课后练习：
基于多层感知机（MLP）的噪声数据过拟合现象探究

刘昭阳 2021211656
中山大学数学学院 (珠海)

2026 年 1 月 6 日

摘要

本实验旨在深入探究深度神经网络在小样本噪声数据下的过拟合行为及其内在机制。我们构建了一个包含两个隐藏层（共约 1000 个参数）的多层感知机（MLP），用于拟合仅含 100 个样本点的含噪正弦波数据。实验结果揭示了过拟合发生的三个显著阶段：快速学习期、平台期和噪声记忆期。进一步的分析表明：(1) 标签噪声是导致过拟合的主导因素，当噪声强度 σ_y 达到 0.8 时，模型完全丧失预测能力；(2) 标准 MLP 缺乏对周期性函数的归纳偏置，导致其在训练区间 $[0, 1]$ 之外的外推能力几乎为零；(3) 模型容量与数据密度的匹配至关重要，过深或过宽的网络在小样本下更易陷入高方差状态。本报告详细记录了实验方法、训练动态及结果分析，并结合偏差-方差权衡理论，讨论了正则化策略及模型结构选择对泛化性能的影响。代码仓库

目录

1 引言

深度学习模型因其强大的函数逼近能力而广受关注。理论上，只要有足够的隐藏单元，一个单隐层的前馈神经网络可以以任意精度逼近任何连续函数。然而，这种强大的能力是一把双刃剑。当模型参数量远超样本信息量时，模型容易陷入过拟合陷阱，即“死记硬背”训练数据中的噪声，导致泛化能力下降。

在实际应用中，我们往往面临着有限且带有噪声的数据。如何在模型复杂度与泛化能力之间找到平衡，是机器学习的核心问题之一。本实验通过一个经典的一维回归问题，旨在：

1. **可视化过拟合现象**：直观展示模型在过度训练后如何拟合噪声。
2. **探究噪声影响**：分析不同类型和强度的噪声对模型学习的干扰。
3. **评估外推能力**：测试模型在训练分布之外的预测表现。
4. **分析超参数敏感性**：系统研究模型结构和训练设置对性能的影响。

通过这些实验，我们希望能够深入理解神经网络的行为特性，为设计更鲁棒的深度学习模型提供经验依据。

2 实验设置与方法

2.1 数据生成

为了模拟真实观测环境，我们生成了基于正弦函数的合成数据。数据集 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ 生成过程如下：

1. 在区间 $[0, 1]$ 上均匀采样 $N = 100$ 个点作为输入 x。
2. 计算真实信号 $y_{true} = \sin(2\pi x)$。
3. 添加高斯白噪声 $\epsilon \sim \mathcal{N}(0, \sigma^2)$，其中 $\sigma = 0.4$。

最终观测值为 $y = y_{true} + \epsilon$。数据分布如图 1 所示。

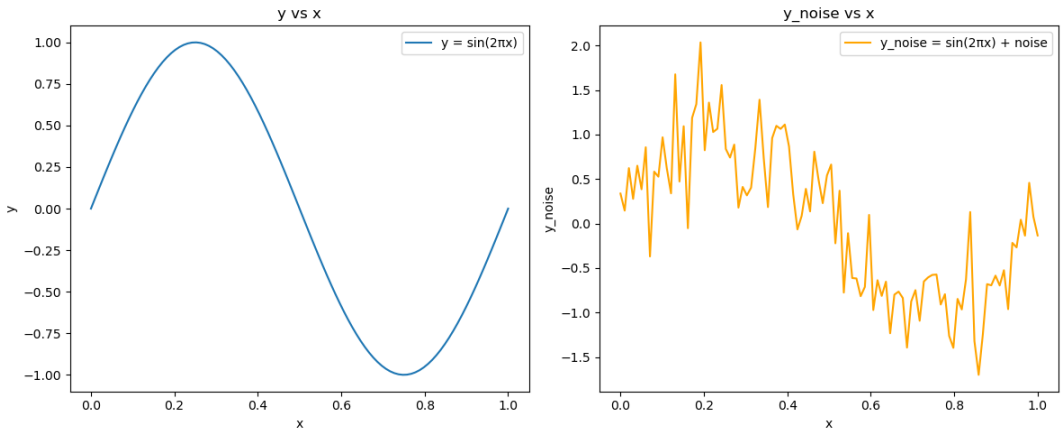

图 1: 数据分布可视化：左图为未受污染的真实信号，右图为用于训练的含噪数据。

2.2 模型架构

实验采用全连接神经网络（MLP），其数学表达如下：
设输入为 $x \in \mathbb{R}$，模型的计算过程可表示为：

$$h_1 = \tanh(W_1 x + b_1), \qquad W_1 \in \mathbb{R}^{32\times 1}, b_1 \in \mathbb{R}^{32} \tag{1}$$

