第三章函数拟合实验总结

樊建琦

October 2024

1 Introduction

第二章、第三章的内容集中讨论函数逼近的问题，插值亦或拟合，或理论或应用，总归落在了拟合二字。实验三系统性的从随机采样到切比雪夫采样；从插值函数到拟合函数，最终得到 $P^*(x), P^*(x) \in H_n$. 为了评价拟合优劣，我们用 matlab 在画布中描出准确值 (gt) 与预测值 $(pred)$, 用二者的平均误差和整体重合程度评价拟合程度。但最小二乘中二范数对噪声有着本源的低抗性，所以我们在每个实验方法后都添加了扰动环节，考察引入扰动后，偏移程度是否能被接受。

2 Newton Interpolation

方案一选取随机采样点，方案二选取切比雪夫多项式零点作为采样点，采用牛顿插值法进行插值，画出插值函数和原函数曲线，对比分析插值结果的龙格现象差异。Fig.1-1,1-2 分别是采样方式为切比雪夫零点、随机取点，

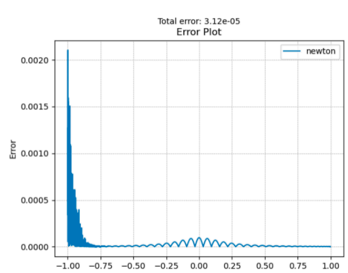


Fig1-1:Chebyshev

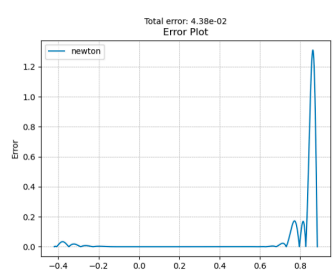
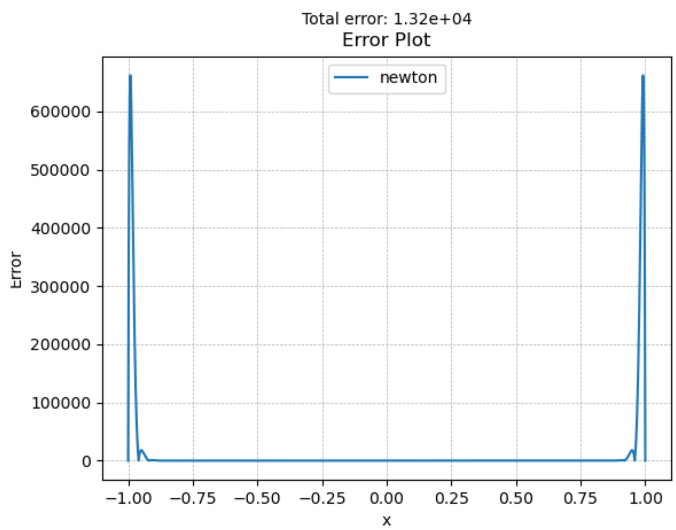
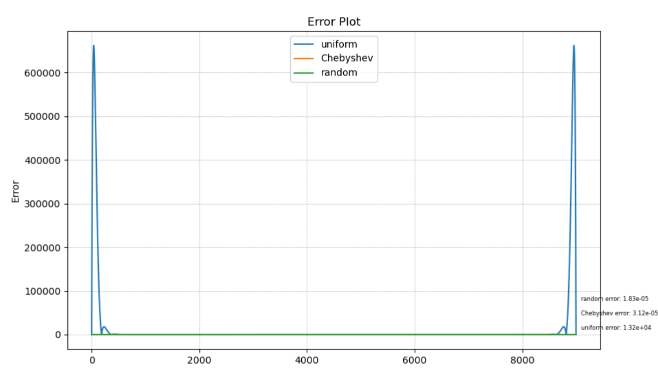


Fig1-2:Random

可以看到，这两种采样方式得到的误差都相对较小，切比雪夫零点的误差小是因为它把 $R_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x-x_i)$ 中的 $\prod_{i=0}^{n}(x-x_i)$ 取到了最小。但是均匀采样如Fig.1-3所示，误差极大，一方面是为了迎合多变的目标函数，多项式最高次数必须有一定的规模，因此导致的边界龙格效应，另一方面是因为由于均匀分布的采样点在区间边界附近的密度较低，导致误差项在区间边界附近较大，从而出现龙格现象。



三种方法对比图如Fig.1-3 所示，切比雪夫与随机取样显然优于均匀采样，且龙格效应被大幅度抑制。



3 Least Square

a,b 为输入参数区间，目标函数形如 $c \cdot \sin(d \cdot x) + e \cdot \cos(f \cdot x)$，参数 n 作为采样点的个数，参数 m 作为实验点的个数。全部使用切比雪夫正交多项式作为基函数以提高方法稳定性。

