这一既有理论基础之上，但在概念上有所不同：我们不再通过解之间的距离刻画遗忘，而是提出一种基于优化动力学的稳定性视角。通过分析机器遗忘更新算子的渐近线性稳定性，我们揭示保留集与遗忘集之间的相互作用——由曲率和数据相干性介导——如何决定优化器是保持在原始极小值附近还是偏离该点，进而决定机器遗忘是否可行。

3 Theory

3.1 Background

极小值附近的线性稳定性。 线性稳定性为分析临界点（例如局部极小值或鞍点）附近的迭代优化局部动力学提供了有理有据的视角：通过对更新映射进行线性化，并研究由此得到的线性时变系统，从而刻画随机迭代算法在该临界点处的收敛／发散行为。这一视角构成了随机梯度下降及其变体的现代收敛性分析的基础。近年来，它也被证明能有效刻画与泛化相关的现象，例如从尖锐极小值中快速逃逸（Wu 等人，2018年；Dexter 等人，2024年）。在本文的研究背景下，我们考虑关于模型参数 $w \in \mathbb{R}^d$ 的损失 $L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(w)$。设 w^* 为一个局部极小值。对于 δ 在 w^* 附近的一个小扰动，一阶泰勒展开可得

$$\nabla L(w^* + \delta) \approx \nabla^2 L(w^*)\,\delta,$$

因为 $\nabla L(w^*) = 0$。这一线性化表明，在 w^* 附近，梯度近似等于海森矩阵 $H = \nabla^2 L(w^*)$ 乘以该扰动。由于我们只关注优化器的动力学（而非其绝对位置），不失一般性，我们取 $w^* = 0$ ，与先前研究一致。关于本文所用假设的更多讨论，请参阅附录 A.2。

随机梯度更新。 我们关注随机梯度下降（SGD）在 w^* 附近的动态。一般的 SGD 更新可以写作

$$w_{t+1} = w_t - \eta\, g_t,$$

其中， $\eta > 0$ 是学习率， g_t 是第 t 步的随机梯度。在 w^* 附近，利用线性近似，可写为 $g_t \approx H_t\, w_t$，其中 H_t 是随机海森矩阵（ H 的小批量估计）。因此，更新变为

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

这里， $H_t = \frac{1}{B}\sum_{i\in D_t} H_i$ 表示大小为 B的小批量 D_t 上的平均海森矩阵，且$H_i = \nabla^2 \ell_i(w^*)$。根据定义，$\bar{H} = \frac{1}{n}\sum_{i=1}^{n} H_i$。遵循 Dexter 等人（2024年）的做法，我们通过伯努利抽样对小批量采样进行建模（每个数据点都以 B/n的概率独立地纳入批次）。

逐样本梯度。 在上面的公式中，我们假设w^* 处的逐样本梯度为零，即，

$$\begin{aligned}\nabla l_i(w) &= \nabla l_i(w^*) + H_i(w - w^*). \\ &= H_i(w - w^*).\end{aligned} \tag{2}$$

这意味着所有梯度都由曲率产生。这一假设遵循先前研究中采用的标准线性插值机制，例如（Dexter 等人，2024；Wu与Su，2023），并已通过实证和理论研究得到验证（Tang等人，2023；Chizat与Bach，2020）。在这一机制下，局部动力学由海森矩阵曲率主导，这与我们的研究重点相符。关于本文所采用假设及其局限性的更多讨论，请参阅附录 A.2。

遗忘更新规则。为了分析机器遗忘（即遗忘训练数据中的一个子集，称为遗忘集，同时保留其余数据保留集），我们采用 Kurmanji 等人（2023）提出的更新规则。在这一方案中，每一步都会同时对保留集执行梯度下降，并对遗忘集执行梯度上升。直观而言，这意味着我们采取一个步骤，降低保留数据上的损失，同时提高应被遗忘数据上的损失。许多基于梯度的遗忘算法都可以视为这一方法的变体，只是对这些组成部分采用了不同的权重。我们使用n_f， f_r 分别表示遗忘样本和保留样本的数量。在我们的线性化框架中，带有遗忘重要性超参数 $\alpha \in [0,1]$ 的更新为：

$$w_{k+1} = w_k - \eta\Big[(1-\alpha)\frac{1}{B}\sum_{i\in D_{r,k}} H_i\, w_k \;-\; \alpha\,\frac{1}{B}\sum_{i\in D_{f,k}} H_i\, w_k\Big], \tag{3}$$

其中$D_{r,\ k}$ 和 $D_{f,\ k}$ 分别表示步骤 k中保留集和遗忘集样本的小批量。将其改写为算子形式可得

$$w_{k+1} = J_k\, w_k, \quad J_k \;=\; I \;-\; \eta(1-\alpha)\frac{1}{B}\sum_{i\in D_{r,k}} H_i \;+\; \eta\alpha\,\frac{1}{B}\sum_{i\in D_{f,k}} H_i\,. \tag{4}$$

随机线性算子 J_k 捕捉了步骤 k中保留梯度和遗忘梯度的综合效应。需要注意的是，由于分别从两个集合中抽取小批量， J_k 本身也是随机的。

线性稳定性分析的一个核心问题是，迭代是否会保持在原始最优解附近，还是会发散远离该点。为量化这一点，我们考察经过 k 步后参数的期望平方范数 $\mathbb{E}\|w_k\|^2 = \mathbb{E}[w_k^T w_k]$。从各向同性小扰动$w_0$开始（不失一般性，我们假设 $w_0 \sim \mathcal{N}(\mathbf{0},\ I)$），可以展开 $w_k = J_{k-1}\cdots J_0\, w_0$。由此得到

$$\begin{aligned}\mathbb{E}\|w_k\|^2 \;&=\; \mathbb{E}[w_0^T (J_0^T\cdots J_{k-1}^T J_{k-1}\cdots J_0) w_0] \\ &=\; \mathbb{E}\,\mathrm{Tr}\big(J_{k-1}\cdots J_0 J_0^T\cdots J_{k-1}^T\big)\,,\end{aligned} \tag{5}$$

其中最后一个等式用到了 $\mathbb{E}[w_0 w_0^T] = I$。式 (5) 是我们将用于分析稳定性的关键量：如果 $\mathbb{E}\|w_k\|^2$ 在 $k\to\infty$时保持有界（或衰减），那么机器遗忘过程在 w^*附近是稳定 （收敛）的；而如果 $\mathbb{E}\|w_k\|^2 \to \infty$，该过程则不稳定（发散并逃离 w^*）。有关本文所用假设的更多讨论，请参阅附录 A.2；有关发散与机器遗忘的更多讨论，请参阅附录 A.4。

3.2 Coherence Measures for Unlearning

单数据集随机梯度下降中的相干性。在介绍专为机器遗忘设计的新相干性度量之前，我们先简要回顾 Dexter 等人（2024）针对标准（单数据集）学习提出的原始海森相干性概念。相干性量化训练集中逐样本海森矩阵之间的对齐程度。直观而言，如果所有样本产生的曲率方向都高度对齐，随机梯度下降就会经历较少的“随机性”，更新也更稳定；反之，如果每个样本的损失景观朝向各不相同，优化动力学就会更加不稳定。

定义 3.1（相干性，单数据集（Dexter 等人，2024））。给定由 n 个训练样本对应的半正定（PSD）海森矩阵 $\{H_i : i \in [n]\}$ 构成的集合，相干矩阵 $S \in \mathbb{R}^{n\times n}$ 定义为

$$S_{ij}^{\text{single}} \;=\; \| H_i^{1/2}\, H_j^{1/2} \|_F \,,$$

样本 i 和 j的海森矩阵平方根乘积的弗罗贝尼乌斯范数。相应的相干性度量 为

$$\sigma^{\text{single}} \;=\; \frac{\lambda_{\max}(S^{\text{single}})}{\max_{i\in[n]} \lambda_{\max}(H_i)} \,,$$

即 S^{single} 除以单个样本海森矩阵特征值的最大值后得到的最大特征值。

直观而言， σ^{single} 接近 1 表明所有样本的主曲率方向高度对齐（相干性高）；而较小的 σ^{single} 则表明不同样本间的曲率各异或相互正交。先前研究表明，较高的相干性 σ^{single} 与随机梯度下降的稳定性更高相关。换言之，当不同样本的损失景观“指向”相似方向时，梯度步会相互强化，随机梯度下降也就更不容易偏离最优点。关于相干矩阵计算在实践中的可行性，请参阅附录 A.5。

保留集与遗忘集的相干性。 在机器遗忘场景中，我们有两个互不相交的样本集合：保留集 D_r （大小为 n_r）和遗忘集 D_f （大小为 n_f）。仅了解各集合内部（保留集与遗忘集）的相干性，不足以描述组合上升-下降动态的行为。我们还需要量化这两个集合之间 的相互作用。因此，我们提出一系列定义，将相干性扩展到多集合设定。
首先，我们定义保留集—遗忘集对的海森矩阵的加权组合，将其作为有效的“混合海森矩阵”：

定义 3.2 （混合海森矩阵）。

$$\begin{aligned} D &:= \frac{1}{n_r n_f} \sum_{r\in D_r, f\in D_f} \frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}} + C_f^{\frac{1}{2}}} H_r + \frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}} + C_f^{\frac{1}{2}}} H_f \\ &= \frac{1}{n_r n_f} \sum_{rf} D_{rf}, \end{aligned} \tag{6}$$

这里 $C_r = \eta^2(1-\alpha)^2 \frac{1}{n_r}\left(\frac{1}{B} - \frac{1}{n_r}\right)$， $C_f = \eta^2\alpha^2 \frac{1}{n_f}\left(\frac{1}{B} - \frac{1}{n_f}\right)$， $D_{rf} = \frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}} + C_f^{\frac{1}{2}}} H_r + \frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}} + C_f^{\frac{1}{2}}} H_f$。
常数 C_r 和 C_f反映了保留集海森矩阵与遗忘集海森矩阵对 w_k 的二阶矩动态的相对贡献（这些常数源自后文引理 3.5 中的随机梯度下降噪声分析）。混合海森矩阵 D 汇总了保留集与遗忘集的成对影响；当在更新中将这两个集合共同考虑时，它有效地概括了二者如何共同影响曲率。接下来，类似于单集情形，我们定义一个相干矩阵，用于捕捉保留集/遗忘集样本对 之间的对齐：

定义 3.3 （混合相干矩阵）。 为所有保留–遗忘对构造一个索引集：$\mathcal{P} = \{(r,\ f):\ r \in D_r,\ f \in D_f\}$，其大小为 $|\mathcal{P}| = n_r n_f$。混合相干矩阵 $S \in \mathbb{R}^{|\mathcal{P}|\times|\mathcal{P}|}$逐项定义为

$$S_{(r,f),(r',f')} \;=\; \left\| D_{rf}^{1/2}\, D_{r'f'}^{1/2} \right\|_F \,,$$

对于任意 $(r,\ f),\ (r',\ f') \in \mathcal{P}$。

换言之， S 度量每一对混合海森矩阵 D_{rf} 和 $D_{r'f'}$之间的对齐程度。最后，我们定义一个适用于机器遗忘的整体相干性度量，将单集 σ推广如下：**定义 3.4** （机器遗忘相干性度量）。 遗忘相干性为

$$\sigma \;=\; \frac{\lambda_{\max}(S)}{\max_{(r,f)\in\mathcal{P}}\ \lambda_{\max}(D_{rf})} \,,$$

混合相干矩阵 S 的主特征值除以所有单个 D_{rf} 矩阵中的最大特征值。

当其中一个集合不存在时（例如， $n_f = 0$ 或 $\alpha = 0$ 仅产生一个保留集），该定义便退化为原始的相干性度量。它同时捕捉集合内的相干性和集合间的耦合。直观而言，如果保留集与遗忘集在曲率方向上高度对齐，那么混合相干性 σ 就会很大。在这种情况下，对遗忘集进行上升、对保留集进行下降，往往会相互抵消：两个集合的更新方向相似但符号相反，因而远离 w^*的移动幅度极小。这预示当前最优点具有稳定性（即对机器遗忘具有抵抗力）。反之，如果两个集合不相干（ σ较小），它们的海森矩阵会将优化推向不同方向；对遗忘集进行上升不会被对保留集进行下降所抵消，因此迭代 w_k 更容易逃离原始极小值。总之，我们的多集合相干性度量 σ 量化了数据几何在机器遗忘过程中对发散或收敛的促进程度。据我们所知，这是首项在优化稳定性分析中明确纳入多个数据子集的研究。

3.3 Linear Stability Analysis of Unlearning

在理论部分，我们研究一个更基础的问题：通过优化而非微调阶段，某个集合是否渐近地可被机器遗忘。现在，我们利用上述框架，推导机器遗忘动力学 (3) 收敛或发散的条件。在我们的工作中，一个关键的技术挑战在于，与单集情形不同，随机梯度下降噪声的影响无法用简单的闭式递推关系来刻画。这是因为梯度噪声现在来自两个交错的来源（保留集与遗忘集），且其大小可能不同。

我们先给出一个引理，描述二阶矩 $\mathbb{E}\|w_k\|^2$ 如何根据递归矩阵序列 N_k演化。该引理将 Dexter 等人（2024）的稳定性条件推广到考虑交替上升／下降更新的情形。

引理 3.5 (机器遗忘的稳定性递推)。 考虑在 (4) 中定义的机器遗忘更新算子J_k 。定义半正定矩阵序列 $\{N_k\}_{k\geq 0}$，使得 $N_0 = I$，并且对于 $k \geq 1$:

$$N_k \;=\; C_f \sum_{i\in D_f} H_i\, N_{k-1}\, H_i \;+\; C_r \sum_{i\in D_r} H_i\, N_{k-1}\, H_i\,, \tag{7}$$