$$h_2 = \mathrm{ReLU}(W_2 h_1 + b_2), \qquad W_2 \in \mathbb{R}^{32\times 32}, b_2 \in \mathbb{R}^{32} \tag{2}$$

$$\hat{y} = W_3 h_2 + b_3, \qquad W_3 \in \mathbb{R}^{1\times 32}, b_3 \in \mathbb{R}^{1} \tag{3}$$

其中 W_l 和 b_l 分别表示第 l 层的权重矩阵和偏置向量。具体结构设计如下：

- **输入层**：1 维。
- **隐藏层 1**：32 个神经元，激活函数为双曲正切 $\tanh(\cdot)$，旨在捕捉数据的平滑特征。
- **隐藏层 2**：32 个神经元，激活函数为线性整流单元 $\mathrm{ReLU}(\cdot)$，引入稀疏性并缓解梯度消失问题。
- **输出层**：1 维线性输出。

该模型总参数量约为 $1\times 32 + 32\times 32 + 32\times 1 + 32 + 32 + 1 = 1121$ 个，远多于 100 个训练样本点，具备发生过拟合的理论容量。

2.3 训练策略

模型的参数 $\theta = \{W_l, b_l\}_{l=1}^{3}$ 通过最小化均方误差（Mean Squared Error, MSE）进行优化：

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f(x_i;\theta)\right)^2 + \lambda \sum_{l} \|W_l\|_F^2 \tag{4}$$

其中 $f(x_i;\theta)$ 为模型的预测输出。在本实验中，为了观察纯粹的过拟合现象，我们设置正则化系数 $\lambda = 0$（即不使用权重衰减）。

具体的训练超参数设置如下：

- **优化器**：AdamW，一种基于梯度的随机优化算法，结合了 Adam 的自适应学习率和解耦的权重衰减。
- **学习率**：$\eta = 0.001$，并在训练过程中保持恒定。
- **训练轮次**：1000 Epochs。我们特意不设置早停（Early Stopping）机制，以便完整记录从欠拟合到过拟合的全过程。

3 结果与分析

3.1 训练动态分析

图 2 展示了训练过程中的损失下降曲线。仔细观察曲线形态，可以将其大致划分为三个阶段，反映了模型学习的不同状态：

1. **快速下降阶段（Early Learning）**：在训练初期，损失函数值迅速下降。这是因为模型从随机初始化状态开始，迅速捕捉到了数据中的低频主成分（即正弦波的整体趋势和均值）。此时梯度较大，学习效率最高。

2. **平台期（Plateau）**：随后，损失曲线进入一个相对平缓的区域。这通常意味着模型陷入了优化地形的鞍点（Saddle Point）或局部极小值附近。此时，模型已经学会了“容易”的正弦规律，但尚未找到进一步降低误差（即拟合高频噪声）的有效参数路径。此外，Tanh 和 ReLU 激活函数的混合使用也可能导致部分神经元进入饱和区，暂时减缓了梯度的传播。

3. **继续下降阶段（Overfitting）**：随着训练的持续，优化器（AdamW）积累了足够的动量或找到了参数空间的特定方向，使得模型突破了平台期。此时损失的进一步降低并非源于对真实信号的更好拟合，而是模型开始利用其过剩的参数容量，精细调整权重以“死记硬背”个别的随机噪声点。这一阶段标志着过拟合的加剧，模型从“拟合信号”转向了“拟合噪声”。

