Linear Regression with Legendre Polynomials

Answers

a)

We need to normalize because levelling the noise would let σ^2 to have real amount of the noise. Without doing this, our function would discover only error impacted pattern but not the actual behavior of the function. When we look at the suggested solutions for normalizing[2], we can deduct that mean = 0, standard deviation = 1 works well for populations that are normally distributed[4]. Also, we can deduct that we can only have a meaningful σ^2 when the energy of f(x) equals to 1, in accordance with signal-to-noise ratio[5].

b)

To obtain best fit hypothesis, we firstly split our data into test-train sets. However, in order to find best fit hypothesis, we should use all of the datapoints that we have. I preferred do this splitting with manual number in each call. Then as we decided to work with legendre polynomials, we need to add polynomial features to the data, which will be followed by scaling of the data as next step. As following to this we can fit our training data, where our linear model would learn the behavior of our target function. When we use our test data with this model, we would obtain predicted results for g(x).

c)

In given problem we know the target function and also the best fit hypothesis. If we compare our centroids from polynomial linear regression with our target function and hypothesis by considering error, we can find expected error for our best fit hypothesis.

To begin with, analytical calculation of expected value of out of sample error equals to:

$$\mathbb{E}_D\left[(g_{10}(x) - f(x))^2\right] = \mathbb{E}_D\left[\underbrace{g_{10}(x) - \mathbb{g}_{10}(x))^2}_{variance}\right] + \underbrace{(g_{10}(x) - f(x))^2}_{bias}]$$

Since we are working with noisy data, we should also include error to our formula. Required transformations are as follows. We will start by transforming *f(x) to y* which are our polynomial centroids.

$$y = f(x) + \varepsilon(x)$$

$$\mathbb{E}_{D,\varepsilon}[(g_{10}(x) - y)^2] = \mathbb{E}_{D,\varepsilon}\left[(g_{10}(x) - f(x) - \varepsilon(x))^2\right]$$

$$= \mathbb{E}_{D,\varepsilon}\left[(g_{10}(x) - \mathbb{g}_{10}(x) + \mathbb{g}_{10}(x) - f(x) - \varepsilon(x))^2\right]$$

Here the interpretation of the parenthesizes;

$$= \underbrace{(g_{10}(x) - \mathbb{g}_{10}(x))^2}_{variance} + \underbrace{(g_{10}(x) - f(x))^2}_{bias} + \underbrace{(\varepsilon(x))^2}_{\sigma^2}$$

Here variance tells us the value for moving from our hypothesis to centroids. Bias is the parameter for distance between centroids to target proper, this is our deterministic error. Finally, noise level is the difference between target proper and actual output, this is called as stochastic error.

As a short definition regarding the difference between deterministic and stochastic errors; stochastic noise is the irreducible error that afflicts our data, and it contains the noise that adds on each value of our data. The deterministic noise is indeed the bias of our learning model and can be reduced by choosing a different hypothesis set.

Finally we can formulize our expected value for out-of-sample error as shown below.

$$\mathbb{E}_D\left[(g_{10}(x)-f(x))^2\right]=\mathbb{E}_{D,x}[(g_{10}(x)-\mathbb{g}_{10}(x))^2]+\mathbb{E}_x\left[(\mathbb{g}_{10}(x)-f(x))^2\right]+\mathbb{E}_{\varepsilon,x}[(\varepsilon(x))^2]$$

d)