其中 C_r、C_f如定义 *3.2* 所述。另设 $M_k = J^{2k} + N_k$。*(*$J = I - \eta(1-\alpha)H_R + \eta\alpha H_F$，其中 H_R 和 H_F 是保留集和遗忘集的完整海森矩阵。参见定义 *A.1*。*)* 然后：

1. （下界） $\mathbb{E}\,\mathrm{Tr}\big(J_0^T \cdots J_{k-1}^T J_{k-1} \cdots J_0\big) \geq \mathrm{Tr}(M_k)$。此外，如果 $\mathrm{Tr}(N_k) \to\infty$ 当 $k \to \infty$时，那么 $\mathbb{E}\|w_k\|^2 \to \infty$ 也成立。*2.* （上界）如果在每一步， J_k的谱有界，即满足$(1-\epsilon)I \succeq J \succeq -(1-\epsilon)I$，其中某个 $\epsilon \in (0,\ 1)$（即 J的所有特征值均位于 $[-(1-\epsilon)\ \ 1-\epsilon]$内），则

$$\begin{aligned}\mathbb{E}\,\mathrm{Tr}\big(J_0^T \cdots J_{k-1}^T J_{k-1} \cdots J_0\big) \;\leq\; \\ \sum_{r=0}^{k-1} \binom{k}{r} (1-\epsilon)^{2(k-r)}\ \mathrm{Tr}(N_r)\,.\end{aligned} \tag{8}$$

此外，如果对所有 r，$\mathrm{Tr}(N_r) \leq \epsilon$，则当 $k \to \infty$ 时， $\mathbb{E}\|w_k\|^2 \to 0$ （遗忘更新在均方意义下收敛）。

讨论。引理 3.5 的第 (1) 部分给出了发散的充分条件：如果“噪声累积”矩阵 N_k （用于刻画随机梯度下降方差在迭代过程中如何累积）的迹无界，那么模型最终会爆炸（偏离最优点）。第 (2) 部分给出了收敛的充分条件：如果每个 J 都是压缩映射（谱范数 < 1至少小于 1 达到 ϵ的幅度），且累积噪声保持较小，那么模型的参数范数将趋于零（也就是说， w_k 会回到最优点）。这些结论将经典稳定性结果推广到了遗忘情形。重要的是， N_k 的递推关系 (7) 没有简单的闭式解，因为 N_{k-1} 同时出现在对集合 D_r 和 D_f求和的表达式中。

保留集与遗忘集之间的这种耦合，使得分析机器遗忘颇具挑战。通过引入相干性度量（定义 3.4 及相关定义），我们克服了这一难题：相干性使我们能够将 Tr(N_k) 与数据相关量（如$\lambda_{\max}(D)$）联系起来，从而推导出易于解释的稳定性判据。

借助相干性框架，我们现在可以陈述主要的稳定性阈值。第一个结果给出了一个条件：在该条件下，机器遗忘动力学会发散 （无法保持在原始极小值处）：

定理 3.6 (机器遗忘的发散判据)。 在引理 *3.5* 的设定下，如果混合海森特征值超过由相干性决定的阈值，机器遗忘过程就会发散。具体而言，如果

$$\lambda_{\max}(D) \;\geq\; \frac{\sqrt{2}\,\sigma}{\eta\Big((1-\alpha)\,n_f\,\sqrt{\frac{n_r}{B}-1}\;+\;\alpha\,n_r\,\sqrt{\frac{n_f}{B}-1}\Big)}\,, \tag{9}$$

那么$\lim_{k\to\infty}\mathbb{E}\|w_k\|^2 = \infty$。等价地，条件(9) 保证机器遗忘算法会因随机动力学而逃离 原始极小值（即发生发散）。

通俗地说，定理 3.6 表明，如果遗忘-保留交互的影响（由 $\lambda_{\max}(D)$衡量）相对于相干性 σ 的稳定作用（以及批大小 B 和相对集合大小等其他因素）足够大，那么遗忘集上的梯度上升就会压过保留集上的下降，导致不稳定性。不等式 (9) 可以看作机器遗忘的定量稳定性极限，或“混沌边缘”：一旦越过这一界限，原始解 w^* 就无法维持。

我们的下一个定理建立了匹配下界，表明上述发散条件本质上是紧的。它保证：当 $\lambda_{\max}(D)$ 低于某一阈值（与(9)同阶）时，可以找到一种使遗忘过程收敛的情形，从而表明该阈值在一般情况下无法被显著改进：

定理 3.7 （收敛条件（匹配下界））。 设 $\lambda_{\max}(D)$ 与 σ满足

$$\lambda_{\max}(D) \;\leq\; \frac{2\,\sigma}{\eta\,C_r'\left(\sigma + n_f\left(\frac{n_r}{B}-1\right)\right)}\,, \tag{10}$$

其中 $C_r' = \sqrt{C_r}/(\sqrt{C_r}+\sqrt{C_f})$（ C_r、 C_f来自定义 *3.2*）。于是，存在保留集与遗忘集的一组半正定海森矩阵 $\{H_i\}$，使得遗忘更新在这些海森矩阵下收敛（即 $\lim_{k\to\infty}\mathbb{E}\|w_k\|^2 = 0$）。

收敛条件(10)在对 σ、n_r， n_f和 B的依赖上与发散条件相对应。当(10)成立时存在一种实现收敛的构造，这表明定理 3.6 中的发散判据在常数因子意义下是紧的。总之，定理 3.6 和 3.7 共同确定了数据相干性与算法参数空间中的一条理论阈值曲线，用以区分稳定（收敛）的机器遗忘与不稳定（发散）的机器遗忘。现在我们可以解释一些常见的机器遗忘策略：

朴素负梯度。一种直接的机器遗忘基线方法是将 $\alpha = 1$ 设为特定值，并仅针对遗忘集进行梯度上升。我们的框架解释了这种方法为何常常失效。如果遗忘集内部相干性较高，其梯度会与 w^* 处的曲率方向对齐，因此上升会沿着单一稳定方向进行，且由于缺乏随机性而无法摆脱极小值。若未体现随机性，较小的学习率可能导致缓慢的发散行为。如果遗忘集和保留集也具有很高的相干性，那么即使不包含保留集，整体相干性仍然很高。在这两种情况下，发散都会受到抑制或减缓，这与实证研究的报告一致：朴素负梯度机器遗忘通常会停滞或振荡，损害保留数据，却几乎无法降低遗忘集上的性能（Ding et al., 2025; Fan et al., 2025; Ding et al., 2025）。

随机标签扰动。 另一种策略是在遗忘过程中引入随机性，例如使用错误标记的数据，或向遗忘集的梯度中注入噪声（参见随机标签遗忘）。用我们的术语来说，这会有意破坏相干性

：如果标签被随机化，遗忘集样本产生的梯度实际上会变得彼此不相关，大幅降低遗忘集内部的 σ。这反过来使模型能够更快地逃离原始极小值。此外，随机化遗忘标签也会减弱遗忘集与保留集之间的耦合（因为此时遗忘集梯度本质上是与保留集海森矩阵正交的随机噪声）。因此，随机标签方法通过降低相干性度量 σ 来改善机器遗忘，使得更容易满足发散判据。（Graves 等人，2020年）

最小-最大（定向遗忘）方法。更复杂的方法会选取对遗忘集性能最“负责”的部分模型权重或方向，然后在这些组成部分上执行上升/下降操作。这可以看作分别将投影矩阵 P_F和 P_R应用于海森矩阵 H_f，H_r，从而将更新集中于特定的特征方向上。此类投影有效减少了遗忘集与保留集更新方向之间的重叠（因为 $P_F H_f$ 和 $P_R H_r$ 作用于不同的子空间），进而降低两组之间的交叉相干性。在我们的框架中，这也对应于更小的整体σ。通过隔离遗忘动力学，最小-最大方法因此降低了保留集对遗忘的干扰能力（反之亦然），使遗忘过程更加有效。（Tang 和 Khanna，2025年；Fan 等人，2024年）

3.4 Memorization and Forgetting

到目前为止，我们的分析一直聚焦于随机梯度噪声（来自小批量采样）的作用。现在，我们转向另一个关键因素：数据本身固有的信号与噪声结构。我们要探讨的是：如果模型已经记忆了某些训练样本（而不是学习到共享信号），这会使得遗忘这些样本更加容易，还是更加困难？我们将展示，模型记忆倾向（当数据的信噪比较低时会出现这种倾向）与机器遗忘难易程度之间存在理论联系。我们的工作旨在识别由高度正交的成分（噪声、离群特征）所导致的记忆现象，以及其与遗忘之间的关系（有关进一步讨论，请参见附录 A.6）。为具体说明这一点，我们考虑一个受 Kou 等人（2023年）理论构造启发的特定数据模型和网络。数据分布经过设计，使得每个样本都包含公共信号与独立噪声的混合成分。具体形式化如下：

定义 3.8（数据设定）。设 $\mu \in \mathbb{R}^d$ 为一个固定的单位范数信号向量。每个训练样本由一个特征对 $x = [x^{(1)}$组成；$x^{(2)}] \in \mathbb{R}^{2d}$（由两个 d 维部分拼接而成）以及一个标签 $y \in \{-1, +1\}$。样本按以下方式生成：

1. 将 y 采样为拉德马赫随机变量（$\Pr(y = +1) = \Pr(y = -1) = \frac{1}{2}$）。
2. 在 $\mathbb{R}^d$中采样一个噪声向量 $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$，其中 σ^2 为噪声方差。
3. 以相同概率，设定为 $x^{(1)} = y\,\mu$ 和 $x^{(2)} = \xi$，或 $x^{(1)} = \xi$ 和 $x^{(2)} = y\,\mu$。换言之，x 的两个半部中，一个携带信号 $y\,\mu$，另一个携带独立噪声。

接下来，我们考虑一个在这些数据上运行、具有 ReLU 激活的两层卷积神经网络（CNN）。该网络有两组卷积滤波器（分别对应正类和负类），并输出一个分数 $f(W, x)$，其符号决定预测标签。具体而言，设 $W^{(+1)}$和 $W^{(-1)}$ 为两个类别的权重矩阵，每个矩阵的形状为 $m \times d$（含有 m 个滤波器）。网络的输出为

$$
\begin{aligned}
& f(W, x) \\
&= \frac{1}{m} \sum_{r=1}^{m} \Big(\mathrm{ReLU}(\langle w_r^{(+1)}, x^{(1)} \rangle) + \mathrm{ReLU}(\langle w_r^{(+1)}, x^{(2)} \rangle) \Big) \\
&- \frac{1}{m} \sum_{r=1}^{m} \Big(\mathrm{ReLU}(\langle w_r^{(-1)}, x^{(1)} \rangle) + \mathrm{ReLU}(\langle w_r^{(-1)}, x^{(2)} \rangle) \Big)
\end{aligned}
\tag{11}
$$

模型采用逻辑损失 $L_S(W) = \frac{1}{n} \sum_{i=1}^{n} \log\big(1 + \exp(-y_i\, f(W,\ x_i))\big)$ 进行训练。我们关注网络能够完美拟合训练数据（插值状态）并可能出现过拟合的情形。

在这种设定下，我们可以分析训练后的解处海森矩阵的相干性。以下结果根据数据的信噪比（SNR）给出了相干性的上界。信噪比定义为 $\mathrm{SNR} = \frac{\|\mu\|}{\sigma\sqrt{d}}$（它衡量每个样本中共同信号相对于噪声的强度）：

定理 3.9（CNN 记忆模型中的相干性界）。 在定义 *3.8* 的数据模型和上文定义的双层 *ReLU* 卷积神经网络下，假设网络经过训练后训练损失接近于零。那么，以至少 $1-8\delta$ 的概率（在数据集的随机抽样下），保留/遗忘划分的相干矩阵 S 的最大特征值满足

$$\lambda_{\max}(S) \;\le\; \mathcal{O}\Big(n_r\, n_f\, d\sigma^2\Big[(\sqrt{C_r'}+\sqrt{C_f'})^2\,(\mathrm{SNR})^2+(C_r'+C_f')\Big]\Big) \tag{12}$$

$$\max_{rf}\lambda_{\max}(D_{rf}) \le \mathcal{O}((C_r'+C_f')(d\sigma^2(\mathrm{SNR})^2+1)), \tag{13}$$

其中， C_r' 和 C_f' 是定理 *3.7* 中定义的归一化保留/遗忘权重比例。考察两个量的比值，我们可以发现，在小信噪比极限和大信噪比极限下：

$$\lim_{SNR\to 0}\frac{\lambda_{\max}(S)^{upper}}{\max_{rf} D_{rf}^{upper}} = \mathcal{O}(n_r n_f)\ , \tag{14}$$

$$\lim_{SNR\to \infty}\frac{\lambda_{\max}(S)^{upper}}{\max_{rf} D_{rf}^{upper}} = \mathcal{O}(n_r n_f(1+\frac{2\sqrt{C_r'C_f'}}{C_r'+C_f'})). \tag{15}$$

讨论 定理 3.9 展示了信噪比通过控制相干性对优化器稳定性所产生的出人意料的作用。特别是，如果数据的信噪比非常低（意味着 μ 相对于噪声 σ 较小），那么网络很可能会记忆噪声。在这种情况下，高维随机噪声向量之间近似正交，因此不同样本对应的海森矩阵对齐程度较差。我们的界表明，与小信噪比极限相比，大信噪比极限下的相干性度量更大，因此较小的信噪比会带来更小的相干性。由此，当模型已经记忆了数据（信噪比较低）时，机器遗忘过程会变得更加容易：模型只需较小的阻力即可偏离原始拟合，这一点已由我们之前的稳定性判据形式化。相反，如果数据具有较高的信噪比（不同样本之间共享的信号占主导），模型就会抓住该信号，从而产生较大的相干性；由于所有样本在方向上达成一致，模型会抗拒离开最优点。

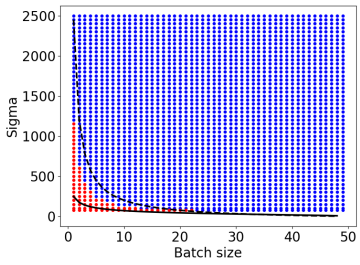
这为一个或许有些违背直觉的格言提供了严谨依据：记忆得越多，就越容易遗忘。换言之，严重依赖单个数据点个体特征（记忆）的模型，在这些数据点上实际上稳定性较差，因此只需较少努力便能将其遗忘；而学到了强大全局结构（信号）的模型则更加稳定，不易移除单个样本的影响。我们的工作首次正式确立了记忆（从数据几何的角度来看）与机器遗忘之间的联系。我们认为，这为理解机器遗忘中固有的权衡提供了有价值的见解。

4 Experiments

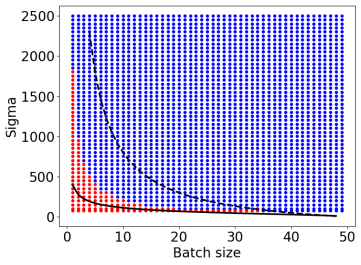
4.1 Diverging and converging condition.