这种“先快、后慢、再下降”的现象是过拟合过程的典型特征。

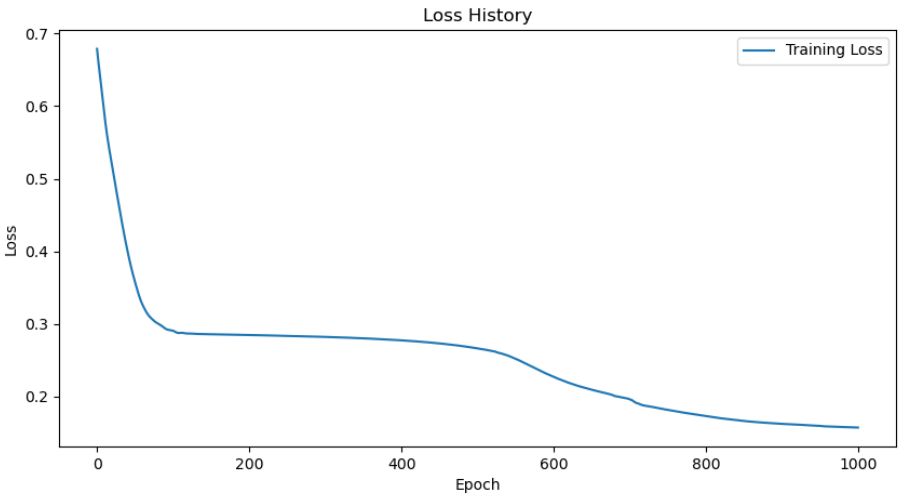

图 2: 训练损失（MSE）随 Epoch 变化的曲线。

3.2 拟合结果定性评估

图 3 展示了模型在 $x \in [0, 1]$ 区间上的预测曲线（虚线）与真实正弦曲线（实线）的对比。
观察与分析：

- **局部波动**：预测曲线并非平滑的正弦波，而在多个局部区域出现了剧烈的震荡和扭曲。
- **噪声拟合**：模型试图穿过那些偏离真实正弦曲线较远的噪声点。例如，当某个区域的噪声使数据点向上偏移时，预测曲线也随之向上突起。

这种现象直观地证实了过拟合的发生：模型将高频的随机噪声误认为是低频的信号特征进行了学习。

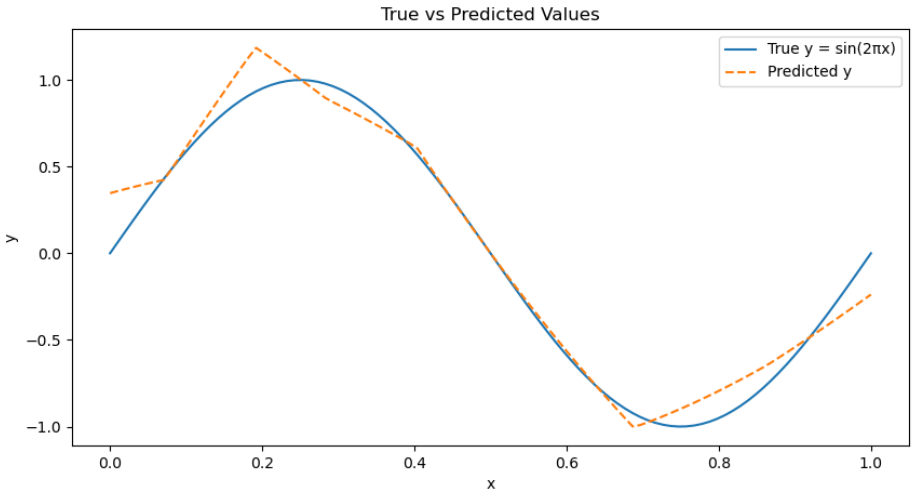

图 3: 模型预测结果对比：虚线展示了模型在过度训练后的拟合形态，可见明显的过拟合特征。

4 噪声类型与强度对模型的影响分析

为了进一步探究噪声特性对模型训练的影响，我们设计了对比实验，分别在输入端（x）和输出端（y）引入不同程度的高斯噪声。实验设置了四种场景，结果如图 8 所示。

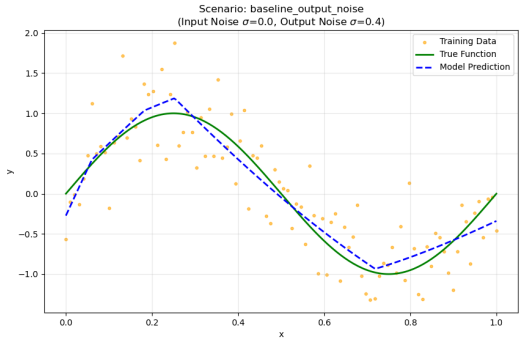

图 4: *
场景 A: 基准输出噪声 ($\sigma_y = 0.4$)