3.1 Global Fitting

在./data/prob2/prob2.21/prompts.json 中可以看到未加扰动的原始输入，误差与函数对比由 Fig2.1 所示

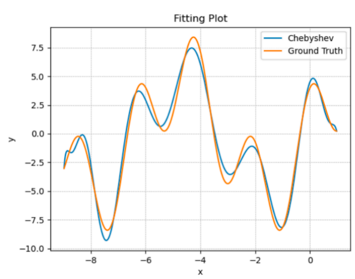


Fig2.1 Contrast between gt and pred

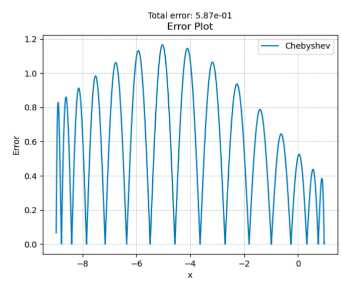


Fig2.2 sub error

此时误差处于可控范围内有两个原因，首先我们拟合的最高次数选的大，而且不过大，所以他有了足够又不过多的伸缩许可，其次所有原始输入均未引入扰动。

为了测试野点对于二范数拟合的干扰程度，我们引入两种野点，第一种是全部点集在进入目标函数法则后都加入微小扰动，第二种是只引入少部分点但是扰动幅度大，这两种测试分别在代码的./data/prob2/prob2.21 及.../prob2.22 存有详细的原始输入信息及测试结果。Fig.2.3, Fig.2.4 是两种干扰实施后的函数值对比.

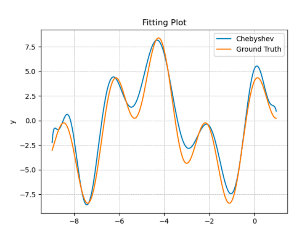


Fig2.3 pertubate all

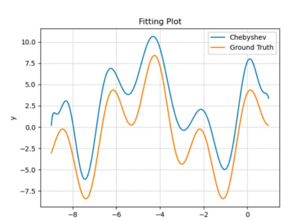


Fig2.4 pertubate a few

当引入少点数大扰动时，二范数的低抗噪性展露无遗。为了达到最小的最大误差，这种方法不可避免的存在着这样的局限。

3.2 Local Fitting

我们一直追求采样点与最高次数的独立性，能够解决这个依赖关系的只有“分段低次”。我们考虑段数为 2，4，8 下拟合情况，其中每个段数遍历最高次数 [1,2,3,4]。同样，我绘制了很多详尽的图像，对每一种情况都进行了绘制，每一张图片都包含扰动点位置、gt 位置、第 k 端的误差、目标函数走势、拟合函数走势...，均保存在./data/prob3/。为了更好的说明分段的优越性，我们以拟合最高次数来区别每一个图像，因此有四个图像分别对应了所有分段情况下的一次拟合（线性拟合）、二次拟合、三次拟合以及四次拟合，如Fig:3-1,3-2 Fig:3-3,3-4.

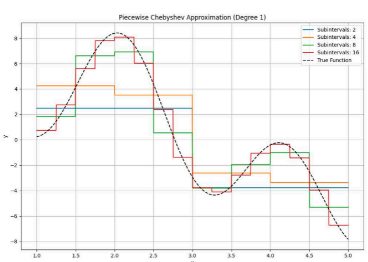


Fig3.1 线性拟合

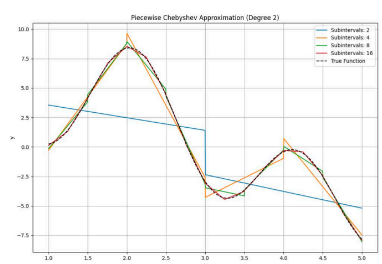


Fig3.2 抛物线拟合

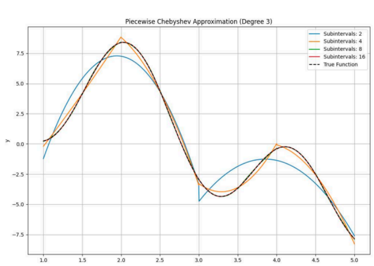


Fig3.3 三次拟合

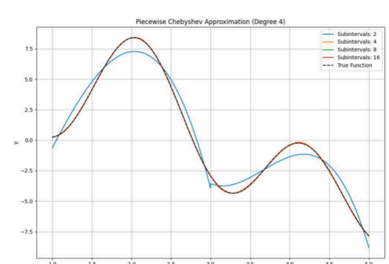


Fig3.4 四次拟合

当引入扰动时，分段多次数小抗噪性略强，拟合度略高，如Fig.3-5所示，从这里我们也可以看出分段的优势，减少了龙格的冗余权衡，用最简单的办法拟合最短的区域是非常好的想法。

