哈爾濱工業大學

模式识别与机器学习实验报告

实验 一

题　　目	多项式拟合正弦函数实验
学　　院	未来技术学院
专　　业	人工智能
学　　号	2023112419
学　　生	陈铠
任课教师	刘扬

哈尔滨工业大学计算机科学与技术学院

2025 年秋季

一、 实验内容

（一）样本数据生成与模型建立

本次实验拟合的目标函数为正弦函数：

$$y = sin(x), x \in [0,2\pi]$$

首先在区间$[0,2\pi]$上均匀采样 N 个点作为输入数据：

$$x_i = \frac{2\pi}{N-1} \cdot i, i = 0,1,\dots,N-1$$

真实输出为

$$y_i^{true} = \sin(x_i)$$

然后在真实值上加入均值为 0、方差为σ^2的高斯噪声：

$$y_i = y_i^{true} + \epsilon_i, \epsilon_i \sim \mathcal{N}(0,\sigma^2)$$

1. 数据归一化与设计矩阵构建

拟合函数为

$$f(x;w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d = \sum_{j=0}^{d} w_j x^j$$

将所有训练点$x_i, y_i (i = 1,\dots,N)$使用设计矩阵表示：

$$\Phi(x_i') = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & & & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^d \end{bmatrix} \in \mathbb{R}^{N\times(d+1)} \; i \in [1,N] \; \mathbf{w} = (w_0,\dots,w_d)^\top$$

那么预测为$y_i = \Phi(x_i')\mathbf{w}$。而在实际求解运算时，随着拟合函数阶数提高，设计矩阵的数值也会增大，不利于求解方程。为了提高数值稳定性，将输入数据线性映射到$[-1,1]$区间：

$$x_i' = \frac{2x_i - (a+b)}{b-a}, a = 0, b = 2\pi$$

再利用多项式基展开构造设计矩阵：

$$\Phi(x_i') = [1, x_i', (x_i')^2, \dots, (x_i')^d]$$

其中 d 为多项式的阶数。

2. 损失函数设计

损失函数采样均方误差（MSE）：

$$E(w) = \frac{1}{N}\sum_{i=1}^{N} (y_i - \Phi(x_i')\mathbf{w})^2$$

若考虑 L2 正则化，则误差函数为；

$$E(w) = \frac{1}{N}\sum_{i=1}^{N} (y_i - \Phi(x_i')\mathbf{w})^2 + \lambda \sum_{j=1}^{d} \mathbf{w}_j^2$$

其中正则化系数$\lambda = 10^{-3}$，且不对偏置项w_0正则化

3. 优化方法

3.1 闭式解（Ridge 回归）

通过解析解可直接求得最优权重向量，令梯度为 0：

$$\nabla_w E(w) - \frac{2}{N} X^{\top}(\mathbf{y} - X\mathbf{w}) + 2\lambda R\mathbf{w} = 0$$

化简得

$$X^{\top}X\mathbf{w} + N\lambda\mathbf{w} = X^{\top}\mathbf{y}$$

于是得到解析解

$$w^* = (\Phi^T\Phi + N\lambda R)^{-1}\Phi^T y$$

其中R为正则化矩阵，目的是只约束权重参数，而不约束偏置项w_0

3.2 梯度下降法

对权重$\mathbf{w}$求梯度：

$$\nabla_w E(w) = -\frac{2}{N}\Phi^T(y - \Phi w) + 2\lambda w$$

则权重更新公式为：

$$w^{(t+1)} = w^{(t)} - \eta\nabla_w E(w^{(t)})$$

式中η为学习率。

实际中学习率的确定较难抉择，因此本实验采用自适应学习率 Adam（Adaptive Moment Estimation）优化器。Adam 通过引入一阶动量（梯度的指数加权平均）和二阶动量（梯度平方的指数加权平均），自适应调整每个参数的学习率，从而在收敛速度和稳定性上优于标准梯度下降。Adam 更新公式为：

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2$$
$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$
$$w_{t+1} = w_t - \eta\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

其中$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 10^{-8}$。

3.3 共轭梯度法（Conjugate Gradient, CG）[1]

当求解对称正定线性方程组$A\mathbf{w} = \mathbf{b}$（本实验中$A = \Phi^{\top}\Phi + N\lambda R$，$\mathbf{b} = \Phi^{\top}\mathbf{y}$）时，可以将其转化为二次函数极小化问题：

$$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\top}A\mathbf{w} - \mathbf{b}^{\top}\mathbf{w}$$