实验设置。 在本节中，我们通过模拟实验来检验定理 3.6 和 3.7。我们固定 $n_f = n_r = 50$ 并设置 $\alpha = 0.1$。设 Q 为超参数常数。我们将 Q 设为不同的值，以控制实验中的各项量。对于保留集，海森矩阵定义为：当 $i \in [Q]$时为 $H_i = me_1e_1^T$，否则为 $H_i = me_{i-Q+1}e_{i-Q+1}^T$，其中$m = 2n_r/Q$；遗忘集采用相同的构造。这样可确保 $\lambda_1(H_R) = \lambda_1(H_F) = \lambda_1(D) = 2$，从而控制锐度。我们选择 $\eta \le 1$ 以避免标准判据 $\eta \ge 2/\lambda_1$导致发散，因此任何逃逸行为都完全源于随机性，这与我们的定理一致。

为了改变相干性，我们调整批大小 Q ，并计算（B， σ）对。对于每一对，我们随机初始化 w，运行 1000 次更新，并记录 $\|w_{1000}\|$。运行结果中



(a) 学习率 $\eta = 0.5$



(b) 学习率 $\eta = 0.8$

图1：**紧致的上界和下界**。蓝色表示 =收敛，红色表示 =发散。虚线为下界（定理 3.7），实线为发散判据（定理 3.6）。两者都与真实边界高度吻合。

$\|w_{1000}\|/\|w_0\| \geq 1000$ 的情况标记为发散。每项实验重复 10 次，并以多数结果判定收敛或发散。

发散界限。图1显示，我们的上界和下界都能准确预测发散区域；对于批量大小为 ≥ 10 的情况，界限更紧。尤其是发散判据与真实边界吻合，表明我们的基于相干性的度量能够准确捕捉优化动力学的本质。这凸显了相干性作为理解机器遗忘动力学的一种有效视角，并具有超出本文范围的潜在应用。更多详情请参阅附录A.3。

4.2 Relation between memorization and forgetting.

实验设置。我们按照定义 3.8 生成数据，并使用两层卷积神经网络。数据集包含 50 个没有标签噪声的训练样本。我们设置 $\boldsymbol{\mu} = \|\boldsymbol{\mu}\|_2[1, 0, \ldots, 0]$，并添加高斯噪声 $\xi \sim \mathcal{N}(0,\ \sigma^2 I_d)$，其中 $\sigma = 1.0$。为了控制实验中的信噪比，我们在固定噪声幅度的同时改变信号强度，以获得不同的信噪比。我们改变 $d \in [100, 1100]$ 的取值，以验证结果在不同过参数化程度下是否成立。该卷积神经网络包含 $m = 10$ 个滤波器，并以 0.1 的学习率进行全批量梯度下降，训练 100 轮，确保训练损失 ≤ 0.1。

我们记录训练损失和测试误差（在 1000 个未见样本上）。对于机器遗忘，在总共 50 个样本中，我们使用 25 个样本构成遗忘集，另外 25 个构成保留集。我们采用负梯度法进行小批量遗忘（批大小为 5），学习率为 0.1，并使用 $\alpha = 0.3$ 执行 90 步，然后记录平均遗忘损失。每项实验重复 20 次并取平均值。

记忆化与遗忘的重叠。图2展示了信号强度和维度上的热图。当训练损失较低但测试误差较高时（左图、中图），我们将其识别为记忆现象。值得注意的是，这些区域与遗忘集上的高损失区域（右图）重合，证实了我们的预测：记忆对应于低相干性，使解不稳定且更容易逃离，从而使机器遗忘更容易。据我们所知，这是首个通过相干性将记忆与遗忘联系起来的研究，为我们的框架提供了有力证据。有关实验设置及其相应目的的更多详细讨论，请参阅附录A.3。

更大规模的真实世界数据。为验证我们的结果在真实世界场景中的有效性，我们在更现实的设置下开展了额外实验：使用 CIFAR-10 和 ResNet-18 模型。我们首先将模型训练至收敛（训练准确率达到 100%）。随后，按照公式 3 所述执行机器遗忘步骤。我们将步长设为 0.01，遗忘集占训练集的 10%。我们将 α 遗忘集与训练集之间的权重设为 0.3。我们记录前 500 个步骤中遗忘集上的损失，每隔 50 个步骤记录一次。为探究我们的理论所预测的记忆与机器遗忘之间的关系，我们向输入中注入不同方差（0.1、0.3、0.5）的高斯噪声。更高的噪声方差会产生更强的记忆，因为网络会过拟合这些特异性的

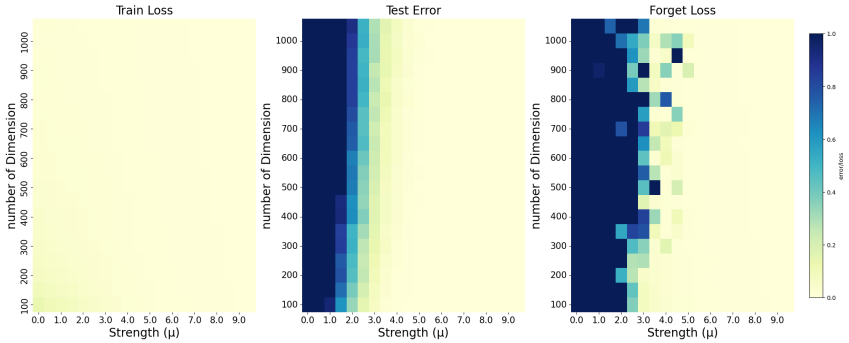


图2：（左）训练损失。（中）测试误差。（右）遗忘损失。如蓝色区域所示，记忆区域与遗忘区域高度重叠。

噪声模式。如表1所示，并且正如我们的相干性框架所预测的那样，具有更强记忆（更大方差）的模型在机器遗忘期间表现出遗忘集上的损失增长更快。

Step	方差 = 0.1	方差 = 0.3	方差 = 0.5
0	0.0016±0.0005	0.0019±0.0004	0.0016±0.0004
50	0.0024±0.0014	0.0016±0.0003	0.0700±0.0730
100	0.0694±0.0858	0.0428±0.0539	0.0715±0.0667
150	0.0477±0.0508	0.0356±0.0401	0.1575±0.1692
200	0.0722±0.0762	0.1582±0.1744	0.6279±0.5556
250	0.1758±0.2054	0.1671±0.1508	1.1366±0.6084
300	0.3888±0.4347	0.5626±0.5598	1.4868±0.9849
350	0.6515±0.7707	0.8544±0.7993	2.4464±1.5124
400	1.4362±1.5605	2.3895±1.2796	5.5148±1.8706
450	2.5029±2.3457	2.9324±1.2882	5.5811±1.8397

表1：不同噪声水平下机器遗忘过程中的遗忘集损失（均值 ± 标准差）。

5 Impact Statement

本文提出了一种基于稳定性的机器遗忘原理性理论，并通过新的保留/遗忘相干性指标推导出紧致的数据依赖阈值，用以预测标准的基于梯度的机器遗忘何时会收敛或发散（即成功或失败）。本文严谨地将记忆与可遗忘性联系起来（较低的信噪比 →较低的相干性→ 更容易遗忘），并通过实证验证所预测的稳定性前沿，从而对数据几何、记忆与有效机器遗忘之间的权衡给出了机理性解释。

References

Agarwala, A. 和 Dauphin, Y. N.《SAM 远离归宿运行：作为一种动力学现象的特征值正则化》，2023年。网址：https://arxiv.org/abs/2302.08692。

Andreyev, A. 和 Beneventano, P.《随机稳定性的边缘：重新审视随机梯度下降的稳定性边缘》，2025年。网址：https://arxiv.org/abs/2412.20553。

Attias, I.、Dziugaite, G. K.、Haghifam, M.、Livni, R. 和 Roy, D. M.《随机凸优化的信息复杂度：在泛化与记忆中的应用》，2024年。网址：https://arxiv.org/abs/2402.09327。

Bhatia, R.《矩阵分析》，第169卷。Springer Science & Business Media，2013年。

Biderman, S., Prashanth, U. S., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S. 和 Raff, E. 大语言模型中涌现且可预测的记忆，2023年。网址：https://arxiv.org/abs/2304.11158。

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D. 和 Papernot, N. 机器遗忘，2020年。网址：https://arxiv.org/abs/1912.03817。Cao, Y. 和 Yang, J. 探索通过机器遗忘让系统忘记。*2015*年*IEEE*安全与隐私研讨会，第463–480页，2015年。网址：https://api.semanticscholar.org/CorpusID:5945696。´Carlini, N., Ulfar Erlingsson 和 Papernot, N. 机器学习中的分布密度、尾部与离群值：度量指标及应用，2019年。网址：https://arxiv.org/abs/1910. 13427。Chang, W.-K. 和 Khanna, R. SAM 与随机梯度下降的统一稳定性分析：数据相干性与简单性偏差的涌现作用，2025年。网址：https://arxiv.org/abs/2511. 17378。Chien, E., Wang, H., Chen, Z. 和 Li, P. 朗之万遗忘：噪声梯度下降在机器遗忘中的新视角，2025年。网址：https://arxiv.org/abs/2401.10371。Chizat, L. 和 Bach, F. 使用逻辑损失训练的宽双层神经网络中梯度下降的隐式偏差，2020年。网址：https://arxiv.org/abs/2002.04486。Dexter, G., Ocejo, B., Keerthi, S., Gupta, A., Acharya, A. 和 Khanna, R. 基于损失曲面几何对随机梯度下降稳定性的精确刻画，2024年。网址：https://arxiv. org/abs/2401.12332。Ding, M., Sharma, R., Chen, C., Xu, J. 和 Ji, K. 理解近似机器遗忘中的微调：理论视角，2025年。网址：https://arxiv.org/abs/2410. 03833。Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D. 和 Liu, S. SalUn：通过图像分类和生成中的基于梯度的权重显著性赋能机器遗忘，2024年。网址：https://arxiv.org/abs/2310.12508。Fan, X., Wu, J., Zhou, M., Liang, P. 和 Phung, D. IMU：影响力引导的机器遗忘，2025年。网址：https://arxiv.org/abs/2508.01620。Feldman, V. 学习是否需要记忆？一个关于长尾的简短故事，2021年。网址：https://arxiv.org/abs/1906.05271。Foret, P., Kleiner, A., Mobahi, H. 和 Neyshabur, B. 尖锐度感知最小化：高效提升泛化能力，2021年。网址：https://arxiv.org/abs/2010.01412。Ghorbani, B., Krishnan, S. 和 Xiao, Y. 基于海森特征值密度探究神经网络优化，2019年。网址：https://arxiv.org/abs/1901.10159。Ginart, A., Guan, M. Y., Valiant, G. 和 Zou, J. 让人工智能忘记你：机器学习中的数据删除，2019年。网址：https://arxiv.org/abs/1907.05012。Golatkar, A., Achille, A. 和 Soatto, S. 纯净网络的永恒阳光：深度网络中的选择性遗忘，2020年。网址：https://arxiv.org/abs/1911.04933。Golatkar, A., Achille, A., Ravichandran, A., Polito, M. 和 Soatto, S. 深度网络中的混合隐私遗忘，2021年。网址：https://arxiv.org/abs/2012.13431。Graves, L., Nagisetty, V. 和 Ganesh, V. 失忆式机器学习，2020年。网址：https: //arxiv.org/abs/2010.10981。Hatua, A., Nguyen, T. T., Cano, F. 和 Sung, A. H. 使用遗忘神经网络实现机器遗忘，2024年。网址：https://arxiv.org/abs/2410.22374。Izzo, Z., Smart, M. A., Chaudhuri, K. 和 Zou, J. 从机器学习模型中近似删除数据，2021年。网址：https://arxiv.org/abs/2002.10077。

Kou, Y., Chen, Z., Chen, Y. 和 Gu, Q. 两层 ReLU 卷积神经网络的良性过拟合，2023年。网址：https://arxiv.org/abs/2303.04145。Kurmanji, M., Triantafillou, P., Hayes, J. 和 Triantafillou, E. 迈向无界机器遗忘，2023年。网址：https://arxiv.org/abs/2302.09880。Lee, S. 和 Jang, C. 基于一种能够感知批次梯度分布的尖锐度度量，对稳定性边缘进行新的刻画。载于第十一届国际学习表征会议，2023年。网址：https://openreview.net/forum?id=bH-kCY6LdKg。Li, H., Xu, Z., Taylor, G., Studer, C. 和 Goldstein, T. 神经网络损失景观的可视化，2018年。网址：https://arxiv.org/abs/1712.09913。Li, H., Huang, D., Wang, Z. 和 Rahmani, A. M. 大语言模型中的偏斜记忆：量化与分解，2025年。网址：https://arxiv.org/abs/2502. 01187。Ma, C. 和 Ying, L. 随机梯度下降的线性稳定性与神经网络的输入平滑性，2021年。网址：https://arxiv.org/abs/2105.13462。Martens, J. 和 Grosse, R. 使用克罗内克因子化近似曲率优化神经网络，2020年。网址：https://arxiv.org/abs/1503.05671。

