Foundations of Overparameterized Machine Learning
Final Project

Shakked River and Eshed Gal

For our final project in the course Foundations of Overparameterized Machine Learning, we chose the paper "Benign, Tempered, or Catastrophic: A Taxonomy of Overfitting" [Mallinar et al., 2022].
This final project has 3 sections: first, we provide a summarizing of our chosen paper. Next, we explain the relation between the paper and the topics taught in the course lectures. Finally, we aim to extend the paper and provide additional experimental results.

Summary of the Paper

The paper we chose discusses different types of overfitting: first, **benign overfitting**, which is the case where the number of training samples approaches infinity, the interpolating methods approach Bayes optimality (meaning, the limit is the optimal irreducible error). The other case is **catastrophic overfitting**, where the test error will grow larger and larger (and have a limit of infinity) as the number of training samples grows. The third option, which is a major focus of this paper, is **tempered overfitting** - a case where the method does not overfit benignly, and yet may be distinguished from the catastrophic case.
The contribution of the paper includes the formal identification of the intermediate regime between benign and catastrophic overfitting - the tempered overfitting situation. The paper shows that tempered overfitting appears in deep neural networks (DNNs) trained for interpolation and in ridgeless (no ridge regularization) kernel regression (KR). This leads the authors to claim that this definition completes the taxonomy of overfitting - each learning procedure will either be benign, tempered, or catastrophic. The paper provides both formal definitions and theoretical analysis as well as experiments over different kernels and DNNs to support these claims.
Section 2 of the paper provides a formal taxonomy of overfitting. It states the definitions relevant to in-distribution supervised learning settings, and the formal definition of learning procedures. The taxonomy suggested by the authors for the 3 different types of overfitting is defined in the following way (which is presented in Table 1 in the paper):

- Benign overfitting is when the limit of the error approaches the irreducible risk R^* (the optimal error) as the number of training samples goes to infinity.
- Catastrophic overfitting is the case whereas the number of training samples grows to infinity, the error goes to infinity in the regression case and to $1 - \frac{1}{K}$ for K-class classification (error of the predictor choosing a uniformly random label).
- Tempered overfitting is when the limit of the error when the training samples number grow to infinity is in between the last stated options.

The error is measured using MSE (mean squared error) for regression or with classification error measure for classification tasks. Technically there might be a fourth option, that the

limit does not exist, but the paper states that this appears to happen in pathological cases only. The learning considers noise profile, where there is noise added to the training samples. In the regression setting the noise will be in the form of Gaussian noise, and for classification, there will be label-flip probability. The paper provides a list of different models and categorizes them into 3 possible overfitting options. The list appears in Table 2 of the paper.
In section 3, the paper shows a study of kernel regression. It presents the formal KR predicted function, and the related mathematical definitions needed to formally state the theoretical result. The main result of this section is theorem 3.1 for KR trichotomy, which gives formal conditions for measuring which of the 3 overfitting types would accrue. Furthermore, the theorem states an expression for the limit of the tempered case, which is affected by the kernel properties. This section elaborates on the implications of this theorem: for KR with fixed positive ridge regression parameters (and appropriate target function), the model will have benign overfitting. For ridgeless KR, there might be tempered overfitting or catastrophic overfitting, depending on the kernel (for example, the Laplace kernel shows tempered overfitting while the Gaussian kernel shows catastrophic one on the ridgeless case). Lastly, early stopping on a wide network should be expected to fit benignly, and when using non-ReLU type kernels there is a dependency on the activation function. The full proof of the theorem is provided in the paper's appendix.
Next, the paper provides experiments to support the taxonomy presented above. Starting with kernel regression experiments, the authors use synthetic data distribution where the samples are drawn from the unit sphere and the targets y are Gaussian: $y_i \sim \mathcal{N}(0, 1)$. This is in fact a regression setting for learning the constant 0 function (with Gaussian noise). The experiments included different kernels, and provided the expected results (which are shown in Figure 4 in the paper): For the Gaussian kernel with ridge regression, the test MSE converges to 0 as the number of training samples grows. For ridgeless Laplacian kernel, the test MSE converges, but to a constant limit larger than 0. And for ridgeless Gaussian kernel, the convergence is indeed catastrophic as we can see that the test error keeps growing as the number of train samples grows. Experiment also show support of the mathematical limit expression of the tempered overfitting case stated in the theoretical part, as presented in Figure 5 of the paper.
An additional set of experiments used deep neural networks. Experiments used the CIFAR-10 dataset (two-class and ten-class versions) as well as a synthetic dataset that aims to learn the constant function 1. Results show that the networks interpolate in a tempered manner, as expected (the experiment is presented in Figure 6 in the paper). Though this is not a supporting result to the theoretical theorem (which discussed kernels), this experiment shows it is heuristically consistent. Finally, experiments show that while trained interpolating DNN exhibits tempered overfitting, early stopped DNNs exhibit fairly benign fitting. In fact, we can observe that looking at the MSE as a function of training time, early training shows benign overfitting while later we will have tempered overfitting (this is shown in figures 7 and 8 in the paper).
In conclusion, this paper presented a study of the nature of overfitting in learning methods that interpolate their training data. It showed a complete taxonomy of overfitting - each learning procedure will be either benign, tempered, or catastrophic. The paper presented a theoretical analysis of kernel regression for the different 3 types and provided experiments conducted on different kernels and deep neural networks.