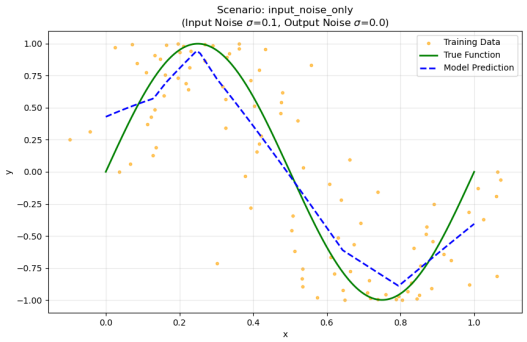

图 5: *
场景 B: 仅输入噪声 ($\sigma_x = 0.1$)

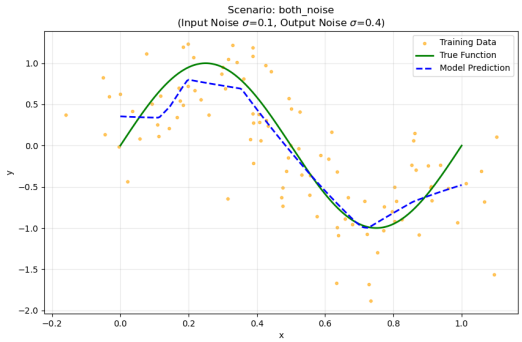

图 6: *
场景 C: 混合噪声 ($\sigma_x = 0.1, \sigma_y = 0.4$)

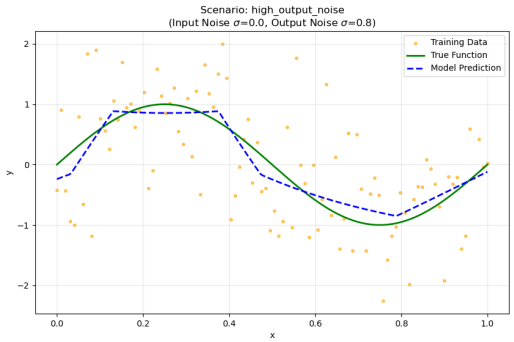

图 7: *
场景 D: 高强度输出噪声 ($\sigma_y = 0.8$)

图 8: 不同噪声配置下的模型拟合结果对比

4.1 结果分析

1. **输出噪声（Label Noise）的影响（场景 A vs D）：**
   - 当仅存在适度的输出噪声（场景 A）时，模型表现出明显的过拟合，预测曲线呈现出试图捕捉噪声点的“摆动”。
   - 当输出噪声强度增加到 $\sigma = 0.8$（场景 D）时，数据分布变得极度混乱，信噪比极低。此时模型为了最小化 MSE，被迫在数据点之间剧烈震荡，导致预测曲线完全失去了正弦波的平滑特征，泛化能力几乎为零。这说明**标签噪声是导致过拟合的主要驱动力**，且噪声越大，模型越容易迷失在随机性中。
2. **输入噪声（Input Noise）的影响（场景 B）：**
   - 在仅有输入噪声的情况下（场景 B），虽然 x 轴上的采样点发生了偏移，但 y 值依然严格遵循 $y = \sin(2\pi x_{true})$（注：此处模拟的是观测误差，即我们观测到的 x 是有误的，但对应的 y 是基于真实 x 产生的）。

- 有趣的是，模型在这种情况下生成的预测曲线相对平滑，并未出现严重的过拟合。这是因为输入噪声在某种程度上起到了**数据增强（Data Augmentation）**的作用，相当于在真实流形附近进行了“抖动”，迫使模型学习更鲁棒的特征，而非死记硬背具体的坐标点。

3. **混合噪声的影响（场景 C）：**

- 当同时存在输入和输出噪声时，任务难度最大。模型不仅要应对标签的随机性，还要处理输入坐标的不确定性。结果显示，预测曲线既有整体趋势的偏移，又有局部的剧烈波动，整体拟合效果最差。