Mavrothalassitis, I.、Puigdemont, P.、Levi, N. I.与Cevher, V.。上升无法遗忘，2025年。URL https://arxiv.org/abs/2509.26427。

Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J. 和 Schoenholz, S. S. Neural Tangents：Python 中快速易用的无限宽神经网络，2019年。网址：https://arxiv.org/abs/1912.02803。Prashanth, U. S., Deng, A., O'Brien, K., V, J. S., Khan, M. A., Borkar, J., Choquette- Choo, C. A., Fuehne, J. R., Biderman, S., Ke, T., Lee, K. 和 Saphra, N. 复述、重构、回忆：语言模型中的记忆是一种多面现象，2025年。网址：https://arxiv.org/abs/2406.17746。Sekhari, A., Acharya, J., Kamath, G. 和 Suresh, A. T. 记住你想要遗忘的内容：机器遗忘算法，2021年。网址：https://arxiv.org/abs/2103.03279。Shen, S., Zhang, C., Bialkowski, A., Chen, W. 和 Xu, M. Camu：解耦深度模型反学习中的因果效应，2024a。网址：https://arxiv.org/abs/2401.17504。Shen, S., Zhang, C., Zhao, Y., Bialkowski, A., Chen, W. T. 和 Xu, M. 标签无关遗忘：深度模型中的一种无监督反学习方法，2024b。网址：https://arxiv. org/abs/2404.00506。Singh, S. P. 和 Alistarh, D. WoodFisher：神经网络压缩的高效二阶近似，2020年。网址：https://arxiv.org/abs/2004.14340。Tang, H. 和 Khanna, R. 尖锐度感知机器反学习，2025年。网址：https://arxiv.org/abs/2506.13715。Tang, Q., Shpilevskiy, F. 和 L´ecuyer, M. DP-AdamBC：你的 DP-Adam 实际上是差分隐私随机梯度下降（除非你应用偏差校正），2023年。网址：https://arxiv.org/abs/2312.14334。Thudi, A., Deza, G., Chandrasekaran, V. 和 Papernot, N. 展开随机梯度下降：理解影响机器遗忘的因素，2022年。网址：https://arxiv.org/abs/2109. 13398。Vershynin, R.《高维概率：数据科学应用导论》。《剑桥统计与概率数学丛书》。剑桥大学出版社，2018年。Warnecke, A., Pirch, L., Wressnegger, C. 和 Rieck, K. 特征与标签的机器遗忘，2023年。网址：https://arxiv.org/abs/2108.11577。

Wen, K., Li, Z. 和 Ma, T. 尖锐度最小化算法并非仅通过最小化锐度来实现更好的泛化，2023年。网址：https://arxiv.org/abs/2307. 11007。Wu, L. 和 Su, W. J. 随机梯度下降中动力学稳定性的隐式正则化，2023年。网址：https://arxiv.org/abs/2305.17490。Wu, L., Ma, C. 等。过参数化学习中 SGD 如何选择全局极小值：一种动力学稳定性视角。神经信息处理系统进展，31，2018年。Wu, L., Wang, M. 和 Su, W. SGD 噪声的对齐性质及其如何帮助选择平坦极小值：稳定性分析，2022年。网址：https://arxiv.org/abs/2207.02628。Yao, Z., Gholami, A., Keutzer, K. 和 Mahoney, M. PyHessian：通过海森矩阵的视角看神经网络，2020年。网址：https://arxiv.org/abs/1912.07145。Yu, L., Gao, X.-S., Zhang, L. 和 Miao, Y. 记忆型神经网络的泛化能力，2024年。网址：https://arxiv.org/abs/2411.00372。

A Appendix

A.1 Additional related works

机器遗忘与记忆。许多机器遗忘方法被提出，旨在有效擦除特定样本的信息。一些基础但广为人知的方法，例如对遗忘集进行随机标记（Graves 等人，2020年）以及对遗忘集显式执行梯度上升（Warnecke 等人，2023年），为当前的机器遗忘方法奠定了基础。近期研究在这些工作的基础上进一步拓展，以提升机器遗忘的整体效果。例如，SCRUB（Kurmanji 等人，2023年）同时对遗忘集执行梯度上升，并对保留集执行梯度下降，以便在机器遗忘过程中更好地维持保留集上的性能。基于影响力的机器遗忘方法（Izzo 等人，2021年）提出了一种思路：考虑数据集的海森信息，以更新模型权重。显著性遗忘（Fan 等人，2024年）通过梯度幅值识别对遗忘集反应强烈的权重，并仅对这些权重执行机器遗忘，以获得更好的性能。还有一些研究从差分隐私的角度探讨机器遗忘，并提供性能保证。例如，朗之万遗忘（Chien 等人，2025年）研究了通过投影噪声梯度下降实现带隐私保证的机器遗忘。Sekhari 等人（2021年）研究了机器遗忘问题，给出了性能保证及相应的样本复杂度。另有研究讨论记忆与泛化之间的关系。Attias 等人（2024年）在信息论框架下探讨了泛化与记忆之间的根本权衡。Carlini 等人（2019年）讨论了用于识别不同类型样本（已记忆、典型样本等）的各种度量指标。Feldman（2021年）通过理论和实验分析指出，记忆对于实现最优性能是必要的。还有一些研究探讨了不同任务和模型架构中的记忆现象（Biderman 等人（2023年）；Li 等人（2025年）；Prashanth 等人（2025年））。

A.2 Discussion about assumptions

线性化动力学与二次近似：本节将全面而详细地讨论为何这一建模选择是（1）标准的、（2）有实证依据的，以及（3）推动机器遗忘理论进展所必需的。

首先，局部二次近似是深度学习理论中成熟且经过广泛验证的建模框架。近年来大量研究成功地运用这一近似解释了多种优化行为，包括稳定性边缘现象（Andreyev 和 Beneventano，2025年；Lee 和 Jang，2023年）、隐式偏差与泛化（Wu与Su，2023；Wu 等人，2022年）、特征值动力学与曲率结构（Agarwala 和 Dauphin，2023年），以及随机梯度下降的稳定性与发散（Dexter 等人，2024；Chang 和 Khanna，2025年）。尽管这一近似确实存在局限（所有理论框架皆如此），但它仍能为现实世界的系统提供有意义的见解，而目前其他方法尚无法做到。这些研究的成功表明，二次建模在理论上富有成果，在实证上也具有相关性。沿用同一近似遵循了这一传统，并使我们能够通过几何相互作用揭示遗忘集与保留集之间的新联系。

其次，多项实证研究表明，训练良好的极小值附近的损失景观较为平滑，且可用二次型在局部进行良好近似（Li 等人，2018年；Singh & Alistarh，2020年；Ghorbani 等人，2019）。这一现象已在不同架构、数据集和训练机制中反复得到观察。重要的是，机器遗忘是一种微调过程，它从一个已经收敛的极小值开始，而这恰恰是此类局部近似最准确的区域。因此，这一理论基础不仅得到既有实证证据的支持，也特别适用于对机器遗忘动力学建模。

第三，一个关键问题是：为什么理论上的机器遗忘需要近似？现有关于机器遗忘的理论研究也依赖简化系统或线性化系统，以获得可分析的结果。例如，Golatkar 等人（2020年）在二次模型下研究机器遗忘，并分析优化轨迹的漂移，这正是

我们采用的测度类型。Ding 等人（2025年）使用线性特征—权重点积模型和权重空间距离来刻画近似机器遗忘。近期的研究（Mavrothalassitis 等人，2025年）采用线性逻辑回归，从而能够得到闭式解，并从理论上刻画遗忘难度。Thudi 等人（2022年）通过展开随机梯度下降递推来分析机器遗忘，这同样依赖于对梯度动力学的线性化。

所有这些理论框架的一个共同点是，它们都需要采用简化假设，使问题变得可分析。我们的二次近似属于这一成熟方法体系。尽管没有任何一种抽象能够涵盖深度网络的全部复杂性，但每种抽象都能揭示这一现象的不同方面。我们的框架通过提供一种有原则的几何描述，说明保留集与遗忘集如何通过曲率和相干性相互作用，从而揭示了一种新的机制解释，阐明了基于梯度的机器遗忘在何种情况下会变得困难或可行。

Sam逐样本梯度 我们工作中采用的这一假设要求逐样本梯度 w^* 处为零，即，

$$\begin{aligned} \nabla l_i(w) &= \nabla l_i(w^*) + H_i(w - w^*) \\ &= H_i(w - w^*) \end{aligned} \tag{16}$$

这意味着最优点附近的所有梯度都严格取决于曲率。这与我们的研究重点一致：探究曲率如何影响优化行为。这一假设遵循先前研究采用的标准线性插值机制，例如 Dexter 等人，2024；Wu与Su，2023。这种机制在过参数化网络中广泛存在，在收敛附近，训练损失和逐样本损失都会趋近于零。实证和理论研究（Tang等人，2023；Chizat与Bach，2020）表明，现代神经网络能够精确拟合所有训练标签，这意味着损失接近零是由逐样本梯度消失导致的，而非大梯度相互平衡所致。在这种状态下，局部动力学由海森矩阵曲率主导，因此，在极小值附近进行线性近似既是标准做法，也是合理的。

A.3 More discussion about the experiments

下面我们将阐明实验的设计与目的，并在工作修订版中加入相关讨论。实验有两个不同目标：（1）在受控的合成设置下验证理论发散判据和匹配下界；（2）使用两层卷积神经网络，通过实证检验记忆与遗忘之间的预测联系。

合成验证。 我们构造逐样本海森矩阵，使整体数据集海森矩阵受限于 $2 / \eta$，从而确保任何发散行为都纯粹源于随机性和数据几何。我们在保持全局尖锐度不变的同时改变逐样本海森矩阵相干性，然后以不同学习率进行训练。当权重范数增长到 $1000\times$ 大于初始化时，即判定发生发散（重复运行五次）。如图 1 所示，实测的发散边界（红色／蓝色交界处）与我们的理论判据和匹配下界均高度吻合，证实该理论能够准确预测优化稳定性。

记忆—遗忘测试。第二部分，我们希望验证，正如理论所指出的那样，更强的记忆是否会带来更强的遗忘或机器遗忘。当模型记忆数据时，它会将数据映射到损失曲率方面与主数据集正交的空间中，并产生较低的相干性度量。因此，较低的相干性度量会如理论所预测的那样，使遗忘过程更容易。为了控制记忆强度，我们生成了信噪比（SNR）各不相同的数据集。低信噪比数据会迫使模型记忆噪声（在损失曲率空间中学习正交噪声方向），从而产生较低的相干性。训练完成后，我们应用遗忘程序并跟踪遗忘损失。训练损失表明，所有模型都能正确学习数据并收敛。测试损失用于说明模型是否记忆了数据。这是因为，当模型在训练集上表现良好、但在测试集上表现不佳时，就表明存在记忆。因此，我们可以观察到

低信噪比区间表现出较强的记忆。最后，遗忘损失表明我们能否摆脱当前极小值并成功实现机器遗忘。结果表明，这与我们的理论预测一致。记忆程度较高的区间也会带来更强的遗忘效果，这与我们的相干性度量相符。

A.4 Discussion about divergence and unlearning.

本节讨论在不同机器遗忘范围下，发散与机器遗忘之间的关系，以及先前研究如何探索这种关系。此外，我们希望回答为什么基于损失或距离的度量仍然具有价值，并继续作为机器遗忘问题的一项标准。

首先，发散和距离是许多先前研究广泛使用的机器遗忘度量指标。尽管目前跨领域尚无统一的机器遗忘定义：一些应用强调隐私（成员推断攻击抵抗能力），另一些则关注混淆移除、偏差移除、安全性或效用保持。尽管存在这些差异，基于损失和距离的度量仍是最常用的评估工具之一，而我们的工作也延续了这一广泛的研究传统。我们采用这种度量，与许多成熟研究的做法一致；这些研究仍通过遗忘集损失或准确率来评估机器遗忘的有效性。例如，机器遗忘准确率（Fan 等人，2024年）、遗忘误差（Kurmanji 等人，2023年）、遗忘性能（Kurmanji 等人，2023年）、遗忘损失（Graves 等人，2020年；Golatkar 等人，2020年）以及基于距离的度量（Thudi 等人，2022年）。此外，还有研究刻画了由不同算法得到的解之间的距离（Mavrothalassitis 等人，2025年）。之所以仍然使用这些度量，是因为它们能够捕捉移除遗忘集所产生的影响，并且与实际隐私风险（例如成员推断攻击成功率）具有很强的相关性。从这个意义上说，分析决定模型如何偏离旧解的发散量或距离量，与标准实践是一致的。

其次，为什么发散在理论上有意义，并且与以往的机器遗忘理论相吻合？一些奠基性研究通过考察优化轨迹及其对机器遗忘中不同组成部分的响应来分析机器遗忘。例如，Golatkar 等人（2020）通过分析基于二次损失的优化轨迹漂移来研究机器遗忘，并确定时间趋于无穷时的权重差异。Ding 等人（2025）使用线性模型和权重空间距离来刻画近似机器遗忘。具体而言，他们将损失作为衡量遗忘过程所造成差异的标准。近期研究（Mavrothalassitis 等人，2025）采用线性逻辑回归，通过研究基于逻辑损失的闭式解，以原始模型与遗忘后模型之间的权重差异来描述遗忘过程。Thudi 等人（2022）通过展开随机梯度下降递推来分析机器遗忘，这同样依赖于梯度动力学的线性化。他们对机器遗忘误差的定义也直接采用权重空间中的差异，并将成员推断攻击与这一理论定义的量联系起来。