Relation of the Paper to Class Topics

In this section, we are looking for similarities and differences between course lectures and the paper we chose. First, we note that as far as formal mathematical definitions go, there is

a similarity between the formal definition of learning and test error given in section 2.1 of the paper and the definitions and formalization we saw in class. For example, looking at the definitions in part 1 of lecture 2 (General Notations and Definitions for Supervised Learning), we see similar definitions. This is unsurprising, as this formal language of learning definitions is standard.

In class, we saw a very comprehensive analysis of linear regression (least squares). The paper mentioned that many studies of linear regression show that in case the input dimension grows faster than the sample size, there will be benign overfitting. This is stated in the prior work section of the paper (section 1.2), and the authors refer to [Bartlett et al., 2020] several times in the paper.

In lecture 6, we saw the analytic characterization of least squares regression in the Gaussian case. In the experiments section on kernel regression, 4.1, it is mentioned that the experiment that is presented in Figure 5 of the paper has an equivalent setup to linear regression with random Gaussian covariates. The setup, which is elaborated in Appendix B of the paper, shows that similarly to the lecture, the samples are drown in an i.i.d. manner with isotropic Gaussian distribution. In section 2.4 of the paper, it is noted that if the number of features is the same as the number of samples (meaning it is at the double descent peak), the learning procedure will have catastrophic overfitting. We note that in the analysis we saw in class in lecture 6, when we analyze the case where $p = n$ we indeed see infinite test error. This is shown in slide 34 of lecture 6.

An important part of the paper discusses kernel regression, which we did not study specifically in the course lectures. However, the paper uses ridge regression which we did see in lecture 7. As stated in theorem 3.1, for fixed positive ridge regression (and appropriate target function) the kernel regression will have benign overfitting.

The paper considers both regression setup and classification setup. Our main focus in the lectures was on regression, and yet we had lecture 8 which discussed classification models. We see that the setup and data model shown in the paper have a similar form to the ones from the lecture, for example, the training data noise which is shown in slide 11 of lecture 8.

On section 4.2 of the paper, we see the experiments conducted on deep neural networks. The paper uses ResNets that are trained to interpolate on the CIFAR-10 dataset. In lecture 9, which discusses double descent in DNNs, we see details of ResNets and CIFAR-10 (as well as CIFAR-100 which was not used in the paper). In the lecture, additional information is provided regarding the optimization of the network, for example, the GD algorithm and the learning rate. This type of discussion does not appear in the paper, as this is not the focus of the experiment. However, experimental details of the deep learning part are found in Appendix C.2 and C.3 of the paper, and the optimization details, among other network details, are provided there.