4.2 结论

实验表明，**输出端的标签噪声对模型的危害远大于输入端的观测噪声**。高强度的标签噪声会直接诱导模型过拟合，而适度的输入噪声在某些条件下甚至能起到正则化的效果，提升模型的鲁棒性。

5 模型外推能力探究

为了评估模型在训练数据分布之外的泛化能力（即外推能力），我们将预测范围从训练区间 $[0, 1]$ 扩展到了 $[-1, 2]$。实验结果如图 9 所示。

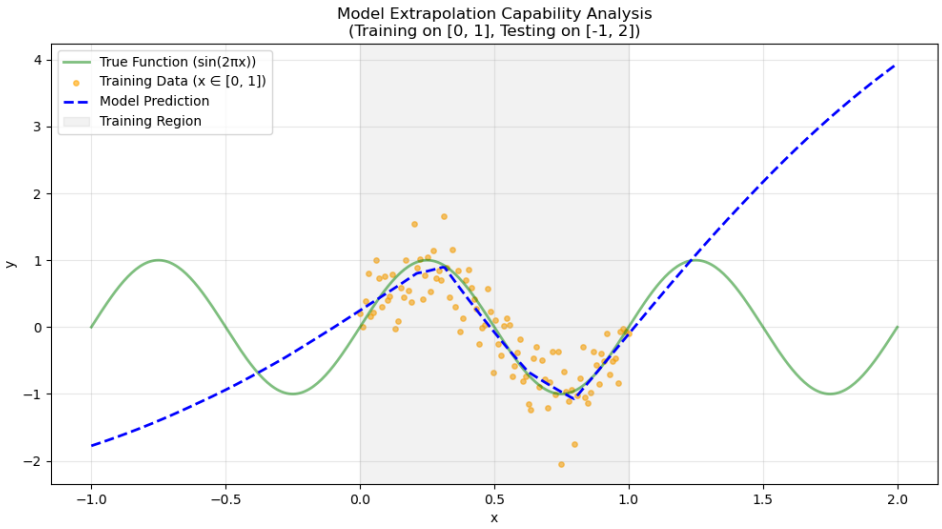

图 9: 模型外推能力分析：灰色区域为训练数据覆盖范围 $[0, 1]$，两侧为外推区域。

5.1 结果分析

从图中可以清晰地观察到：

- **内插表现（Interpolation）**：在训练区间 $[0, 1]$ 内（灰色背景区域），模型能够较好地拟合数据的整体趋势（尽管存在过拟合噪声的现象），预测曲线大致跟随正弦波波动。
- **外推表现（Extrapolation）**：一旦超出训练区间（即 $x < 0$ 或 $x > 1$），模型的预测能力迅速失效。

– 在 $x \in [1, 2]$ 区间，模型并没有重复正弦波的周期性规律，而是呈现出某种线性的延伸趋势。这是由 **ReLU 激活函数的性质**决定的：ReLU 网络本质上是分段线性函数（Piecewise Linear Function）。在远离训练数据的区域，激活状态保持不变，导致输出退化为线性函数。

– 在 $x \in [-1, 0]$ 区间，情况类似，模型无法预测出未曾见过的周期性变化。

5.2 结论

实验结果有力地证明了**多层感知机（MLP）通常缺乏外推能力**。MLP 本质上是一个复杂的函数逼近器，它通过组合非线性激活函数来“记忆”和“插值”训练数据分布内的映射关系。对于周期性函数（如正弦波），除非显式地在特征工程中引入周期性特征（如 $\sin(x), \cos(x)$ 作为输入），否则标准的 MLP 无法自动学习到“周期性”这一归纳偏置（Inductive Bias）。它只能保证在训练数据覆盖的流形附近给出合理的预测，而在远离训练数据的区域，其行为是不可控且不可靠的。