A.5 Feasibility of computing data coherence

审稿人提出的一个实际问题是，该理论的“核心指标依赖于逐样本海森矩阵或格拉姆矩阵”，而在现代大型模型上计算这些量的成本高昂，甚至不可行。这一点确实成立。深度网络中的精确逐样本海森矩阵可能规模极其庞大，但这并不是该方法的致命缺陷。在机器学习理论中，有大量先例表明，一些起初难以处理的量能够带来新的洞见，并最终产生实用的近似方法。费 Fisher 信息矩阵（FIM）和网络的完整海森矩阵就是典型例子：早期理论研究尽管认识到它们规模庞大，仍将其视为重要对象，并由此推动了近似或约束这些矩阵的方法发展（Martens & Grosse，2020；Yao 等人，2020）。自然梯度方法以及 K-FAC（克罗内克因子化近似曲率）等二阶优化器（Martens & Grosse，2020）会显式近似 Fisher 信息矩阵/海森矩阵，以获得接近最优的下降方向。事实上，锐度感知最小化（SAM）（Foret 等人，2021）这一近期提出的、能够改善泛化的正则化方法，正是受到惩罚海森矩阵最大特征值这一思想的启发（即最小化锐度）。SAM 并不计算完整海森矩阵，而是采用巧妙的一阶近似方法（通过扰动权重来间接测量曲率），但其思想源于海森矩阵谱具有重要意义这一原则。同样，神经

切线核最初是数据点上的 N×N 格拉姆矩阵，对于玩具数据集之外的场景而言，这似乎并不实用，但它促成了核代理方法，并启发了实用的诊断工具。例如，研究人员开发了针对部分数据估计神经切线核（Novak 等人，2019）或相关格拉姆矩阵的方法，用于监测训练动力学，并利用常数神经切线核理论解释宽网络为何表现得更加可预测。

在我们的研究中，引入海森对齐/相干性，是为了将其作为理解机器遗忘的概念工具。虽然我们确实在一个小型、受控的卷积神经网络中计算了这一指标，以验证理论，但这与许多理论分析的做法类似：先在一个可以清晰计算精确度量指标的“玩具”设定中进行验证，再逐步扩大规模。值得注意的是，许多大规模理论研究已经找到了近似海森矩阵相关度量的方法。例如，Ghorbani 等人（2019）开发了数值线性代数技术，用于估计 ImageNet 规模网络的完整海森特征值密度。他们采用随机矩阵草图和幂迭代方法，高效地生成海森矩阵谱；而几年前，这似乎还是不可能完成的任务。这表明，蛮力方法无法实现的事情，借助算法上的巧思也可能变得可行。深度学习研究的历史表明，最初“仅用于解释”的方法，最终可能变得切实可用。神经切线核（NTK）曾经纯粹是理论概念，但如今从业者会讨论“神经切线核条件性”，并运用核类比来选择架构。同样，我们预计相干性度量可能启发新的诊断方法（例如，在一个小型留出批次上计算相干性分数，作为代理指标），或新的训练流程（例如，促使遗忘梯度与保留梯度去相关）。在我们的投稿中，我们承认，当前要为一个巨型模型直接计算逐样本海森矩阵并不现实；但我们有意在一个能够精确计算这些矩阵的设定中验证理论，从而确立了明确的真实基准。这是有价值的第一步。今后，我们的框架可以指导对可扩展近似方法的研究：例如，对海森矩阵进行低秩分解，或计算分块相干性（按层计算，或计算特定神经元之间的相干性），以获得成本更低的度量指标。因此，我们有充分理由坚持在分析中使用线性稳定性和局部线性化：这是一种有原则的方法，立足于深度学习理论领域既有的成功经验，并为理解模型何时会——或不会——放下已学内容提供了强有力的解释框架。

A.6 More discussion about memorization and forgetting.

我们对记忆的定义基于如下观察：为了最小化训练损失，模型往往会过拟合到数据中高度正交的分量以及与主信号不相关的方向。这一观点与我们基于相干性的分析一致：在信号加噪声实验中，噪声分量在期望意义上是正交的，而理论预测，一旦上升开始，此类方向很容易被遗忘。

近期工作中也探讨了类似概念。Wen 等人（2023年）表明，被记忆的样本对应于网络内部的正交激活模式，这些模式会转化为正交的海森矩阵方向；而 Yu 等人（2024年）研究了高度正交子空间中的记忆。这些结果支持我们的几何解释，即记忆源于局部化的低相干模式。

我们的相干性框架所捕捉的记忆类型（拟合正交方向或离群特征）是现代深度学习中最基本、研究最广泛的记忆形式之一。它与优化稳定性、泛化行为乃至遗忘都密切相关。我们认同，大语言模型中的逐字序列记忆可能涉及其他机制；然而，据我们所知，这类记忆背后的根本原因仍是一个开放的研究问题，目前尚无相应的理论表述和／或研究。由于我们的工作聚焦于支配基于梯度的学习的优化几何，将基于相干性的分析扩展到大语言模型中的序列记忆或长上下文记忆，是一个令人振奋的未来研究方向。

A.7 Lemmas and proofs

定义 A.1（完全遗忘海森矩阵和保留海森矩阵）。

$$H_R = \frac{1}{n_r}\sum_{r\in D_r} H_r,\;\; H_F = \frac{1}{n_r}\sum_{f\in D_f} H_f, \tag{17}$$

引理 A.2。 l_1-l_2 范数不等式：对于任意 $x\in\mathbb{R}$, $||x||_2 \le ||x||_1 \le \sqrt{d}||x||_2^-$

引理 A.3。 二项式系数：对于所有满足 $k \le n$ 的 $n, k \in \mathbb{N}$，二项式系数满足

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}. \tag{18}$$

引理 A.4。 对于任意矩阵 $M\in\mathbb{R}^{n\times n}$, $||M||_F \le ||M||_{S_1} \le \sqrt{n}||M||_F$，其中 $||M||_{S_p}$ 是 M 的谱的 p 范数，而该不等式由 l_1-l_2 范数不等式得到。

引理 A.5。 对于矩阵 $M_1...M_k \in \mathbb{R}^{n\times n}$，$\mathrm{Tr}[M_1...M_k] \le ||M_1...M_k||_{S_1}$ （参见 *Bhatia（2013）*）

引理 A.6（Vershynin（2018））。 考虑从 N 独立同分布采样的 n 个随机高斯向量 $x_1...x_n$，其中 C_1 是一个常数，使得以概率 $1-\delta$,

$$\sum_{i=1}^{n}\|x_i\| \le n\sqrt{2}\sigma\frac{\Gamma((d+1)/2)}{\Gamma(d/2)} + \sqrt{\frac{n\sigma^2}{C_1}\log(\frac{2}{\delta})} \;\; \textit{for } n,d \textit{ large enough.} \tag{19}$$

引理A.7（Vershynin（2018年））。 考虑从 N（0，$\sigma^2 I_d$）中独立同分布采样的 n 个随机高斯向量 $x_1...x_n$，存在常数 C_1，使得以概率 $1-\delta$,

$$\sum_{i=1}^{n}\|x_i\|^2 \le n\sigma^2 d + \sqrt{\frac{n\sigma^4 d}{C_2}\log(\frac{2}{\delta})} \;\; \textit{for } n,d \textit{ large enough.} \tag{20}$$

A.8 Proof of lemma 3.5

Lem引理A.8（重述）。 考虑在（4）中定义的机器遗忘更新算子 J_k。定义 ~~列半正定~~矩阵 $\{N_k\}_{k\ge 0}$，其由 $N_0 = I$ 给出，并且对于 $k\ge 1$:

$$N_k \;=\; C_f\sum_{i\in D_f} H_i\,N_{k-1}\,H_i \;+\; C_r\sum_{i\in D_r} H_i\,N_{k-1}\,H_i\,, \tag{21}$$

其中 C_r、C_f 如定义 *3.2* 所示。另设 $M_k = J^{2k} + N_k$。则：

1.（下界） $\mathbb{E}\mathrm{Tr}\big(J_0^T\cdots J_{k-1}^T J_{k-1}\cdots J_0\big) \ge \mathrm{Tr}(M_k)$。此外，如果$\mathrm{Tr}(N_k)\to\infty$ 当 $k\to\infty$ 时，则 $\mathbb{E}\|w_k\|^2\to\infty$ 也成立。*2.*（上界）如果在每一步中，J_k 的谱均有界，即（$1-\epsilon$）$I \succeq J \succeq -$（$1-\epsilon$）I，其中 $\epsilon\in$ 属于（0，1）（即 J 的所有特征值均位于 $[-$（$1-\epsilon$） $1-\epsilon]$ 中），则

$$\mathbb{E}\,\mathrm{Tr}\big(J_0^T\cdots J_{k-1}^T J_{k-1}\cdots J_0\big) \;\le\; \sum_{r=0}^{k-1}\binom{k}{r}(1-\epsilon)^{2(k-r)}\;\mathrm{Tr}(N_r)\,.$$

此外，如果对于所有 r 都有$\mathrm{Tr}(N_r) \le \epsilon$，那么当 $k\to\infty$ 时 $\mathbb{E}\|w_k\|^2\to 0$（遗忘更新在均方意义下收敛）。

证明。 由于我们是在计算过程中取期望值，因此实际上可以将 J_k 转换为以下涉及随机变量的形式：

$$J_k = (I - \eta(1-\alpha)\frac{1}{B}\sum_{i\in D_{r,k}} H_i + \eta\alpha\frac{1}{B}\sum_{i\in D_{f,k}} H_i) = (I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r + \eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f), \tag{22}$$

其中 x_r， x_f 是相应的伯努利随机变量，其概率为 $P(x_r = 1) = \frac{B}{n_r}$，而概率为 $P(x_f = 1) = \frac{B}{n_f}$时则为 0。

为了迈出刻画机器遗忘与通常学习过程之间差异的第一步，我们首先计算 $E[J_1^T J_1]$ 如下:

$$\begin{aligned} E[J_1^T J_1] &= E[(I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r + \eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f)^T (I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r + \eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f)] \\ &= E[(I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r)^T (I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r) \\ &+2(I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r)^T (\eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f) + (\eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f)^T (\eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f)]. \end{aligned} \tag{23}$$

Her我们将上述方程分成三部分，并相应地取期望 :

$$\begin{aligned} &E[(I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r)^T (I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r)] \\ &= E[(I - 2\eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r + \eta^2(1-\alpha)^2\frac{1}{B}^2\sum_{r\in D_r} x_r H_r \sum_{r\in D_r} x_r H_r)], \\ &= I - 2\eta(1-\alpha)\frac{1}{n_r}\sum_{r\in D_r} H_r + E[\eta^2(1-\alpha)^2(\frac{1}{B})^2\sum_{r\in D_r} x_r H_r \sum_{r\in D_r} x_r H_r], \\ &= I - 2\eta(1-\alpha)\frac{1}{n_r}\sum_{r\in D_r} H_r + E[\eta^2(1-\alpha)^2(\frac{1}{B})^2\sum_{r'\in D_r}\sum_{r\in D_r} x_{r'} x_r H_{r'} H_r], \\ &= I - 2\eta(1-\alpha)\frac{1}{n_r}\sum_{r\in D_r} H_r + \eta^2(1-\alpha)^2(\frac{1}{n_r})^2(\sum_{r\in D_r} H_r)^2 + \eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B} - \frac{1}{n_r})\sum_{r\in D_r} H_r^2, \\ &= I - 2\eta(1-\alpha)H_R + \eta^2(1-\alpha)^2 H_R^2 + \eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B} - \frac{1}{n_r})\sum_{r\in D_r} H_r^2, \\ &= (I - \eta(1-\alpha)H_R)^2 + \eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B} - \frac{1}{n_r})\sum_{r\in D_r} H_r^2. \end{aligned} \tag{24}$$

随机变量 x_r 彼此独立，但与自身并不独立，因此最终一行中多出一项。此外，与原始随机梯度下降相比，还多乘了一个 $(1-\alpha)^2$。接下来，我们来看交互项:

$$E[2(I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r)^T (\eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f)] = 2(I - \eta(1-\alpha)H_R)^T (\eta\alpha H_F). \tag{25}$$

由于我们假设保留集和遗忘集的采样过程相互独立，因此可以直接按照上述方式进行表述。最后，遗忘集产生的项:

$$E[(\eta\alpha\frac{1}{B}\sum_{f\in D_f}x_f H_f)^T(\eta\alpha\frac{1}{B}\sum_{f\in D_f}x_f H_f)] = \eta^2\alpha^2 H_F^2 + \eta^2\alpha^2\frac{1}{n_f}(\frac{1}{B}-\frac{1}{n_f})\sum_{f\in D_f}^{n_f} H_f^2. \quad (26)$$

接着，我们将这三部分合并，并重新表述雅可比矩阵：

$$\begin{aligned}
E[J_1^T J_1] &= (I-\eta(1-\alpha)H_R)(I-\eta(1-\alpha)H_R) + 2(I-\eta(1-\alpha)H_R)^T(\eta\alpha H_F) + \eta^2\alpha^2 H_F^2 \\
&\quad + \eta^2\alpha^2\frac{1}{n_f}(\frac{1}{B}-\frac{1}{n_f})\sum_{f\in D_f}^{n_f} H_f^2 + \eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B}-\frac{1}{n_r})\sum_{r\in D_r}^{n_r} H_r^2, \\
&= (I-\eta(1-\alpha)H_R+\eta\alpha H_F)(I-\eta(1-\alpha)H_R+\eta\alpha H_F) \\
&\quad + \eta^2\alpha^2\frac{1}{n_f}(\frac{1}{B}-\frac{1}{n_f})\sum_{f\in D_f}^{n_f} H_f^2 + \eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B}-\frac{1}{n_r})\sum_{r\in D_r}^{n_r} H_r^2, \\
&= J^2 + \eta^2\alpha^2\frac{1}{n_f}(\frac{1}{B}-\frac{1}{n_f})\sum_{f\in D_f}^{n_f} H_f^2 + \eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B}-\frac{1}{n_r})\sum_{r\in D_r}^{n_r} H_r^2,
\end{aligned} \quad (27)$$