The last part of lecture 9 contains many graphs and experiments of the double descent phenomena, first as a function of the DNN width. Next, there are experiments that show double descent during training, as a function of epochs. This is similar to the time dynamic experiments that appear in section 4.2 of the paper, and presented in figure 8. Finally, at the end of lecture 9, there are some slides that show the effect of a larger training dataset, which is one of the main ideas of the paper. Specifically, on slide 43 we see test and train errors as a function of the number of training samples. This is similar to many results of the paper, and in fact we can observe that for example the large model experiences benign overfitting of test error, as we see that the relevant purple line in the graph on slide 43 goes down as the number of training samples grows. This is different from class, where we usually saw (and created on our own on the home assignments) graphs of the test error as a function of the number of parameters p and not the number of training samples n.

To conclude, while our chosen paper uses some of the main ideas we saw in class, such as

linear and ridge regression and the double descent phenomena, it also focuses on additional material we did not learn in the course lectures, such as kernel regression. Furthermore, the definition and taxonomy of overfitting in our paper focus on the limit of the test error as the number of training samples goes to infinity, and we did not focus on this taxonomy in class.

Paper Extension

In this section, we are looking for an experimental extension of the paper. We focus on the experiments conducted in section 4.1, using kernel regression. The code we used is provided in the $OML\_final\_project.py$ file, and available in GitHub as well[1]. The code has some sections, where experiments appear one after the other and each experiment is titled with its sequential number.
We chose to look at this part of the experiment since the paper deals mainly (both theoretically and empirically) with kernel regression, and yet we felt that there is room for more experiments to be made to provide stronger evidence for the theory stated in the paper, as well as some explanations and results that did not appear there.
Challenges of performing this extension involve the fact that we did not learn kernel regression in class (we have seen some of it in different courses, but we needed to refresh our theoretical knowledge to deal with this topic).

Reproduce Paper Results

As a starting point for our experiments, we needed to reconstruct the results shown in section 4.1 of the paper. Specifically, we aimed to reproduce Figure 4, which shows benign fitting for ridged Gaussian kernel, tempered fitting for a ridgeless Laplacian kernel, and catastrophic fitting for a ridgeless Gaussian kernel.
We note that the authors did not publish their code, which means we needed to perform this part on our own. We used some of the experimental details provided in Appendix C of the paper. Similarly to the original figure, we use a regression factor of 0.1, and use synthetic data of point on the 5-dimensional sphere, with label noise $y_i \sim \mathcal{N}(0, 1)$. In our code, we create the data and calculate the test MSE for each of the kernels using our fit function. The MSE is tested for different values of train samples, similar to the plots in Figure 4. We calculate and fit the model directly, as the authors stated they performed in their experiment (in appendix C of the paper). Finally, we have a function that plots the results. Farther technical details of our code for this part are elaborated in Appendix A. The fit_kernel function is performing the mathematical formula (2) presented in section 3 of the paper:

$$\hat{f}(x) = K(x, D_n) \left(K(D_n, D_n) + \delta I_n\right)^{-1} Y$$

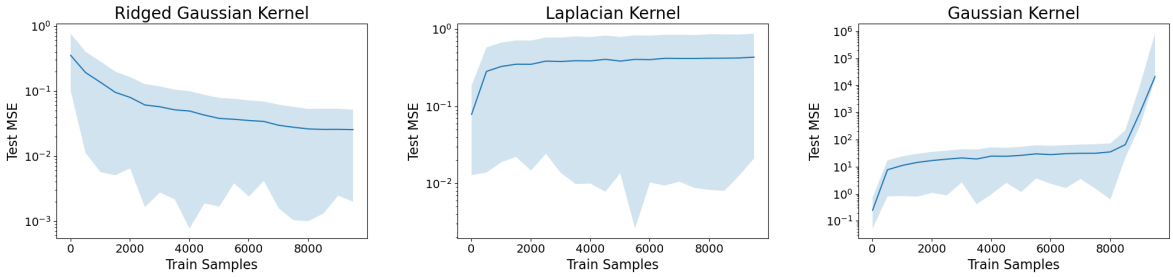
Gaussian and Laplacian kernel[2] are calculated with the formal definitions:

$$K_{Gaussian}(x_i, x_j) = \exp(-\gamma ||x_i - x_j||_2^2)$$

$$K_{Laplacian}(x_i, x_j) = \exp(-\gamma ||x_i - x_j||_1)$$

When γ is a kernel parameter. This part yielded the plots presented in Figure 1, which in fact shows that our reconstruction succeeded and we are ready for additional experiments to be made.