6 超参数敏感性分析

为了全面理解模型行为，我们系统地研究了四个关键超参数对训练动态和预测性能的影响：隐藏单元数量、数据采样密度、激活函数类型以及网络深度。

6.1 隐藏单元数量的影响

我们对比了隐藏单元数量为 4、32 和 128 时的模型表现（图 10）。

- **欠拟合（Hidden=4）**：模型容量不足，无法捕捉正弦波的非线性特征，预测曲线过于平直。
- **适中（Hidden=32）**：能够较好地拟合正弦波，但也开始表现出对噪声的敏感性。
- **过拟合（Hidden=128）**：模型容量过大，预测曲线在数据点之间剧烈震荡，试图穿过每一个噪声点，泛化能力最差。

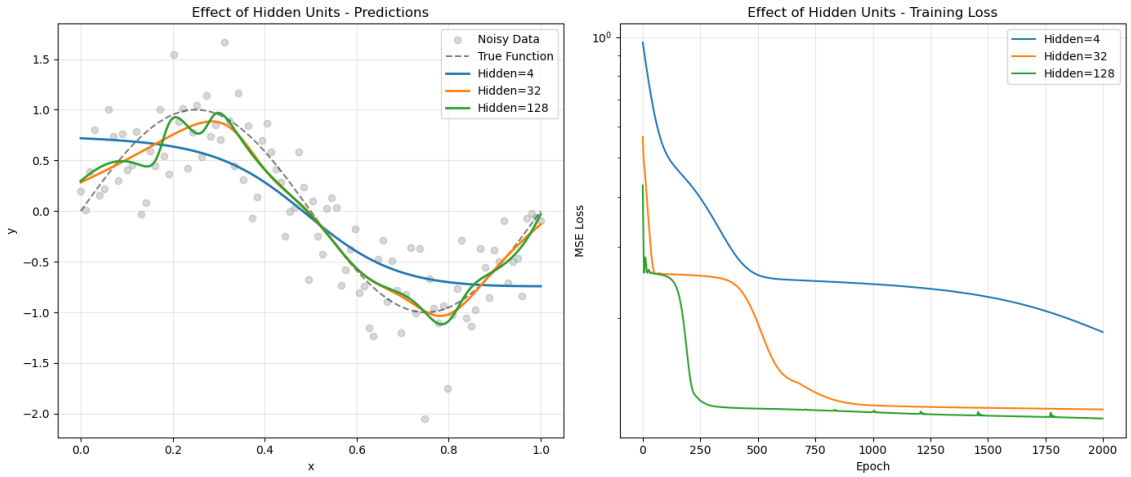

图 10: 不同隐藏单元数量对模型拟合（左）和训练损失（右）的影响

6.2 数据采样密度的影响

我们对比了训练样本数量 N 为 20、100 和 500 时的效果（图 11）。

- **稀疏数据（N=20）**：模型缺乏足够的信息来重建正弦波，容易在数据空隙处产生错误的插值。
- **中等数据（N=100）**：模型能够捕捉整体趋势，但仍受噪声干扰。
- **密集数据（N=500）**：随着样本量的增加，噪声的影响被平均化（大数定律），模型能够学习到更平滑、更接近真实函数的曲线，过拟合现象显著减轻。