共轭梯度法通过在逐步扩大的 Krylov 子空间上做局部精确最小化来求解该二次问题。算法维护残差向量$\mathbf{r}_k = \mathbf{b} - A\mathbf{w}_k$和一组 A-共轭（A-orthogonal）的搜索方向$\mathbf{p}_k$做一维最

小化，更新公式如下：

$$\alpha_k = \frac{\mathbf{r}_k^\top \mathbf{r}_k}{\mathbf{p}_k^\top A \mathbf{p}_k}, \mathbf{w}_{k+1} = \mathbf{w}_k + \alpha_k \mathbf{p}_k, \mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k A \mathbf{p}_k$$

$$\beta_{k+1} = \frac{\mathbf{r}_{k+1}^\top \mathbf{r}_{k+1}}{\mathbf{r}_k^\top \mathbf{r}_k}, \mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta_{k+1} \mathbf{p}_k$$

在精确算术下，CG 在至多d步（d为自由度）内收敛到精确解，而在浮点运算下通常更早达到足够小的残差。与直接求解（如 LU 分解/Cholesky 分解）相比，CG 的优点在于：若矩阵A稀疏或有效乘法 $A\mathbf{p}$代价低，则每一步计算代价低且内存需求小；与梯度下降相比，CG 利用了 A-共轭方向，使得收敛速度通常更快，尤其是在条件数较好的情况下。

（二）不同数据量、超参数、阶数变量设计

表 1 本实验使用工具包

研究环节	固定变量	变化变量
多项式阶数的影响	$N = 10, \lambda = 0, \sigma^2 = 0.3$	$d = 3,5,7,9,11,13$
样本数量的影响	$d = 9/11, \lambda = 0, \sigma^2 = 0.3$	$N = 10,30,50,100$
正则化参数的影响	$N = 10, d = 11, \sigma^2 = 0.3$	$\lambda = 10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 0.1,0,0.5,1.0$
优化方法的影响	$d = 11, \sigma^2 = 0.3$	$N = 10/15, \lambda = 0/10^{-3},$ 方法 $=$ 闭式解/梯度下降/CG

二、 实验环境

操作系统：Windows 11
实验平台：pycharm
解释器版本：Python 3.12
工具包：

表 2 本实验使用工具包

工具包名称	版本
pip	25.2
matplotlib	3.10.6
numpy	2.3.3

三、 实验结果及分析

（一）多项式阶数对拟合结果的影响

固定样本数 N=10，不对误差函数作正则化（lam=0.0），噪声方差为$\sigma^2 = 0.3$，仅改变多项式阶数的拟合结果图如下（degree 分别取 3，5，7，9，11，13）：