[1] https://github.com/shakkedriver/opml_finel
[2] https://scikit-learn.org/stable/modules/metrics.html



(a) Ridged Gaussian Kernel (b) Ridgeless Laplacian Kernel (c) Ridgeless Gaussian Kernel

Figure 1: Kernel regression different fitting regimes, test MSE as function of train samples

Experiment 1: Try new kernels

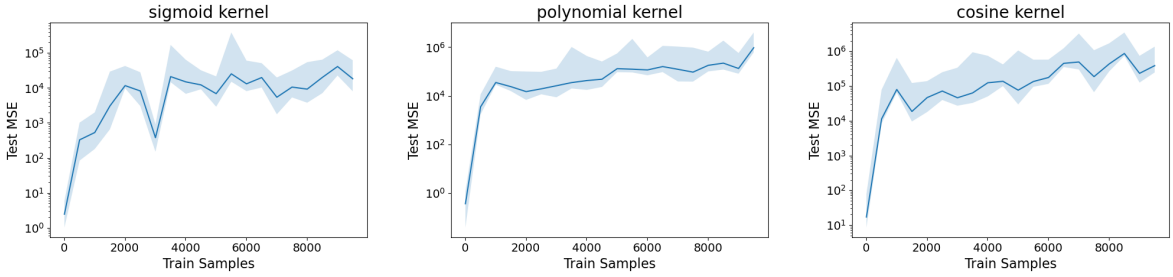
Our first extension will be trying new different kernels. We note that kernel regression is a wide field, and the paper only presented and experimented with Gaussian and Laplacian kernels. Even when we look at Table 2 which shows a taxonomy of known fitting results, we note that there are some possible kernels that are missing and their asymptotic behavior might be worth checking.
Our experiment uses 3 kernels: sigmoid, polynomial, and cosine. We use the formulas:

$$K_{sigmoid}(x_i, x_j) = tanh(\gamma < x_i, x_j > +c)$$

$$K_{polynomial}(x_i, x_j) = tanh(\gamma < x_i, x_j > +c)^{deg}$$

$$K_{cosine}(x_i, x_j) = \frac{< x_i, x_j >}{||x_i|| \cdot ||x_j||}$$

Where c is a constant and deg is the polynomial degree. We calculated test MSE using a fit function with similar methods to the ones we used for the reconstruction step. Our results are presented in Figure 2.
We see that all 3 newly tested kernels perform catastrophic fitting. We suspect that the explanation lies in theorem 3.1, where the new kernels have similar eigandecay as in option (c). However, our focus is on experimental results and mathematical proof of that requires further details of kernels regression theory. This result might suggest that catastrophic behavior is actually common and might be expected when using ridgeless kernel regression.



(a) Ridgeless Sigmoid Kernel (b) Ridgeless Polynomial Kernel (c) Ridgeless Cosine Kernel

Figure 2: Kernel regression with newly tested kernels, test MSE as function of train samples

Experiment 2: Try non-linear target function

Next, we noted that the experiment presented in Figure 4 only considered one target function: the constant function 0. As this is a very "easy" function, we aimed to figure out what would

happen if we tried to learn different and more complicated functions. In other words, we aim to check that the results presented in the paper are wildly true and are not limited to a simple and trivial target function, where all labels are 0. For this experiment, we chose a non-linear target function:

$$f(x) = \sum_{i=1}^{d} x_i^2 \cdot i$$

For d dimensional vector x. This function is calculated in $quad\_func$. The results show that ridgeless Laplacian and Gaussian kernels appear to have similar asymptotic behavior for this target function as well - we observe tempered fitting for Laplacian and catastrophic fitting for the Gaussian kernel. Since those results are similar to previous ones, we present them on Appendix B. However, we note that something different now happens for the ridge Gaussian case. In fact, it appears that the test MSE does not decay to zero, as we expected. This is shown in Figure 3a.
A closer look leads us to suspect that although we used a positive fixed ridge value (of 0.1), we believe this function fails to follow the condition detailed in the first bullet of the theorem implications explanations that guarantee benign fitting: The function needs to be in the kernel's RKHS. As this might be formally proved, a closer look at this definition (Reproducing kernel Hilbert space) [3] leads us to believe that this is the explanation for this result.