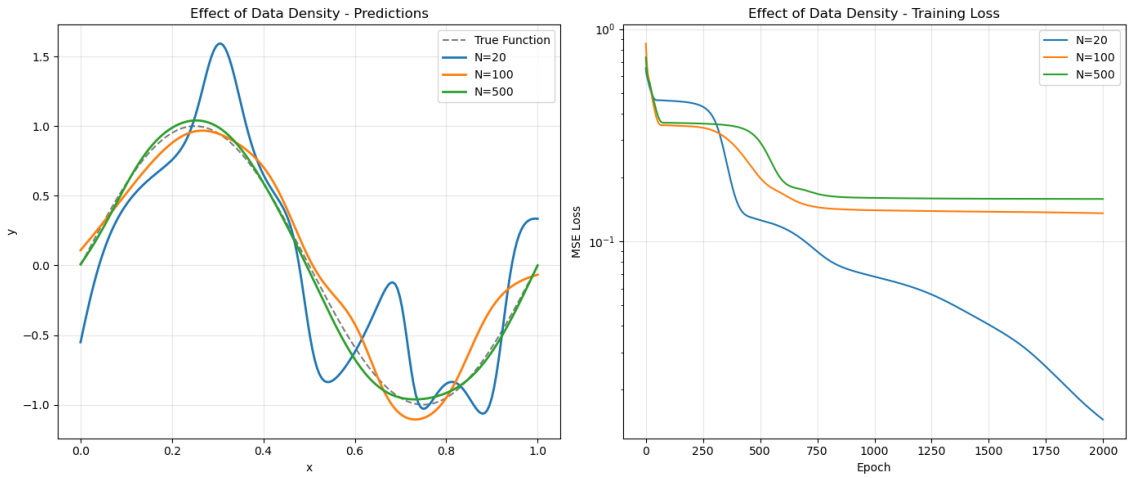

图 11: 不同数据采样密度下的模型表现

6.3 激活函数类型的影响

我们对比了 ReLU、Tanh 和 Sigmoid 三种激活函数（图 12）。

- **ReLU**：拟合能力强，收敛速度快，但生成的曲线由分段线性函数组成，显得较为折线化（Jagged）。
- **Tanh**：由于其平滑且中心化（Zero-centered）的特性，非常适合拟合正弦波这类平滑函数，生成的曲线最自然。
- **Sigmoid**：在深层网络中容易出现梯度消失问题，导致收敛较慢或陷入局部极小值，拟合效果在此实验中略逊一筹。

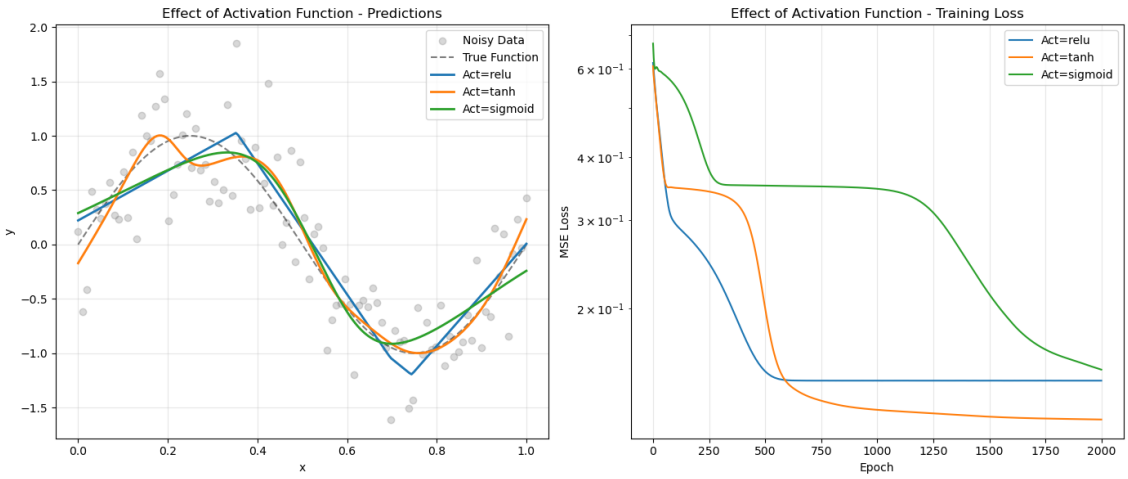

图 12: 不同激活函数对模型拟合（左）和训练损失（右）的影响

6.4 网络深度的影响

我们对比了隐藏层层数为 1、3 和 6 时的模型（图 13）。

- **浅层网络（Depth=1）**：虽然只有一层，但只要隐藏单元足够，根据通用近似定理，仍能拟合正弦波，但可能需要更多参数。
- **中等深度（Depth=3）**：表现均衡，能够学习复杂的非线性映射。
- **深层网络（Depth=6）**：在此简单任务中显得多余。过深的网络不仅难以训练（梯度传播困难），而且更容易过拟合，导致预测曲线出现不必要的波动。