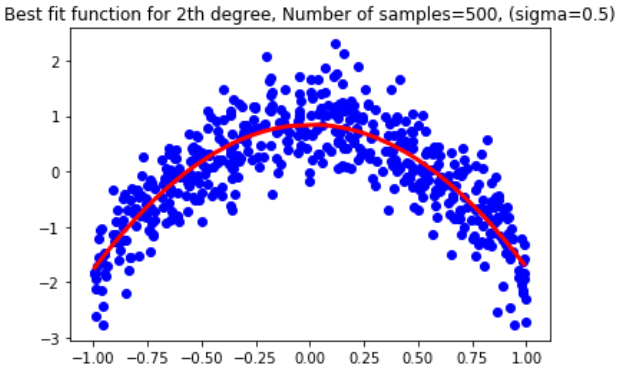
(d) Vary Q_f, N, σ and for each combination of parameters, run a large number of experiments, each time computing $E_{out}(g_2)$ and $E_{out}(g_{10})$. Averaging these out-of-sample errors gives estimates of the expected out-of-sample error for the given learning scenario (Q_f, N, σ) using $\mathcal{H}_2$ and $\mathcal{H}_{10}$. Let

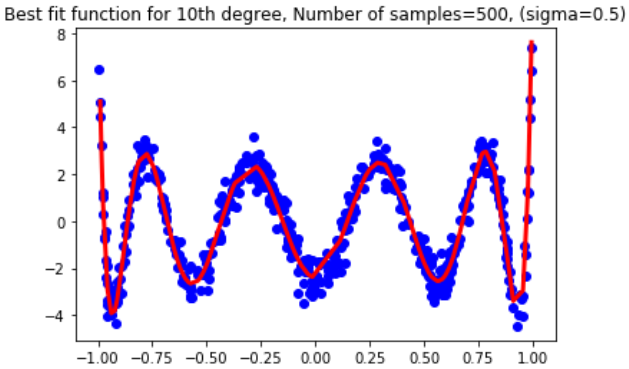
$$E_{out}(\mathcal{H}_2) = \text{average over experiments}(E_{out}(g_2)),$$
$$E_{out}(\mathcal{H}_{10}) = \text{average over experiments}(E_{out}(g_{10})).$$

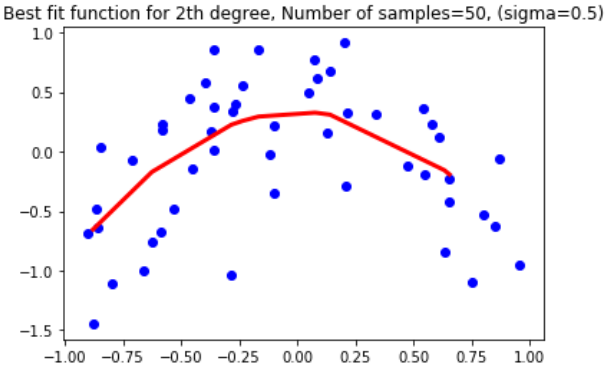
Define the overfit measure $E_{out}(\mathcal{H}_{10}) - E_{out}(\mathcal{H}_2)$. When is the overfit measure significantly positive (i.e. overfitting is serious) as opposed to significantly negative? Try the choices $Q_f \in \{1,2,\ldots,100\}$, $N \in \{20,25,\ldots,120\}$, $\sigma^2 \in \{0,0.05,0.1,\ldots,2\}$. Explain your observations.

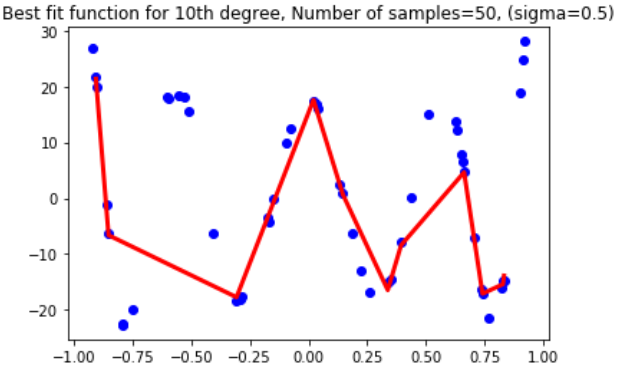
The logic behind the calculation of overfit measure with different degrees is only to compare error change in case we have models that have different complexity. If overfit measure is bigger than 0, we can conclude that our complex model does not fit well. Oppositely, if we find a negative error than it means that more complex model is better.