Experiment 3: Try different target function

Considering experiment 2 results, we wanted to look for new different function, which is not as trivial as the constant 0 function and yet will show the benign behaviour described in the paper. We chose to use the following function:

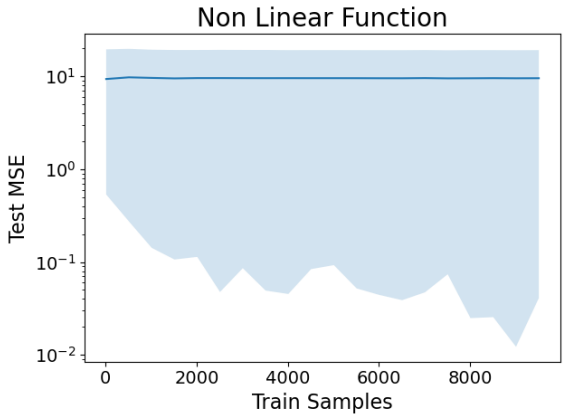
$$f(x) = k_{Gaussian}(x, 1^d)$$

For d dimensional vector x, which is in the RKHS of the Gaussian kernel since it is simply calculating the kernel of the given vector x and a constant vector. We try again the ridged Gaussian kernel using this function, and see in Figure 3b that now we indeed experience benign fitting, which supports our claim that the former result had to do with the RKHS. This also shows that the results presented in the paper will still hold on the synthetic data even for a more complicated function than the constant 0 function.

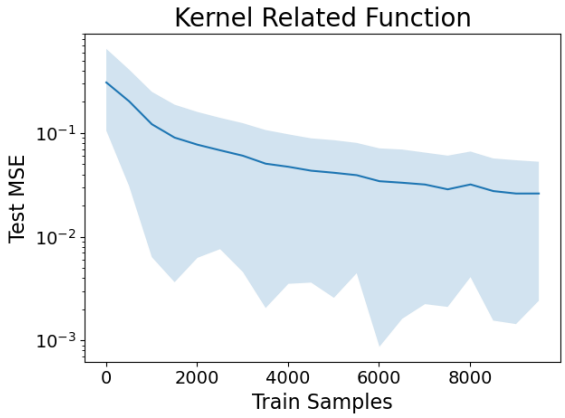
Experiment 4: Different ridge regularization values for zero function

In our next experiment, we took another angle of the results of section 4.1. As stated in the part detailing the relation between the paper and the class lectures, while we did not see kernel regression in class, we did learn about the effect of using ridge regression. Our extension in this part is to try different ridge values and check the effect of using small or large regularization factors. This is not done at all in the paper, as the authors only considered regularization of value 0.1. In a way, this is similar to the coding questions in homework 2, though here we calculate the test error as a function of train sample size, similarly to the plots in the paper.
For this experiment, we consider a few different regularization factors, and show the results in figure 4. We observe that though all plots experience benign fitting as expected (test MSE decays as the number of train samples grows), the regularization factor has an effect - the stronger the regularization, the faster the decay. This makes sense (and similar to class results) since using more regularization reduces the variance of the model thus making it more stable. This kind of observation is not considered in the paper.

[3] https://en.wikipedia.org/wiki/Reproducing_kernel_Hilbert_space



(a) Ridged Gaussian kernel for non linear target function (experiment 2)



(b) Ridged Gaussian kernel for kernel related target function (experiment 3)

Figure 3: Kernel regression with ridgeless kernels and different target functions, , test MSE as function of train samples