其中我们定义 $J = I-\eta(1-\alpha)H_R+\eta\alpha H_F$。在整个工作中，我们将分析这些项，以刻画遗忘过程的行为。

我们对理论的第一部分和第二部分均使用归纳证明，并按如下方式开始证明：

第一部分：

基础情形：$\mathbf{k = 1}$

$$M_1 = J^2 + C_f\sum_{f\in D_f}^{n_f} H_f^2 + C_r\sum_{r\in D_r}^{n_r} H_r^2 = J^2 + N_1 \preceq E[J_1^T J_1], \quad (28)$$

其中左侧项与方程 27 相符，因此基例成立。现在我们继续进行归纳步骤。

归纳步骤：k-1

$$
\begin{aligned}
&E[J_k^T J_{k-1}^T ... J_1^T J_1 ... J_{k-1}] \succeq E[J_k^T M_{k-1} J_k], \\
&= E[(I - \eta(1-\alpha)\frac{1}{B}\sum_{i\in D_r} x_r H_r + \eta\alpha\frac{1}{B}\sum_{f\in D_f} x_f H_f) M_{k-1} (I - \eta(1-\alpha)\frac{1}{B}\sum_{i\in D_r} x_r H_r + \eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f)], \\
&= J M_{k-1} J + C_f \sum_{f\in D_f}^{n_f} H_f M_{k-1} H_f + C_r \sum_{r\in D_r}^{n_r} H_r M_{k-1} H_r, \\
&= J(J^{2(k-1)} + N_{k-1})J + C_f \sum_{f\in D_f}^{n_f} H_f (J^{2(k-1)} + N_{k-1}) H_f + C_r \sum_{r\in D_r}^{n_r} H_r (J^{2(k-1)} + N_{k-1}) H_r, \\
&= J^{2k} + C_f \sum_{f\in D_f}^{n_f} H_f N_{k-1} H_f + C_r \sum_{r\in D_r}^{n_r} H_r N_{k-1} H_r + J N_{k-1} J + C_f \sum_{f\in D_f}^{n_f} H_f J^{2(k-1)} H_f + C_r \sum_{r\in D_r}^{n_r} H_r J^{2(k-1)} H_r, \\
&= M_k + J N_{k-1} J + C_f \sum_{f\in D_f}^{n_f} H_f J^{2(k-1)} H_f + C_r \sum_{r\in D_r}^{n_r} H_r J^{2(k-1)} H_r, \\
&\succeq M_k.
\end{aligned} \tag{29}
$$

最后一个等式成立，是因为根据假设，后面三项均为半正定，因为它们相对于整体乘法的左半和右半是对称的。由于我们可以通过 M_k 求得下界， N_k 的发散将导致 M_k ，并使整个乘积发散。

第二部分：

基础步骤：k=1。

$$
E[J_1^T J_1] = J^2 + N_1 \preceq (1-\epsilon)^2 I + N_1 = \sum_{r=0}^{1} \binom{1}{r} (1-\epsilon)^{2(1-r)} N_r. \tag{30}
$$

The J^2 由于我们的假设，其上界为（$1-\epsilon$）$^2 I$。现在，我们从归纳步骤开始: p

归纳步骤：k-1。

$$
\begin{aligned}
&E[J_k^T J_{k-1}^T ... J_1^T J_1 ... J_{k-1} J_{k-1}] \preceq E[J_k^T (\sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_r) J_k], \\
&= J(\sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_r) J \\
&+ C_f \sum_{i \in D_f}^{n_f} H_i (\sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_r) H_i + C_r \sum_{i \in D_r}^{n_r} H_i (\sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_r) H_i, \\
&= J(\sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_r) J + \sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} (C_f \sum_{i \in D_f}^{n_f} H_i N_r H_i + C_r \sum_{i \in D_r}^{n_r} H_i N_r H_i), \\
&\preceq \sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-r)} N_r + \sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_{r+1}, \\
&= \sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-r)} N_r + \sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_{r+1}, \\
&= (1-\epsilon)^2 N_o + \sum_{r=1}^{k-1} (\binom{k-1}{r} + \binom{k-1}{r-1}) N_r + N_k, \\
&= (1-\epsilon)^2 N_o + \sum_{r=1}^{k-1} \binom{k}{r} N_r + N_k, \\
&= \sum_{r=0}^{k} \binom{k}{r} (1-\epsilon)^{2(k-r)} N_r.
\end{aligned}
\tag{31}
$$

□

第一个和第二个不等式由归纳法中对前一步的假设得出，我们在最后一步通过引理 A.3 合并系数。最后，如果进一步有 $\mathbf{Tr}[N_r] \leq \epsilon \; \forall r$，那么

$$
\begin{aligned}
&E[\mathrm{Tr}[J_k^T J_{k-1}^T ... J_1^T J_1 ... J_{k-1} J_{k-1}]], \\
&= \sum_{r=0}^{k} \binom{k}{r} (1-\epsilon)^{2k-r} \mathrm{Tr}[N_r], \\
&\leq \sum_{r=0}^{k} \binom{k}{r} (1-\epsilon)^{2(k-r)} \epsilon^r, \\
&\leq ((1-\epsilon)^2 + \epsilon)^k \leq (1-\epsilon)^k,
\end{aligned}
\tag{32}
$$

当 $k \to \infty$ 时，该值将收敛至零。

A.9 Proof of theorem 3.6

定理 A.9（重述）。 在引理 *3.5* 的设定下，如果混合海森特征值超过由相干性决定的阈值，遗忘过程将发散。特别地，如果

$$\lambda_{\max}(D) \;\geq\; \frac{\sqrt{2}\,\sigma}{\eta\Big((1-\alpha)\,n_f\,\sqrt{\frac{n_r}{B}-1}\;+\;\alpha\,n_r\,\sqrt{\frac{n_f}{B}-1}\Big)}\,, \tag{33}$$

那么极限 $\lim_{k\to\infty}\mathbb{E}\|w_k\|^2 = \infty$。等价地，条件(9) 保证遗忘算法会由于随机动力学而逃离原始极小值（发散）。

证明。 为简化记号，我们使用以下

$$\begin{aligned} &L_k \in \{r,f\}^k \;\; \text{(a string of length } k \text{ over the alphabet } \{r,f\}),\\ &L_k[i] \;\;\mapsto\;\; \text{the } i\text{-th symbol of } L_k,\;\; 1\leq i\leq k. \end{aligned} \tag{34}$$

根据引理 3.5 的第一部分，我们知道可以对 N_k 给出下界，从而为$E[\mathrm{Tr}[J_k^TJ_{k-1}^T...J_1^TJ_1...J_{k-1}J_k]$给出下界。首先，我们可以将总和写成如下形式:

$$\begin{aligned}
\mathrm{Tr}[N_k] &= \sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}} \Big(\sum_{a_k\in D_{L_k[k]}}\sum_{a_{k-1}\in D_{L_k[k-1]}}\cdots\sum_{a_1\in D_{L_k[k]}}\Big)\,\mathrm{Tr}[H_{a_k}...H_{a_1}H_{a_1}...H_{a_k}],\\
&= \sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}} \Big(\sum_{a_k\in D_{L_k[k]}}\sum_{a_{k-1}\in D_{L_k[k-1]}}\cdots\sum_{a_1\in D_{L_k[k]}}\Big)\|H_{a_k}...H_{a_1}\|_F^2,\\
&\geq \frac{1}{d}\sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}} \Big(\sum_{a_k\in D_{L_k[k]}}\sum_{a_{k-1}\in D_{L_k[k-1]}}\cdots\sum_{a_1\in D_{L_k[k]}}\Big)\|H_{a_k}...H_{a_1}\|_{\mathcal{S}_1}^2,\\
&\geq \frac{1}{d}\sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}} \Big(\sum_{a_k\in D_{L_k[k]}}\sum_{a_{k-1}\in D_{L_k[k-1]}}\cdots\sum_{a_1\in D_{L_k[k]}}\Big)\,\mathrm{Tr}[H_{a_k}...H_{a_1}]^2,\\
&\geq \frac{1}{n_rn_f}\frac{1}{d}\sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}} \Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\Big)\,\mathrm{Tr}[H_{a_{L_k[k]}}...H_{a_{L_k[1]}}]^2,\\
&= \frac{1}{n_rn_f}\frac{1}{d}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\Big)\sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}\,\mathrm{Tr}[H_{a_{L_k[k]}}...H_{a_{L_k[1]}}]^2,\\
&\geq \frac{1}{n_rn_f}\frac{1}{d}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\Big)\frac{1}{2^k}\Big(\sum_{L_k\in\{r,f\}^k}\mathrm{Tr}[C_r^{\frac{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}{2}}\, C_f^{\frac{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}{2}} H_{a_{L_k[k]}}...H_{a_{L_k[1]}}]\Big)^2,\\
&= \frac{1}{n_rn_f}\frac{1}{d}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\Big)\frac{1}{2^k}\Big(\mathrm{Tr}[\sum_{L_k\in\{r,f\}^k} C_r^{\frac{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}{2}}\, C_f^{\frac{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}{2}} H_{a_{L_k[k]}}...H_{a_{L_k[1]}}]\Big)^2,\\
&= \frac{1}{n_rn_f}\frac{1}{d}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\Big)\frac{1}{2^k}(C_r^{\frac12}+C_f^{\frac12})^{2k}\Big(\mathrm{Tr}[(\frac{C_r^{\frac12}}{C_r^{\frac12}+C_f^{\frac12}}H_{a_r}+\frac{C_f^{\frac12}}{C_r^{\frac12}+C_f^{\frac12}}H_{a_f})^k]\Big)^2,\\
&= \Big(\frac{1}{n_rn_f}\Big)^2\frac{1}{d}\frac{1}{2^k}(C_r^{\frac12}+C_f^{\frac12})^{2k}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\mathrm{Tr}[(\frac{C_r^{\frac12}}{C_r^{\frac12}+C_f^{\frac12}}H_{a_r}+\frac{C_f^{\frac12}}{C_r^{\frac12}+C_f^{\frac12}}H_{a_f})^k]\Big)^2.
\end{aligned} \tag{35}$$

对于第一个和第二个不等式，我们使用引理 A.4 和 A.5。对于第三个不等式，我们将求和化简为 $\sum_{a_r\in D_r}\sum_{a_f\in D_f}$。 由于存在不涉及 D_r 或 D_f 的项，我们将整个等式除以 n_fn_r ，以确保不等式成立。对于第四个不等式，我们使用引理 A.2。

在尝试将该量与上述内容联系起来之前，我们先对以下内容重新索引：

$$\frac{1}{n_r n_f}\sum_{a_r\in D_r, a_f\in D_f}\frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_r}+\frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_f}=\frac{1}{n_r n_f}\sum_{rf}D_{rf}=D, \tag{36}$$

其中 $D_{rf}=\frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_r+\frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_f$，下标表示对相应的子集（保留集和遗忘集）求和。现在，

我们继续探讨不同量之间的关系。

$$\begin{aligned}
&\mathrm{Tr}[(\frac{1}{n_r n_f}\sum_{a_r\in D_r, a_f\in D_f}\frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_r}+\frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_f})^k]=\mathrm{Tr}[(\frac{1}{n_r n_f})^k(\sum_{rf}D_{rf})^k],\\
&=\mathrm{Tr}[(\frac{1}{n_r n_f})^k\sum_{rf_1}\sum_{rf_2}...\sum_{rf_k}D_{rf_1}D_{rf_2}...D_{rf_{k-1}}D_{rf_k}],\\
&\le d(\frac{1}{n_r n_f})^k\sum_{rf_1}\sum_{rf_2}...\sum_{rf_k}\|D_{rf_k}^{\frac{1}{2}}D_{rf_1}^{\frac{1}{2}}\|_F\|D_{rf_1}^{\frac{1}{2}}D_{rf_2}^{\frac{1}{2}}\|_F...\|D_{rf_{k-1}}^{\frac{1}{2}}D_{rf_k}^{\frac{1}{2}}\|_F,\\
&=d(\frac{1}{n_r n_f})^k\sum_{rf_1}\sum_{rf_2}...\sum_{rf_k}S_{rf_k,rf_1}S_{rf_1,rf_2}...S_{rf_{k-1},rf_k},\\
&=d(\frac{1}{n_r n_f})^k\,\mathrm{Tr}(S^k),\\
&\le d^2(\frac{1}{n_r n_f})^k\lambda_1(S)^k.
\end{aligned} \tag{37}$$

因此，我们可以说

$$\mathrm{Tr}[D^k]\le d^2(\frac{1}{n_r n_f})^k\lambda_1(S)^k, \tag{38}$$

并且可以得到

$$\begin{aligned}
\frac{(n_r n_f)^k\,\mathrm{Tr}[D^k]}{d^2\sigma^k}&\le\frac{(n_r n_f)^k\,\mathrm{Tr}[D^k]}{d^2\lambda_1(S)^k}\max_{i\in D_r j\in D_f}\lambda_1(\frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_i+\frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_j)^k,\\
&\le\sum_{a_r\in D_r}\sum_{a_f\in D_f}\mathrm{Tr}[(\frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_r}+\frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_f})^k].
\end{aligned} \tag{39}$$