Results for 10th and 2nd degree hypothesis with the same amount of data as shown below.





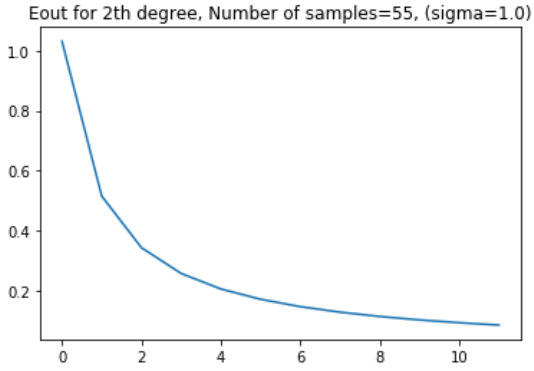


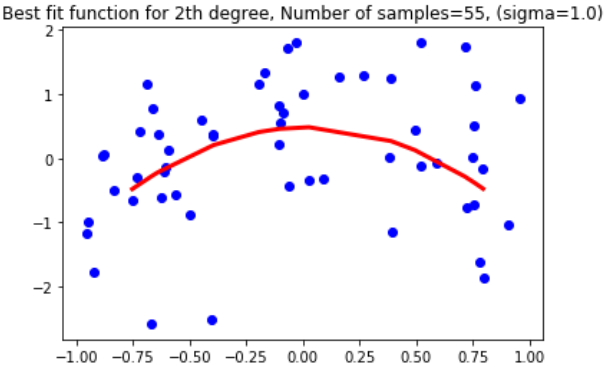


This graphs clearly shows that number of when error complexity increases, if number of the samples are also high, model fits very well. However, when number of sample is relatively low and model is too complex, overfitting occurs.

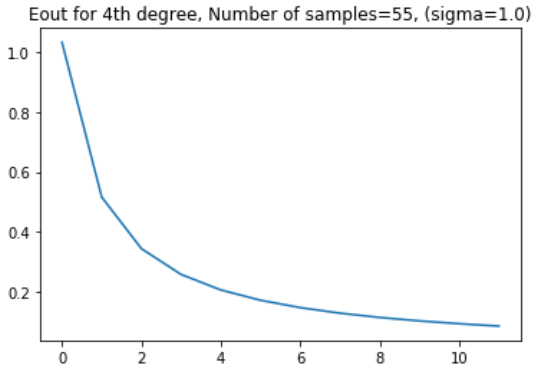
In order to have more generalized idea I split all of options for N, sigma and Qf to 2 parts and then I took the mean value of each half. Then I iterated through each quartiles to see results with incremental values for each element. Results are as shown below.