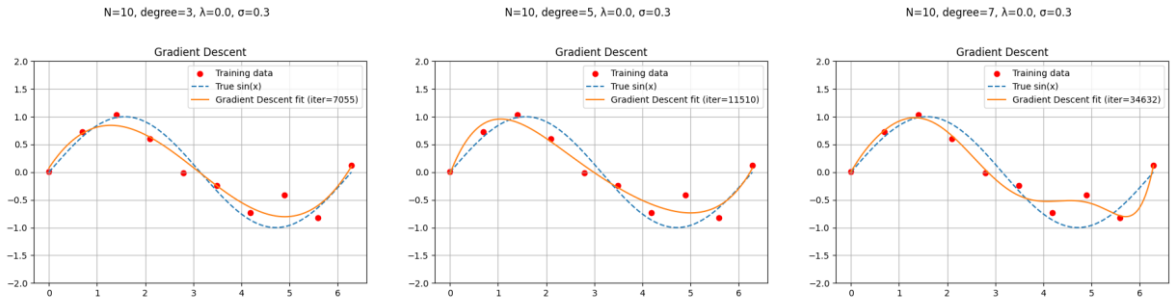

图 1 degree = {3, 5, 7}

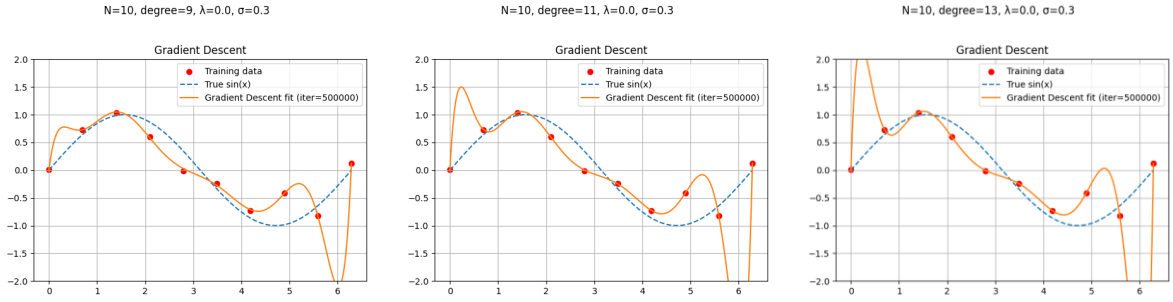

图 2 degree = {9, 11 13}

阶数$d = 3$时，曲线平滑，整体趋势接近$sin(x)$，但在波峰、波谷处偏差较大，略显欠拟合现象，迭代次数为 7055，主要受模型复杂度不足的影响，无法捕捉$sin(x)$的周期性和细节。

阶数$d = 5 - 7$时，曲线贴合训练样本，同时与$sin(x)$的偏差小，没有明显的振荡，拟合情况较优，迭代次数 11510-34632，模型复杂度与样本信息量匹配，既能拟合样本又不被噪声干扰。

阶数$d = 9 - 13$时，曲线在样本点间剧烈振荡，训练误差极小但偏离$sin(x)$严重，出现过拟合现象，迭代次数 500000 达到训练上限，模型复杂度过高，拟合能力极强，因此可以通过极端系数来实现震荡以拟合所有的样本点。对于这种情况的过拟合可以通过增加样本数量或对损失函数添加正则项来改善。

（二）样本数量对拟合结果的影响

选取多项式阶数为$degree = \{9，11\}$，不对误差函数作正则化（lam=0.0），噪声方差为$\sigma^2 = 0.3$。令样本数分别为$N = \{10,30,50,100\}$：