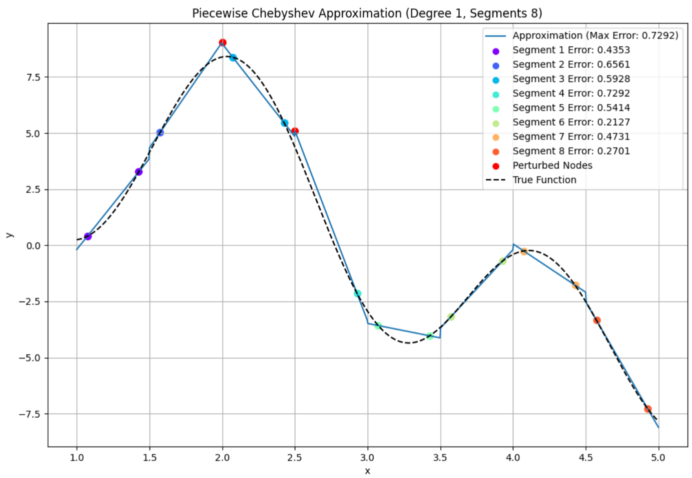


Fig.3.5 degree:1,segments:8

另,我绘制了所有情况下扰动后的详细图如Fig.3-5于./data/prob3/new_prob3.2/，但是次数高了，加上扰动的存在，还是容易出现龙格效应，如Fig.3-6所示

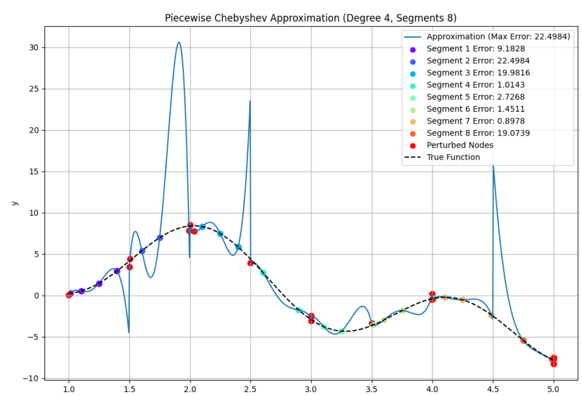


Fig.3.6 degree:4,segments:8

4 RANSAC

算法简述：我们假设有一个池子，里面是 100 个采样点，这些采样点得到的 y 都带有噪声而且噪声程度不确定（代码用高斯分布模拟），我们在预知图像基本分布后决定一组用 7 个采样点（这个很重要），于是我们遍历这 $\binom{100}{7}$ 个独异的组合，每次遍历都会剩余 93 个点，这 93 需要经过 loss 函数判断，如果此时的 x 得到的 pred_y 和 gt_y 的损失可接受（落在阈值以里）那么可信度 ++，最后选出来可信度最高的一个组合推到的模型，这个模型就是我代码里面 fitting_nodes() 得到的一组二元匿名函数 ($S(x) = \sum_{k=0}^{m} c_k T_k(x)$,S 和 T 都已知了用高斯消元解方程即可) 其中耐人寻味的是 loss 函数的设计，一开始我很朴实的用同 x 下的两个 y 相减，但是后来发现这对于 x 方向上偏移量大的曲线来讲误差极大。现构造一函数

$$h(x) = (x - x_0)^2 + (f(x) - y_0)^2$$

$h(x)$ 最小就相当于让曲线上任意一个点距离给定的 (x_0, y_0) 最小了，一步一步走最快的下坡路迭代即可。这里如若选择 `from scipy.optimize import minimize_scalar` 让他快速收敛到最小值的话，可以得到非常好的去噪效果，如Fig.4-1所示，此时内点数仅有 30%，阈值控制在 1，采样点为 7。

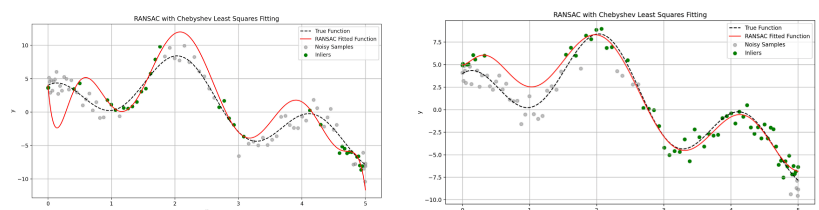


Fig3.3 三次拟合　　Fig3.4 四次拟合

Fig.4-1 left: normal-loss, right: min_scalar