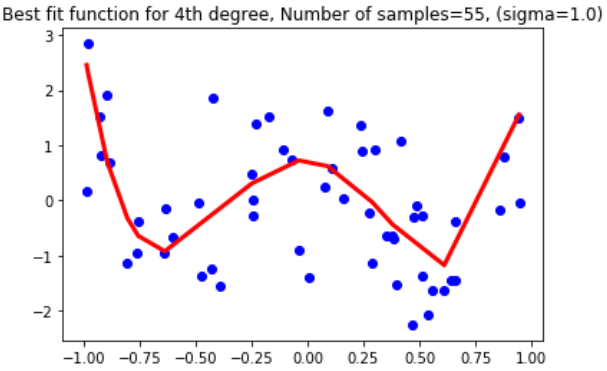
Sample 1



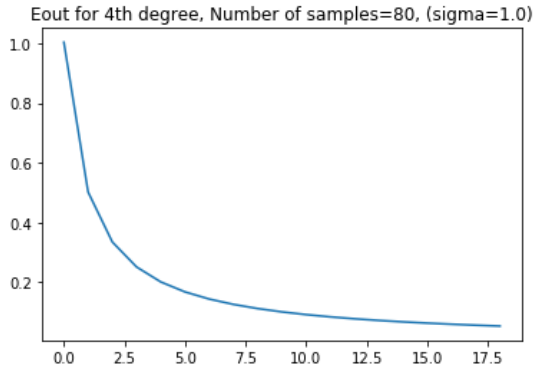


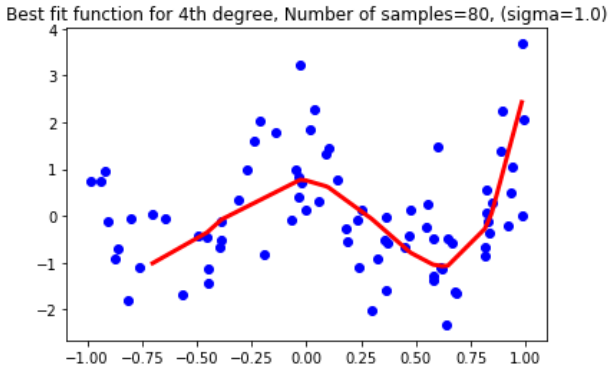
When we hold N and sigma same and increased degree of our function, it fits better.



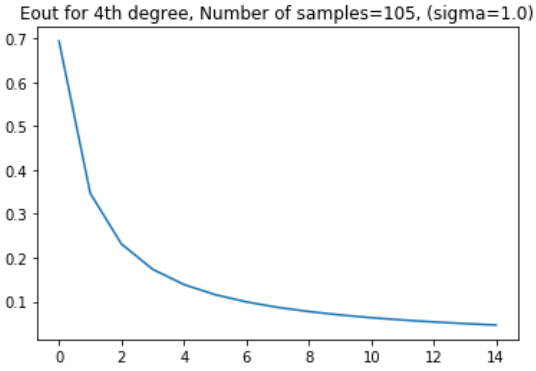


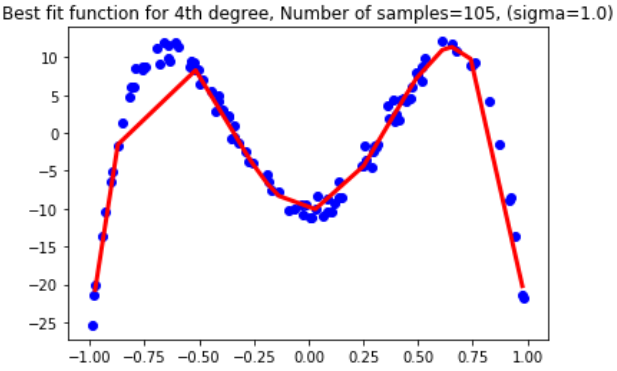
Sample 2



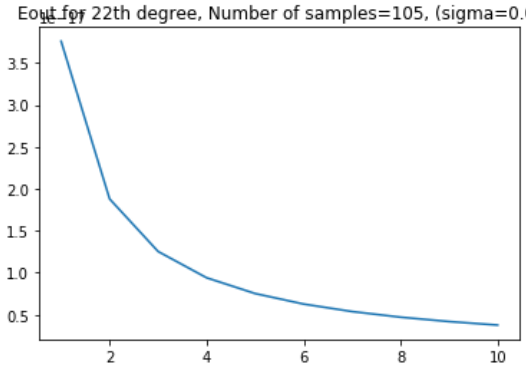


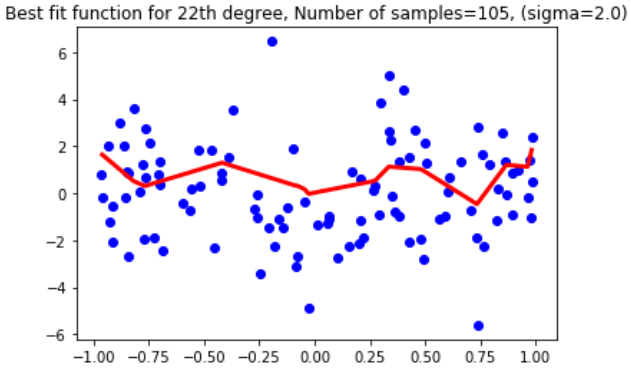
When we increase number of samples, function fits better. Error rate stabilesez in sooner time.