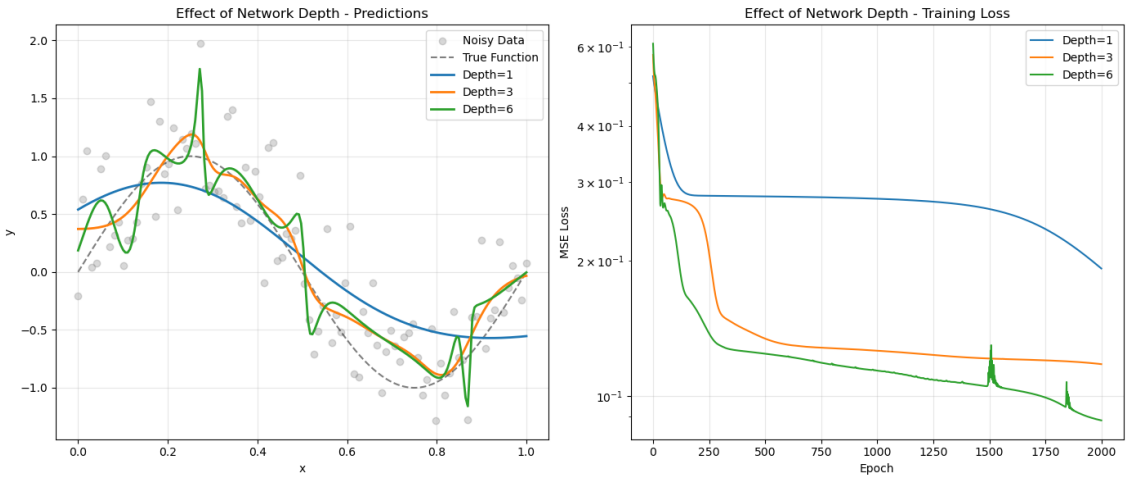

图 13: 不同网络深度对模型拟合（左）和训练损失（右）的影响

7 讨论

7.1 偏差-方差权衡（Bias-Variance Tradeoff）

本实验的系列结果生动地展示了机器学习中的核心概念——偏差-方差权衡。

- **高偏差（High Bias）**：如 Hidden=4 的模型，因容量不足而无法捕捉数据的真实形态，导致欠拟合。
- **高方差（High Variance）**：如 Hidden=128 或 Depth=6 的模型，因容量过大而对训练数据中的随机噪声过于敏感，导致过拟合。

理想的模型应当处于两者的平衡点（如 Hidden=32），既能捕捉信号的主体趋势，又能忽略随机噪声的干扰。

值得注意的是，近年来的研究（如 Deep Double Descent 现象）指出，在参数量极大的过参数化（Over-parameterized）区间，测试误差可能会再次下降。然而，在本实验的小样本（$N = 100$）与中等规模模型（Params ≈ 1000）设置下，我们主要观察到了经典的“U 型”泛化误差曲线，即过拟合导致性能恶化。

7.2 过拟合的本质与成因

实验表明，过拟合的本质是模型将训练数据中的“特异性”（噪声）误认为是“普遍性”（规律）。其主要成因包括：

1. **模型容量过剩**：参数数量远多于独立样本数量，使得模型有能力“记住”每一个样本。
2. **数据信噪比低**：高强度的标签噪声（如 $\sigma_y = 0.8$）掩盖了真实信号，误导了优化方向。
3. **缺乏归纳偏置**：MLP 作为通用逼近器，缺乏对特定问题（如周期性函数）的先验假设，导致外推能力缺失。

7.3 改进策略与未来方向

针对上述问题，未来的研究和实践可以从以下几个维度进行改进：

1. **正则化技术**：引入 L1/L2 正则化、Dropout 或 Batch Normalization，显式地限制模型复杂度。
2. **数据增强与扩充**：通过增加样本量或引入对抗样本（Adversarial Examples）来提高模型的鲁棒性。
3. **结构设计**：针对特定任务设计具有归纳偏置的网络结构，例如使用循环神经网络（RNN）处理序列数据，或在输入特征中加入周期性编码（如 $\sin(x), \cos(x)$）。
4. **早停法（Early Stopping）**：利用验证集监控训练过程，在泛化误差开始上升时及时停止训练。

8 结论

本报告通过一系列控制变量实验，全面分析了多层感知机在拟合噪声数据时的行为特征。我们证实了模型容量、数据规模、噪声特性以及超参数设置均对模型的泛化能力有着决定性影响。实验结果表明，单纯追求训练误差的最小化并不能保证模型的泛化性能，甚至可能导致严重的过拟合。在实际应用中，必须综合考虑数据特性和任务需求，合理设计模型结构，并采用适当的正则化手段，以构建既准确又鲁棒的深度学习系统。