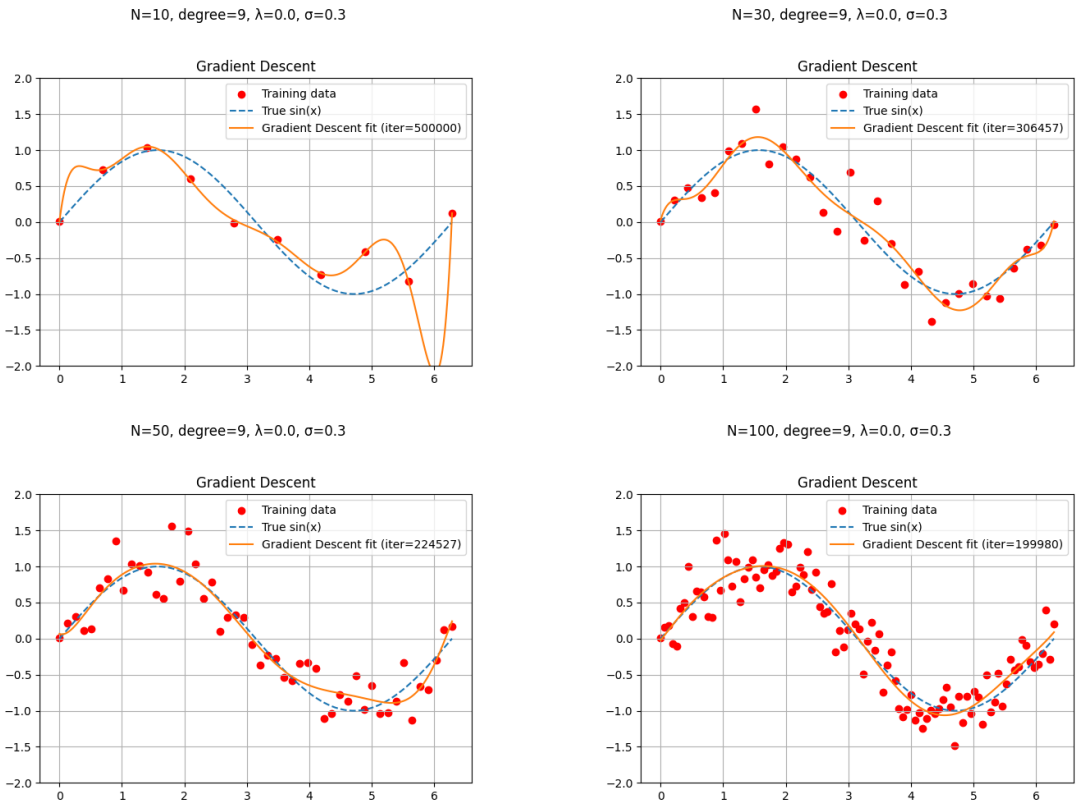

图 3 degree = 9, N = {10, 30, 50, 100}

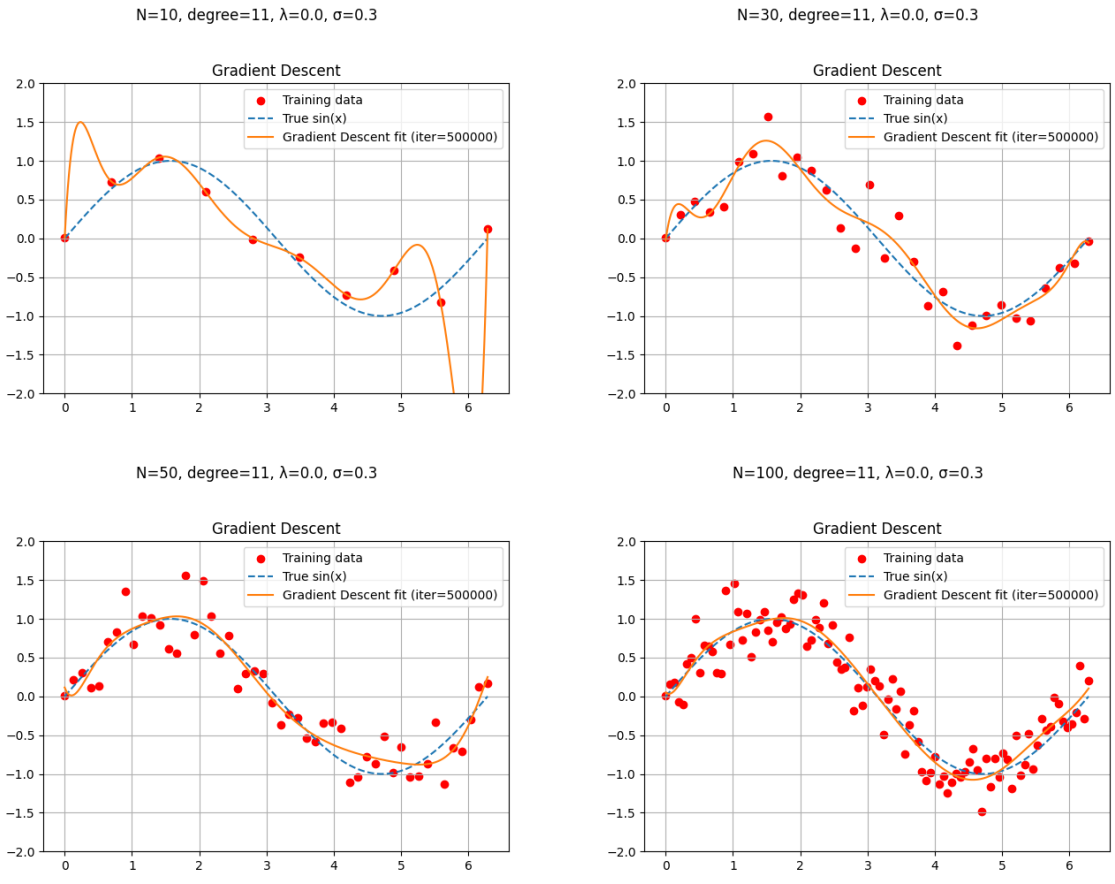

图 4 degree = 11, N = {10, 30, 50, 100}

1. 多项式阶数 d=9 时的拟合情况

样本量$N = 10$，曲线呈现剧烈震荡特征，过拟合现象严重。

样本量$N = 30$，曲线的震荡幅度显著减弱，整体形态接近目标函数 sin(x)，迭代次数 306457。样本量增加至 30 后，为多项式提供了更多信息，参数调整不再过度迎合个别噪声样本，调整过程更稳健，与$sin(x)$的偏差缩小。

样本量$N = 100$，曲线变得平滑，与目标函数$sin(x)$几乎完全重合，无过拟合迹象，迭代次数 199980。核心原因是：样本量充足时，大量样本中的噪声相互“平均抵消”，模型能够充分学习$sin(x)$的周期性与非线性细节，10 个参数被充分约束，无需通过极端调整拟合噪声，最终实现对$sin(x)$真实趋势的精准拟合。