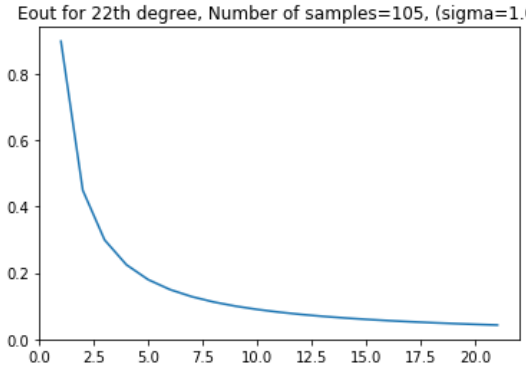


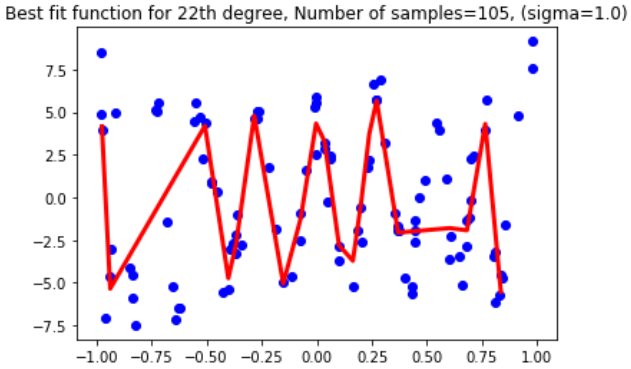
Sample 3



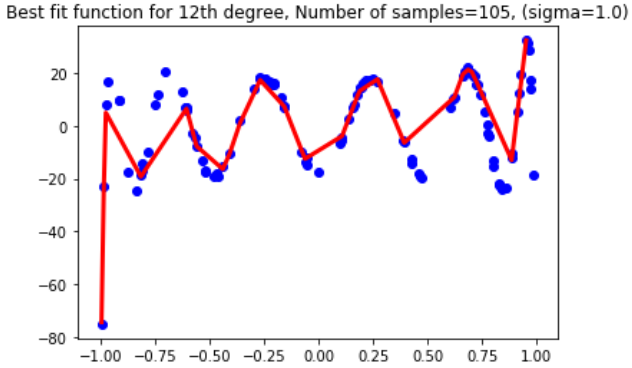


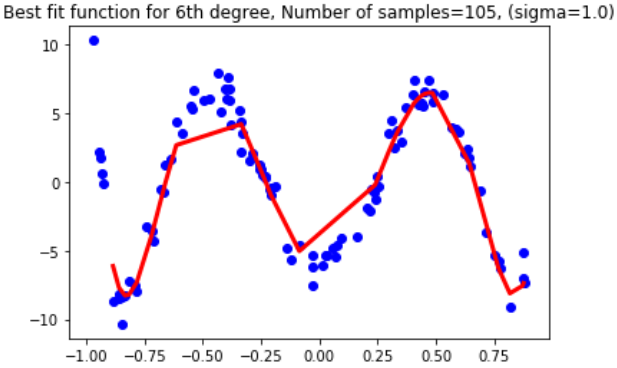
When we decrease sigma we get overfitting. Complexity of the model was too further then the required fit for our data set with smaller noise.





If we fit with less complex model, it firs better.





e) Why do we take average over many experiments? Use the variance to select an acceptable number of experiments to average over.

As we start exploring the weight space by going from one iteration to the next, we explore more and more space of the coefficients. When we explore the whole dataset, in case we have different datasets as well, we would have the effective VC dimension, which is basically total number of the free parameters in the model.