因此，我们可以得出结论

$$\begin{aligned}
\mathrm{Tr}[N_k]&\ge\frac{1}{d}\frac{1}{n_f n_r}\frac{1}{2^k}(C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}})^{2k}(\frac{(n_r n_f)^k\,\mathrm{Tr}[D^k]}{d^2\sigma^k})^2,\\
&\ge\frac{1}{d^5}\frac{1}{n_f n_r}\frac{1}{2^k}(C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}})^{2k}\frac{(n_r n_f)^{2k}\lambda_1(D)^{2k}}{\sigma^{2k}},\\
&\ge\frac{1}{d^5}\frac{1}{n_f n_r}\frac{1}{2^k}(C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}})^{2k}\frac{(n_r n_f)^{2k}\lambda_1(D)^{2k}}{\sigma^{2k}}.
\end{aligned} \tag{40}$$

最后，我们可以看到，迹是否发散取决于那些含有 k 次幂的项。因此，通过重新排列，并将系数的定义代入这些项中，我们可以得到

$$\lambda_1(D)\ge\frac{\sqrt{2}\sigma}{\eta}\Big((1-\alpha)n_f(\frac{n_r}{B}-1)^{\frac{1}{2}}+\alpha n_r(\frac{n_f}{B}-1)^{\frac{1}{2}}\Big)^{-1}, \tag{41}$$

这是发散行为的条件 □

A.10 Proof of theorem 3.7

定理 A.10 （（重述）匹配下界。）。 假设 $\lambda_{\max}(D)$ 且 σ 满足

$$\lambda_{\max}(D) \;\le\; \frac{2\,\sigma}{\eta\, C'_r\left(\sigma + n_f\left(\frac{n_r}{B}-1\right)\right)}, \tag{42}$$

其中 $C'_r = \sqrt{C_r}/(\sqrt{C_r}+\sqrt{C_f})$（$C_r$、$C_f$ 来自定义 3.2）。那么，存在一种针对保留集与遗忘集的半正定海森矩阵 $\{H_i\}$ 的选取方式，使得在这些海森矩阵下，遗忘更新收敛（即 $\lim_{k\to\infty}\mathbb{E}\|w_k\|^2 = 0$）。

证明。 我们通过如下构造来证明。我们通过设置 $H_i = m\cdot e_1 e_1^T \;\forall\, i \in [\frac{\sigma}{n_f}]$来构造保留集。（$m = (C'_r)^{-1}\frac{\lambda_1(D)}{\frac{\sigma}{n_f}}\frac{n_r}{}$以及 $C'_r = \frac{C_r^2}{C_r^2+C_f^2}\frac{1}{}$和 C_r 与 C_f 的定义见定义 3.2。）否则，我们将海森矩阵设为零矩阵。对于遗忘集，我们将所有矩阵都设为零矩阵。

我们首先验证混合海森矩阵的特征值确实是指定值 $\lambda_1($ $D)$.

$$D = \frac{1}{n_r n_f}\sum_{rf} C'_r H_r + C'_f H_f = \frac{1}{n_r n_f}\sum_{rf} C'_r (C'_r)^{-1}\frac{\lambda_1(D) n_r}{\frac{\sigma}{n_f}} e_1 e_1^T = \lambda_1(D) e_1 e_1^T, \tag{43}$$

and 我们得到，该构造确实具有相应的混合海森特征值。

我们现在验证相干性度量为指定值 σ。首先注意，相干矩阵的元素为:

$$S_{rf,r'f'} = \sqrt{\mathrm{Tr}[(C'_r H_r + C'_f H_f)(C'_r H_{r'} + C'_f H_{f'})]} = C'_r m = \frac{\lambda_1(D) n_r}{\frac{\sigma}{n_f}},\;\; \forall r, r' \in [\frac{\sigma}{n_f}]. \tag{44}$$

否则为零。我们知道每行和每列有 $n_f \cdot \frac{\sigma}{n_f} = \sigma$ 个非零元素。注意，我们还需要将相干矩阵除以 $\max_{rf} D_{rf} = \max_{rf} C'_r H_r + C'_f H_f = {}^{\lambda_1(D)}{}_{\frac{\sigma}{n_f}}{}^{n_r}$。最后，经过此除法后，每个元素均为 1，因此我们可以得到该矩阵的特征值为 σ，并验证该构造有效。

现在注意，在我们的构造中，每一步 J_i 都彼此可交换，因为涉及的每个矩阵都是对角矩阵。因此，我们可以只关注一个步骤，计算导致发散或收敛的条件。由于我们有意将矩阵设为一维，因此只需研究单个轴上的行为 e_1，方法是将上述内容代入如下:

$$\begin{aligned} e_1 E[J_1 J_1] e_1 &= e_1[I - 2\eta(1-\alpha)H_R + \eta^2(1-\alpha)^2 H_r^2 + \eta^2(1-\alpha)^2\sum_r H_r^2]e_1, \\ &= 1 - 2\eta(1-\alpha)(C'_r)^{-1}\lambda_1(D) + \eta^2(1-\alpha)^2(C'_r)^{-2}\lambda_1(D)^2 + \frac{(C'_r)^{-2}}{\sigma}\lambda_1(D)^2\eta^2(1-\alpha)^2 n_f(\frac{n_r}{B}-1). \end{aligned} \tag{45}$$

由于我们希望研究收敛行为，因此希望上述量小于 1，使得反复相乘后能够收敛。

$$\begin{aligned} &1 - 2\eta(1-\alpha)(C'_r)^{-1}\lambda_1(D) + \eta^2(1-\alpha)^2(C'_r)^{-2}\lambda_1(D)^2 + \frac{(C'_r)^{-2}}{\sigma}\lambda_1(D)^2\eta^2(1-\alpha)^2 n_f(\frac{n_r}{B}-1) \le 1, \\ &\implies 2 \ge \eta(1-\alpha)(C'_r)^{-1}\lambda_1(D)(1 + \frac{n_f}{\sigma}(\frac{n_r}{B}-1)), \\ &\implies 2 \ge \frac{\eta}{\sigma}(1-\alpha)(C'_r)^{-1}\lambda_1(D)(\sigma + n_f(\frac{n_r}{B}-1)), \\ &\implies \lambda_1(D) \le \frac{2\sigma}{\eta}C'_r\Big((1-\alpha)(\sigma + n_f(\frac{n_r}{B}-1))^{-1}\Big). \end{aligned} \tag{46}$$

□

A.11 Proof of theorem A.11

定理 A.11 (重述)。 在定义 *3.8* 的数据模型和上述定义的双层 *ReLU* 卷积神经网络下，假设该网络经过训练后训练损失接近于零。那么，以至少 $1-8\delta$ 的概率（在数据集的随机抽样下），保留/遗忘划分对应的相干矩阵 S 的最大特征值满足

$$\lambda_{\max}(S) \;\le\; \mathcal{O}\Big(n_r\, n_f\, d\sigma^2\Big[(\sqrt{C_r'}+\sqrt{C_f'})^2\,(\mathrm{SNR})^2+(C_r'+C_f')\Big]\Big)\,, \tag{47}$$

$$\max_{rf}\lambda_{\max}(D_{rf}) \le \mathcal{O}((C_r'+C_f')(d\sigma^2(\mathrm{SNR})^2+1)), \tag{48}$$

其中 C_r' 和 C_f' 是定理*3.7*中所定义的归一化保留/遗忘权重比例。考虑两个量相除，我们可以发现，在小信噪比极限和大信噪比极限下：

$$\lim_{SNR\to 0}\frac{\lambda_{\max}(S)^{upper}}{\max_{rf} D_{rf}^{upper}}=\mathcal{O}(n_r n_f)\;\;,\lim_{SNR\to\infty}\frac{\lambda_{\max}(S)^{upper}}{\max_{rf} D_{rf}^{upper}}=\mathcal{O}(n_r n_f(1+\frac{2\sqrt{C_r'C_f'}}{C_r'+C_f'})). \tag{49}$$

证明。 我们首先计算一个样本相对于其中一个$w_{j,\ r}$的梯度。

$$\frac{\partial \ell(y_i\cdot f(W,x_i))}{\partial w_{j,r}}=\ell_i'\cdot\frac{j}{m}\cdot(\mathbf{1}_{\{\langle w_{j,r},y_i\cdot\boldsymbol{\mu}\rangle>0\}}\cdot\boldsymbol{\mu}+\mathbf{1}_{\{\langle w_{j,r},\xi_i\rangle>0\}}\cdot y_i\cdot\xi_i). \tag{50}$$

上述方程中有若干下标（即 j 和 r），我们用它们对特定的特征权重向量求导。我们将在后续计算中继续使用这一记号。现在，我们转而计算数据 i 关于两个不同权重特征的二阶导数，如下：

$$\frac{\partial^2 \ell(y_i\cdot f(W,x_i))}{\partial w_{j,r}\partial w_{j',r'}}=$$
$$\ell_i''\cdot\frac{jj'}{m^2}\cdot(\mathbf{1}_{\{\langle w_{j,r},y_i\cdot\boldsymbol{\mu}\rangle>0\}}\cdot\boldsymbol{\mu}+\mathbf{1}_{\{\langle w_{j,r},\xi_i\rangle>0\}}\cdot y_i\cdot\xi_i)(\mathbf{1}_{\{\langle w_{j',r'},y_i\cdot\boldsymbol{\mu}\rangle>0\}}\cdot\boldsymbol{\mu}+\mathbf{1}_{\{\langle w_{j',r'},\xi_i\rangle>0\}}\cdot y_i\cdot\xi_i)^T. \tag{51}$$

以上是海森矩阵的一个分块。接下来，为简化记号，我们将指示函数（ReLU 的导数）简写为$\mathbf{1}_{j',r',y_i\cdot\boldsymbol{\mu}}$ 和$\mathbf{1}_{j',r',\boldsymbol{\xi}}$，以避免繁复的记号。要计算相干矩阵，我们需要计算不同样本对应的海森矩阵乘积的迹，

$$
\begin{aligned}
&\mathrm{Tr}[H_i H_k] = \sum_{j,j',r,r'} \mathrm{Tr}[\frac{\partial^2 \ell(y_i \cdot f(W,x_i))}{\partial w_{j,r} \partial w_{j',r'}} \frac{\partial^2 \ell(y_k \cdot f(W,x_k))}{\partial w_{j',r'} \partial w_{j,r}}], \\
&= \frac{\ell_i'' \ell_k''}{m^4} \sum_{j,j',r,r'} (\mathbf{1}_{j,r,y_k \cdot \boldsymbol{\mu}} \cdot \boldsymbol{\mu} + \mathbf{1}_{j,r,\xi_k} \cdot y_k \cdot \xi_k)^T (\mathbf{1}_{j,r,y_i \cdot \boldsymbol{\mu}} \cdot \boldsymbol{\mu} + \mathbf{1}_{j,r,\xi_i} \cdot y_i \cdot \xi_i) \\
&(\mathbf{1}_{j',r',y_i \cdot \boldsymbol{\mu}} \cdot \boldsymbol{\mu} + \mathbf{1}_{j',r',\xi_i} \cdot y_i \cdot \xi_i)^T (\mathbf{1}_{j',r',y_k \cdot \boldsymbol{\mu}} \cdot \boldsymbol{\mu} + \mathbf{1}_{j',r',\xi_k} \cdot y_k \cdot \xi_k), \\
&= \frac{\ell_i'' \ell_k''}{m^4} \Big((\sum_{j,r} \mathbf{1}_{j,r,y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j,r,y_i \cdot \boldsymbol{\mu}}) \|\boldsymbol{\mu}\|^2) + (\sum_{j,r} \mathbf{1}_{j,r,y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j,r,\xi_k}) \boldsymbol{\mu}^T \xi_k + \\
&(\sum_{j,r} \mathbf{1}_{j,r,y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j,r,\xi_i}) \boldsymbol{\mu}^T \xi_i + (\sum_{j,r} \mathbf{1}_{j,r,\xi_k} \mathbf{1}_{j,r,\xi_i}) \xi_k^T \xi_i \Big), \\
&\Big((\sum_{j,r} \mathbf{1}_{j',r',y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j',r',y_i \cdot \boldsymbol{\mu}}) \|\boldsymbol{\mu}\|^2) + (\sum_{j',r'} \mathbf{1}_{j',r',y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j',r',\xi_k}) \boldsymbol{\mu}^T \xi_k + \\
&(\sum_{j',r'} \mathbf{1}_{j',r',y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j',r',\xi_i}) \boldsymbol{\mu}^T \xi_i + (\sum_{j',r'} \mathbf{1}_{j',r',\xi_k} \mathbf{1}_{j',r',\xi_i}) \xi_k^T \xi_i \Big), \\
&\leq 4 \frac{\ell_i'' \ell_k''}{m^2} (\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_k| + |\boldsymbol{\mu}^T \xi_i| + |\xi_i^T \xi_k|)^2, \\
&\leq \frac{4}{m^2} (\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_k| + |\boldsymbol{\mu}^T \xi_i| + |\xi_i^T \xi_k|)^2.
\end{aligned} \tag{52}
$$

现在我们来分析相干矩阵中的各项。

$$
\begin{aligned}
S_{r_1 f_{1'}, r_2 f_{2'}} &= \sqrt{\mathrm{Tr}((C_r' H_{r_1} + C_f' H_{f_{1'}})(C_r' H_{r_2} + C_f' H_{f_{2'}}))}, \\
&= \sqrt{\mathrm{Tr}[C_r'^2 H_{r_1} H_{r_2}] + \mathrm{Tr}[C_r' C_f' H_{r_1} H_{f_{2'}}] + \mathrm{Tr}[C_r' C_f' H_{f_{1'}} H_{r_2}] + \mathrm{Tr}[C_f'^2 H_{f_{1'}} H_{r_{2'}}]}, \\
&\leq \sqrt{\mathrm{Tr}[C_r'^2 H_{r_1} H_{r_2}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{r_1} H_{f_{2'}}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{f_{1'}} H_{r_2}]} + \sqrt{\mathrm{Tr}[C_f'^2 H_{f_{1'}} H_{r_{2'}}]},
\end{aligned} \tag{53}
$$