2. 多项式阶数 d=11 时的拟合情况

样本量$N = 10$，曲线的震荡程度比 d=9、N=10 时更剧烈，过拟合现象更为严重。

样本量$N = 100$，曲线的震荡基本消失，整体形态紧密贴合目标函数 sin(x)，过拟合现象得到有效抑制，迭代次数仍为 500000 次，高于 d=9、N=100 时的 199980 次。核心原因是：11 次多项式的参数数量（12 个）多于 d=9 的 10 个参数，需要更多迭代步骤才能逼近最优参数，因此迭代次数高于相同样本量下的 d=9 模型。

（三）正则化参数对拟合结果的影响

令 N=10，degree=11，噪声方差为$\sigma^2 = 0.3$。正则化参数 lam 依次取$\lambda = 0, 10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-1}, 0.5, 1.0$

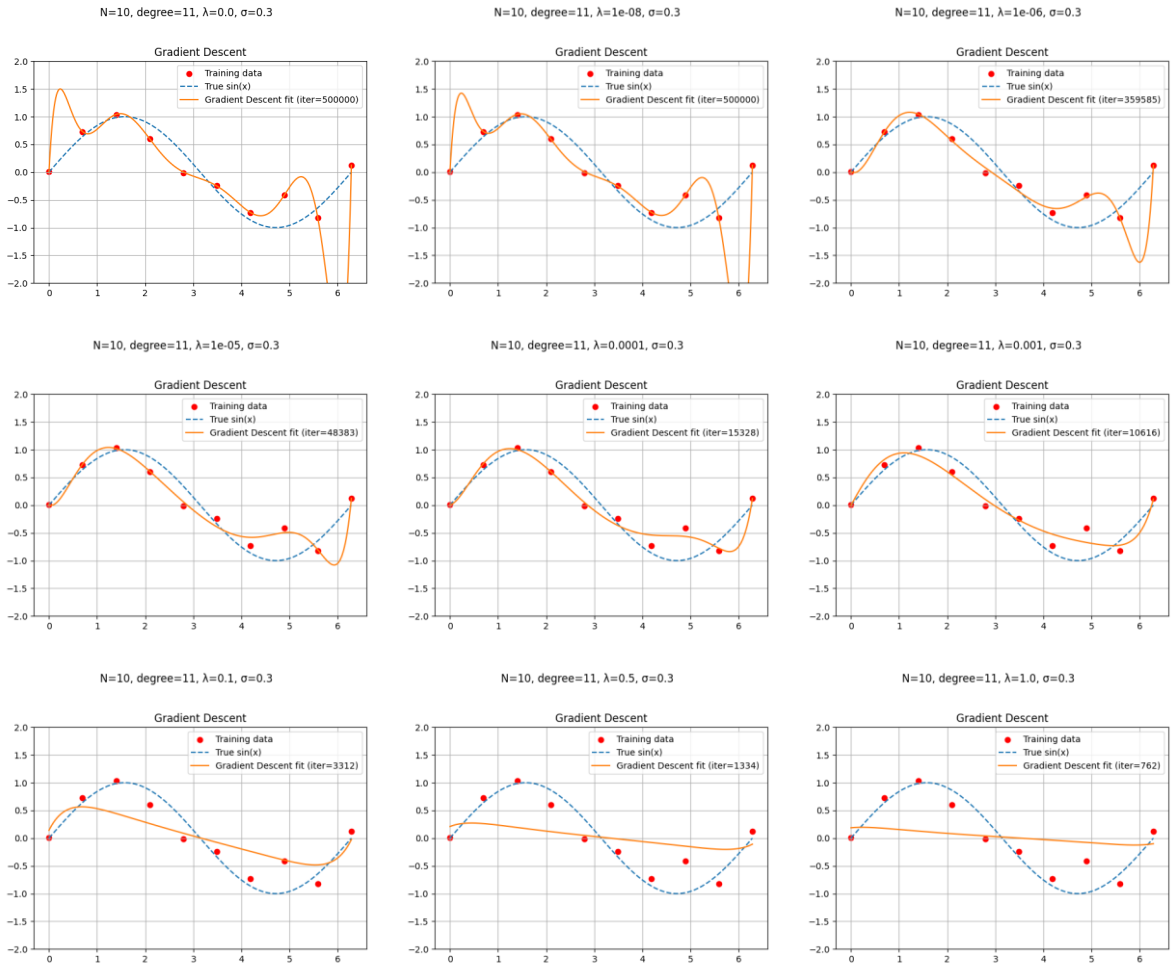

图 5 不同正则化参数拟合结果图

$\lambda = 0 - 10^{-8}$，曲线呈现剧烈震荡特征，过拟合现象严重。

$\lambda = 10^{-5} - 10^{-4}$，曲线的震荡幅度显著减弱，整体形态开始向目标函数$sin(x)$靠近，但样本点间仍存在小幅波动，过拟合现象减轻至“轻度过拟合”水平。迭代次数降至 15328-48383 次。

$\lambda = 10^{-3}$，曲线表现出平滑特征，能够贴合目标函数$sin(x)$，样本点间无明显震荡，拟合效果达到最优。迭代次数进一步降至 10616 次，收敛效率显著提升。核心原因在于：正则化参数 λ 的适中，既能够有效惩罚参数（$w_1 - w_{11}$）的极端数值，压制其过度适应噪声的行为，避免曲线震荡；又不会因惩罚过强导致参数被过度约束，保证模型仍具备学习$sin(x)$周期性与非线性细节的能力。

$\lambda = 0.1 - 1.0$，曲线接近“水平直线”（如 y≈0），与目标函数$sin(x)$的偏差极为严重，表现为严重欠拟合。迭代次数仅为 762-3312 次，虽收敛快但精度极差。原因在于：正则化参数 λ 过大，对多项式参数（$w_1 - w_{11}$）的惩罚强度过强，参数（$w_1 - w_{11}$）被过度压制至接近 0 的水平，多项式退化为仅含常数项w_0的“水平直线”，完全失去表达$sin(x)$非线性趋势的能力，最终导致曲线严重偏离目标函数。