In order to show that representations of overfitting are not just point particular cases, we need to average over number of experiments. Thus, we conclude that overfitting occurs in the majority of the cases.

In terms of selecting acceptable number of experiments, there is no general calculation method. Rule of thumb according to a research says that, “if the underlying population has high variation in outcomes, the evaluation needs a larger sample.”[1]

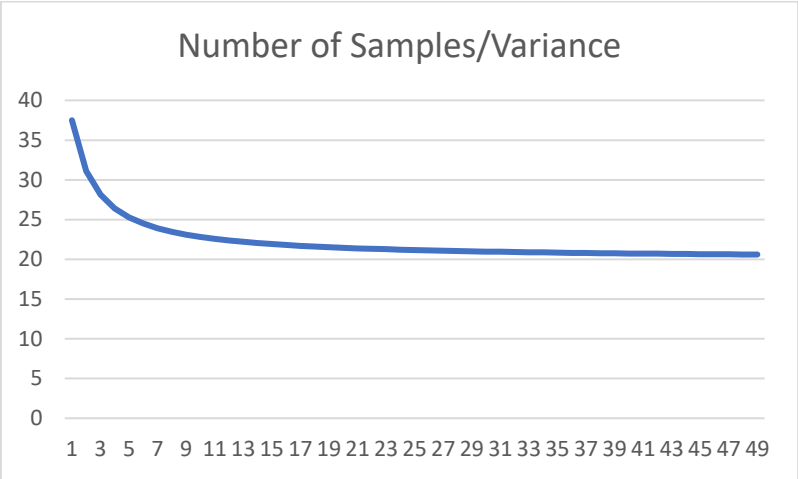
Given formula to calculate variance of the function;

$$Var(f(x)) = \sum_{q=0}^{Qf} \frac{{a_q}^2}{2q+1}$$

According to the book[3] another rule of thumb for this problem says that the ratio between number of parameters and data points should be equal to 10.

Here is the calculated ratio between number of samples and variance:

See calculation of variance for each number of parameters between [2, 50] and their comparison with related number of samples in Appendix A.



Here we can conclude that from the point on 21 parameters and 220 samples, we are starting to have a fixed index. So that, this should be our acceptable numbers for experiments.

Appendix

Number of Parameters	Number of Samples	Variance	N of Sample/Variance
2	30	0,800	38
3	40	1,286	31
4	50	1,778	28
5	60	2,273	26
6	70	2,769	25
7	80	3,267	24
8	90	3,765	24
9	100	4,263	23
10	110	4,762	23
11	120	5,261	23
12	130	5,760	23
13	140	6,259	22
14	150	6,759	22
15	160	7,258	22
16	170	7,758	22
17	180	8,257	22
18	190	8,757	22
19	200	9,256	22
20	210	9,756	22
21	220	10,256	21
22	230	10,756	21
23	240	11,255	21
24	250	11,755	21
25	260	12,255	21
26	270	12,755	21
27	280	13,255	21
28	290	13,754	21
29	300	14,254	21
30	310	14,754	21
31	320	15,254	21
32	330	15,754	21
33	340	16,254	21
34	350	16,754	21
35	360	17,254	21
36	370	17,753	21
37	380	18,253	21
38	390	18,753	21
39	400	19,253	21
40	410	19,753	21
41	420	20,253	21
42	430	20,753	21
43	440	21,253	21
44	450	21,753	21
45	460	22,253	21
46	470	22,753	21
47	480	23,253	21
48	490	23,753	21

References

1)https://www.povertyactionlab.org/sites/default/files/resources/2018.03.21-Rules-of-Thumb-for-Sample-Size-and-Power.pdf

2) https://dataakkadian.com/standardization-vs-normalization/

3) https://royaltymende96.files.wordpress.com/2019/02/learning-from-data.pdf

4)https://books.google.es/books?id=mviJQgAACAAJ&printsec=frontcover&hl=tr&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

5)https://en.wikipedia.org/wiki/Signal-to-noise_ratio