其中， C_r' 和 C_f' 分别是上一节提到的归一化系数。

我们的目标是估计相干矩阵的最大特征值，以及它与设计中不同变量之间的关系。为了估计最大特征值，我们引入 ϵ-net，它采用随机矩阵理论。

$$
\lambda_1 = \sup_{\|x\|=1} \langle x, Sx \rangle. \tag{54}
$$

对于一个向量 x，我们可以将表达式写成求和形式:

$$
\begin{aligned}
&\langle x, Sx \rangle = \sum_{r_1 f_{1'}, r_2 f_{2'}} S_{r_1 f_{1'}, r_2 f_{2'}} x_{r_1 f_{1'}} x_{r_2 f_{2'}}, \\
&\leq \sum_{r_1 f_{1'}, r_2 f_{2'}} (\sqrt{\mathrm{Tr}[C_r'^2 H_{r_1} H_{r_2}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{r_1} H_{f_{2'}}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{f_{1'}} H_{r_2}]} + \sqrt{(\mathrm{Tr}[C_f'^2 H_{f_{1'}} H_{r_{2'}}])} x_{r_1 f_{1'}} x_{r_2 f_{2'}}.
\end{aligned} \tag{55}
$$

我们可以通过随机矩阵理论对上述内容进行估计，并利用我们建立的逐元素计算来给出最大特征值的上界，再对每个随机变量使用尾界，从而得出受控变量与所得最大特征值之间的关系。我们首先将讨论分为几种情况。第一种情况，当我们有四个不同的样本 r_1， r_2， f_1'， f_2' 时，我们可以得到

$$
\begin{aligned}
&(\sqrt{\mathrm{Tr}[C_r'^2 H_{r_1} H_{r_2}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{r_1} H_{f_{2'}}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{f_{1'}} H_{r_2}]} + \sqrt{\mathrm{Tr}[C_f'^2 H_{f_{1'}} H_{r_{2'}}]}) x_{r_1 f_{1'}} x_{r_2 f_{2'}},\\
&\le (C_r' \frac{2}{m}(\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_{r1}| + |\boldsymbol{\mu}^T \xi_{r2}| + |\xi_{r1}^T \xi_{r2}|) + \sqrt{C_r' C_f'} \frac{2}{m}(\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_{r1}| + |\boldsymbol{\mu}^T \xi_{f2'}| + |\xi_{r1}^T \xi_{f2'}|),\\
&+ \sqrt{C_r' C_f'} \frac{2}{m}(\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_{r2}| + |\boldsymbol{\mu}^T \xi_{f1'}| + |\xi_{f1'}^T \xi_{r2}|) + C_f' \frac{2}{m}(\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_{f2'}| + |\boldsymbol{\mu}^T \xi_{f1'}| + |\xi_{f1'}^T \xi_{f2'}|)) x_{r_1 f_{1'}} x_{r_2 f_{2'}},\\
&\le (\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\| + \sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|)(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\| + \sqrt{C_r'}\|\xi_{r2}\| + \sqrt{C_f'}\|\xi_{f2'}\|) x_{r_1 f_{1'}} x_{r_2 f_{2'}}.
\end{aligned}
\tag{56}
$$

我们这样做的目的是建立 CNN网络中所用不同变量之间的关系。由上文可知，我们可以通过向量 $v_{rf} = \sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\| + \sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|$ 的叉积给出特征值的上界，因为相干矩阵逐元素小于等于该向量。（即 $\lambda_1(S) \le \lambda_1(vv^T) = \|v^T v\|^2$），这就将特征值的估计转化为向量大小的估计。

现在，我们分析 $v^T v$,

$$
\begin{aligned}
&v^T v = \sum_{rf}(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\| + \sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|)(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\| + \sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|),\\
&= \sum_{rf}(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)^2 + 2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)(\sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|) + (\sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|)^2,\\
&= n_r n_f(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)^2 + 2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)\sum_{rf}(\sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|) +\\
&\sum_{rf}(C_r'\|\xi_{r1}\|^2 + C_f'\|\xi_{f1'}\|^2 + \sqrt{C_r' C_f'}'\|\xi_{r1}\|\|\xi_{f1'}\|).
\end{aligned}
\tag{57}
$$

我们按如下方式分析各项:

$$
\begin{aligned}
&2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)\sum_{rf}(\sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|) =\\
&\qquad 2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)(n_f \sum_r \sqrt{C_r'}\|\xi_{r1}\| + n_r \sum_f \sqrt{C_f'}\|\xi_{f1'}\|).
\end{aligned}
\tag{58}
$$

我们知道 $\|\xi_{r1}\|$, $\|\xi_{f1'}\|$ 服从卡分布，而卡分布也是次指数分布。我们可以利用次指数随机变量求和的尾界，得到该求和的高概率界。由此可知，以概率 2δ,

$$
\begin{aligned}
&2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)\sum_{rf}(\sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|),\\
&\le 2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)(n_r n_f \sqrt{C_r'}\sigma\sqrt{d} + n_f n_r \sqrt{C_f'}\sigma\sqrt{d} + n_f\sqrt{\frac{n_r\sigma^2}{C_1}\log(\frac{2}{\delta})} + n_r\sqrt{\frac{n_f\sigma^2}{C_1}\log(\frac{2}{\delta})}).
\end{aligned}
\tag{59}
$$

现在，我们转向下一个卡方分布项 $C_r' \sum \|\xi_{r_1}\|^2, C_f' \sum \|\xi_{f_1'}\|^2$。利用引理A.7，我们可以得到，以概率 $1-\delta$,

$$
C_r' \sum_{rf} \|\xi_{r_1}\|^2 \le C_r'(n_f n_r \sigma^2 d + n_f \sqrt{\frac{n_r \sigma^4 d}{C_2}\log(\frac{2}{\delta})}).
\tag{60}
$$

因此 $C'_f \sum \|\xi_{f'_1}\|^2$ 也是如此,

$$C'_f \sum_{rf} \|\xi_{r_1}\|^2 \leq C'_f(n_f n_r \sigma^2 d + n_r \sqrt{\frac{n_f \sigma^4 d}{C_2} \log(\frac{2}{\delta})}). \tag{61}$$

项 $\sum_{rf} \sqrt{C'_r C'_f} \|\xi_{r1}\| \|\xi_{f1'}\|$ 也可以用同样的方式处理,

$$\begin{aligned}
&\sqrt{C'_r C'_f} \sum_{rf} \|\xi_{r1}\| \|\xi_{f1'}\| \leq \sqrt{C'_r C'_f} (\sum_r \|\xi_{r1}\|)(\sum_f \|\xi_{f1'}\|), \\
&\leq \sqrt{C'_r C'_f} (n_r \sqrt{2}\sigma \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} + \sqrt{\frac{n_r \sigma^2}{C_1} \log(\frac{2}{\delta})})(n_f \sqrt{2}\sigma \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} + \sqrt{\frac{n_f \sigma^2}{C_1} \log(\frac{2}{\delta})}).
\end{aligned} \tag{62}$$

为简化分析，我们只保留大小至少为 $n_f n_r$ 的项。我们将以概率 $1 - 6\delta$ 得到这一结果

$$\begin{aligned}
\lambda_1(S) \leq \mathcal{O}\Big(n_f n_r \big((\sqrt{C'_r} + \sqrt{C'_f})^2 \|\boldsymbol{\mu}\|^2 + 2\sqrt{2}(\sqrt{C'_r} + \sqrt{C'_f})^2 \|\boldsymbol{\mu}\| \sigma (\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}) \\
+ (C'_r + C'_f)\sigma^2 d + 2\sigma^2 \sqrt{C'_r C'_f} (\frac{\Gamma((d+1)/2)}{\Gamma(d/2)})^2)\Big).
\end{aligned} \tag{63}$$

为了解信噪比（$\text{SNR} = \frac{\|\boldsymbol{\mu}\|}{\sigma\sqrt{d}}$）如何影响右侧，我们从所有相关项中提取因子 $\sigma^2 d$：

$$\begin{aligned}
\lambda_1(S) &\leq \mathcal{O}\Big(n_f n_r \sigma^2 d \big((\sqrt{C'_r} + \sqrt{C'_f})^2 (\text{SNR})^2 + \frac{2\sqrt{2}}{\sqrt{d}} (\sqrt{C'_r} + \sqrt{C'_f})^2 (\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}) (\text{SNR}) \\
&\quad + (C'_r + C'_f) + \frac{2}{d} \sqrt{C'_r C'_f} (\frac{\Gamma((d+1)/2)}{\Gamma(d/2)})^2)\Big), \\
&\leq \mathcal{O}\Big(n_f n_r \sigma^2 d \big((\sqrt{C'_r} + \sqrt{C'_f})^2 (\text{SNR})^2 + (C'_r + C'_f)\big)\Big).
\end{aligned} \tag{64}$$

其中，在最后一个方程中，我们省略了分母中含有 d 的项，因为在考虑更大的网络时，d 往往会变得很大。

对于证明的第二部分，我们知道 H_i 具有如下块结构:

$$\begin{aligned}
&\frac{\partial^2 \ell(y_i \cdot f(W, x_i))}{\partial w_{j,r} \partial w_{j',r'}} = \\
&\ell''_i \cdot \frac{jj'}{m^2} \cdot (\mathbf{1}_{\{\langle w_{j,r}, y_i \cdot \boldsymbol{\mu} \rangle > 0\}} \cdot \boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j,r}, \xi_i \rangle > 0\}} \cdot y_i \cdot \xi_i)(\mathbf{1}_{\{\langle w_{j',r'}, y_i \cdot \boldsymbol{\mu} \rangle > 0\}} \cdot \boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j',r'}, \xi_i \rangle > 0\}} \cdot y_i \cdot \xi_i)^T.
\end{aligned} \tag{65}$$

我们可以看到，整个 H_i 矩阵可以视为向量 vv^T 的外积，其中我们有 v_{jr} 为

$$v_{jr} = \frac{l''_i j}{m} (\mathbf{1}_{\{\langle w_{j,r}, y_i \cdot \boldsymbol{\mu} \rangle > 0\}} \cdot \boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j,r}, \xi_i \rangle > 0\}} \cdot y_i \cdot \xi_i). \tag{66}$$

我们可以立即知道， H_i 的特征值将以 $v^T v$ 为上界，如下所示：

$$\begin{aligned}
\lambda_{\max}(H_i) \leq v^T v &= \frac{l_i''}{m^2} \sum_{jr} (\mathbf{1}_{\{\langle w_{j,r}, y_i \cdot \boldsymbol{\mu} \rangle > 0\}} \cdot \boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j,r}, \xi_i \rangle > 0\}} \cdot y_i \cdot \xi_i)^2, \\
&\leq 2\frac{l_i''}{m^2} \sum_{jr} \mathbf{1}_{\{\langle w_{j,r}, y_i \cdot \boldsymbol{\mu} \rangle > 0\}} \|\boldsymbol{\mu}\|^2 + \mathbf{1}_{\{\langle w_{j,r}, \xi_i \rangle > 0\}} \|\xi_i\|^2, \\
&\leq 2\frac{l_i''}{m^2} \sum_{jr} \|\boldsymbol{\mu}\|^2 + \|\xi_i\|^2, \\
&\leq 2\frac{1}{m^2} \sum_{jr} \|\boldsymbol{\mu}\|^2 + \|\xi_i\|^2, \\
&= C(\|\boldsymbol{\mu}\|^2 + \|\xi_i\|^2),
\end{aligned} \tag{67}$$

其中我们用 C 涵盖所有常数。

为了界定 $\max_{rf} \lambda_{\max}(D_{rf}) = \lambda_{\max}(C_r' H_r + C_f' H_f)$，我们可以使用以下结果：

$$\lambda_{\max}(D_{rf}) = \lambda_{\max}(C_r' H_r + C_f' H_f) \leq C_r' \lambda_{\max}(H_r) + C_f' \lambda_{\max}(H_f). \tag{68}$$

那么，对于任意 $\delta \in (0,1)$，以至少 $1-\delta$ 的概率，我们可以用以下结果给出 H_r 的上界（$\|\xi_i\|$ 是次指数的）：

$$\begin{aligned}
\max_{1 \leq i \leq n_r} C(\|\boldsymbol{\mu}\|^2 + \|\xi_i\|^2) \;&\leq\; C\left(\|\boldsymbol{\mu}\|^2 + \sigma^2\Big[d \;+\; 2\sqrt{d \log \frac{n_r}{\delta}} \;+\; 2\log \frac{n_r}{\delta}\Big]\right), \\
&\leq \mathcal{O}(\|\boldsymbol{\mu}\|^2 + \sigma^2 d).
\end{aligned} \tag{69}$$

类似地，我们可以得到关于 H_f 的界，其阶相同；结合起来，以概率 $1-8\delta$，我们可以得到

$$\lambda_{\max}(D_{rf}) \leq \mathcal{O}((C_r' + C_f')(\|\boldsymbol{\mu}\|^2 + \sigma^2 d)) = \mathcal{O}((C_r' + C_f')\sigma^2 d(\mathrm{SNR}^2 + 1)) \tag{70}$$

最后进行相除并取极限，即可得到以下结果：

$$\lim_{\mathrm{SNR} \to 0} \frac{\lambda_{\max}(S)^{\mathrm{upper}}}{\max_{rf} D_{rf}^{\mathrm{upper}}} = \mathcal{O}(n_r n_f) \;, \lim_{\mathrm{SNR} \to \infty} \frac{\lambda_{\max}(S)^{\mathrm{upper}}}{\max_{rf} D_{rf}^{\mathrm{upper}}} = \mathcal{O}(n_r n_f (1 + \frac{2\sqrt{C_r' C_f'}}{C_r' + C_f'})). \tag{71}$$

□