（四）不同求解方法对拟合结果的影响

令 N={10，15}，degree=11，lam={0，10^{-3}}，所得拟合结果：

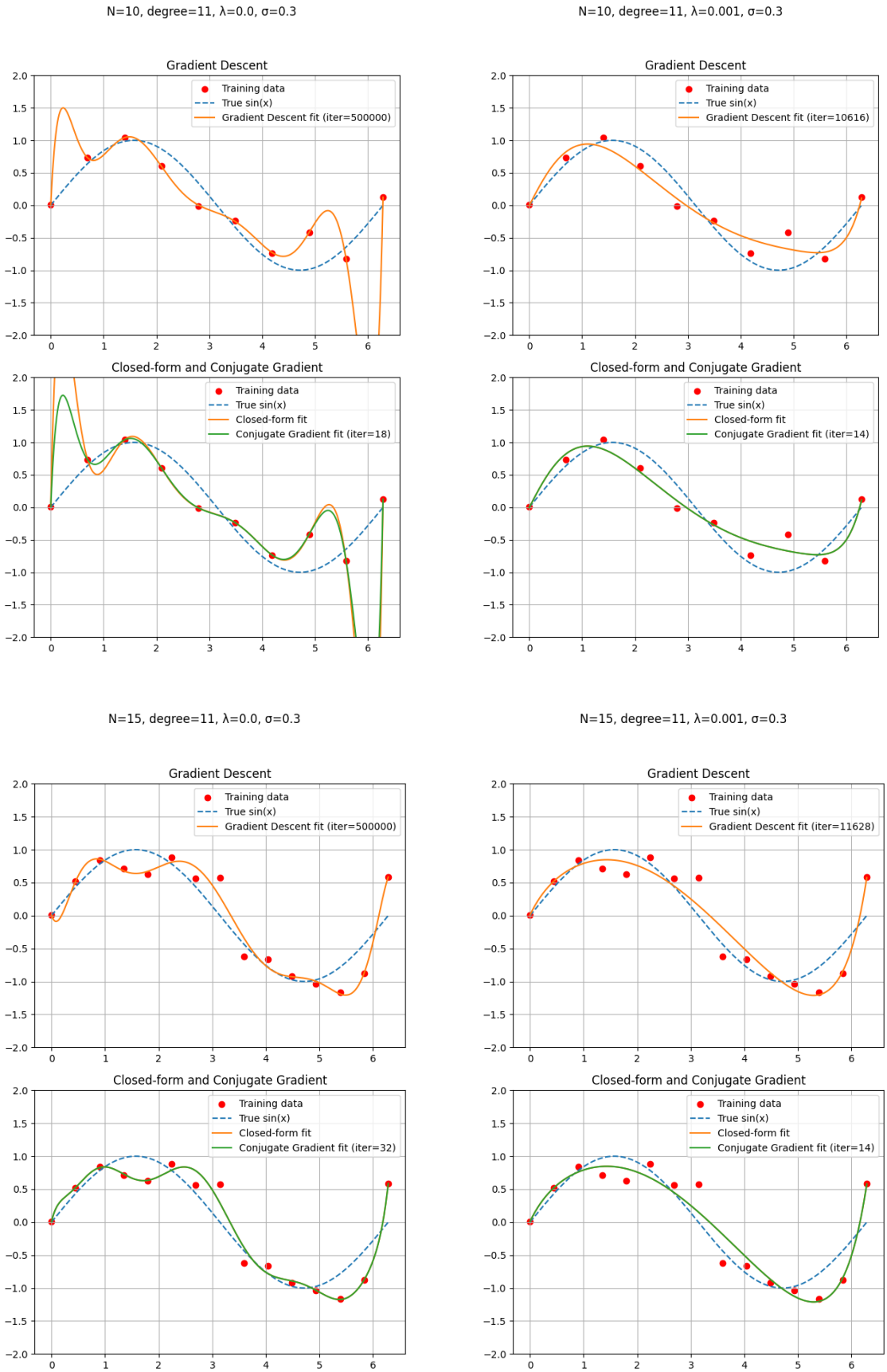

图 6 不同求解方法拟合结果图

表 3 不同求解方法迭代次数

参数	梯度下降迭代次数	共轭梯度迭代次数
$N=10,\lambda=0$	500000	18
$N=10,\lambda=0.001$	10616	14
$N=15,\lambda=0$	500000	32
$N=15,\lambda=0.001$	11628	14

三种求解方法（闭式解、梯度下降、共轭梯度法）在结果精度上差别不大，最终拟合曲线基本一致。但在效率方面表现出显著差异：梯度下降需要大量迭代才能收敛，收敛速度最慢且依赖学习率选择；共轭梯度法的迭代次数最少，能在远低于梯度下降的计算代价下达到与闭式解相同的效果，是本实验中效率最高的方法。综合来看，闭式解适合小规模问题，梯度下降适合复杂或大规模数据，而在当前实验场景下，共轭梯度法兼具精度和效率，表现最优。

四、 结论

本次实验通过控制变量法，系统分析了多项式拟合正弦函数的关键影响因素及优化方法特性，核心结论如下：

1.模型复杂度需与数据信息量匹配：多项式阶数d是决定拟合效果的核心：d过小（如$d<5$）导致欠拟合（无法捕捉$sin(x)$的非线性），d过大（如d>9）导致过拟合（学习噪声）；当$d=5\sim7$且$N=10$时，模型复杂度与样本信息量最优匹配，拟合效果最佳。

2.样本量可有效抑制过拟合：增加样本量可提升数据的真实趋势占比，当$d=9,N=100$时，过拟合现象基本消除；且样本量越大，模型对噪声的鲁棒性越强，泛化能力越好。

3.正则化参数对求解结果影响较大：L2 正则化的最优λ需平衡拟合精度与模型复杂度：本实验中$\lambda=10-3$是最优值（$d=11,N=10$），λ过小无法抑制过拟合，λ过大导致欠拟合；正则化同时能提升优化效率，降低迭代次数。

4.共轭梯度法在该类问题中效率最高，迭代次数远少于梯度下降；梯度下降适合大规模数据或高维复杂模型，无需矩阵求逆。

五、 参考文献

[1] Hestenes M R, Stiefel E. Methods of conjugate gradients for solving linear systems[J].
[2]https://en.wikipedia.org/wiki/Gradient_descent
[3]（美）SHELDON AXLER 著；杜现坤，刘大艳，马晶译. 线性代数应该这样学 第 3 版[M]. 北京：人民邮电出版社, 2016.10.
[4]周志华著. 机器学习[M]. 北京：清华大学出版社, 2016.01.
[5]谢文睿，秦州编著. 机器学习公式详解[M]. 北京：人民邮电出版社, 2021.03.
[6]李航著. 统计学习方法 第 2 版[M]. 北京：清华大学出版社, 2019.05.