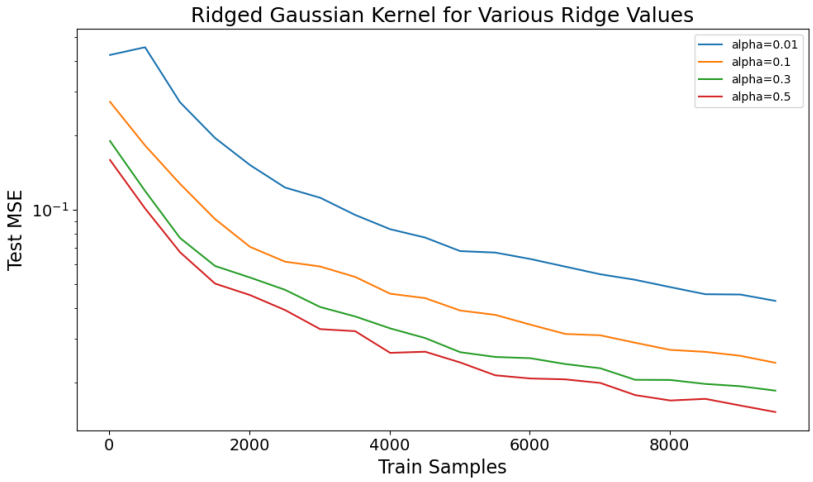


Figure 4: Test MSE for various ridge regression values with Gaussian kernel, as function of train samples

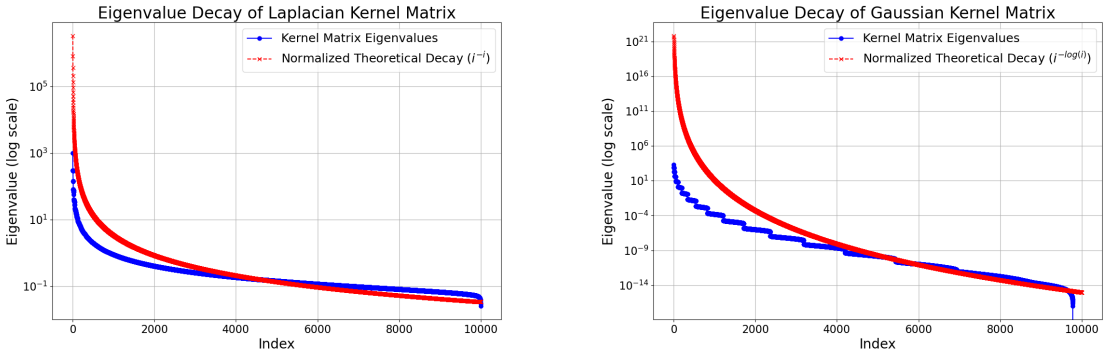
Experiment 5: Different ridge regularization values for non-linear function

In this experiment, we go back to the non-linear function target of experiment 2. As detailed, we saw that in the ridged case, there appeared to be an asymptotic behavior more similar to the tempered case than the benign case. Now, we aim to challenge our thought that this happens since the function is not in the kernel's RKHS, and check if maybe just adding stronger (or weaker) regularization might help. Results support our former claim - since we see that even changing the regularization factor does not make the error decay and the plot is similar to the one observed in Figure 3a. We present this result in Appendix C.

Experiment 6: Showing kernel eigenvalues asymptotic decay

In this last experiment, we try to give the theoretical conditions of theorem 3.1 some experimental supporting results. We observe that the authors only mentioned the asymptotic decay of the λ_i values and did not conduct any specific experiments or elaborated about it.

Hence, we chose to calculate and plot the eigenvalues decay of both Gaussian and Laplacian kernels, and compared each of them with the related asymptotic function of the theorem: for Gaussian, $\lambda_i = i^{-logi}$ and for Laplacian $\lambda_i = i^{-\alpha}$ for $\alpha > 1$, which we chose to be 2. We use a kernel data matrix size of $10,000$ and get the following plots showed in Figure 5, which indeed support the theorem, as we see a similarity between the theoretical function and the eigenvalues decay.



(a) Laplacian kernel eigenvalue decay

(b) Gaussian kernel eigenvalue decay

Figure 5: Kernel eigenvalues decay, eigenvalues as function of their index in descending order

Conclusion

In this part of our project, we tried different experiments and extensions regarding kernel ridge regression. We tested new kernels, tried different target functions, and considered the effect of changing the ridge regularization factor. We also investigated the kernel eigenvalues property to try and understand the difference between different kernels which yields different asymptotic behavior. As there is much left to try and expand on this topic, we believe that we found some interesting results during our experiments.

References

[Bartlett et al., 2020] Bartlett, Long, Lugasi, and Tsigler (2020). Benign overfitting in linear regression. In *Proceedings of the National Academy of Sciences*.

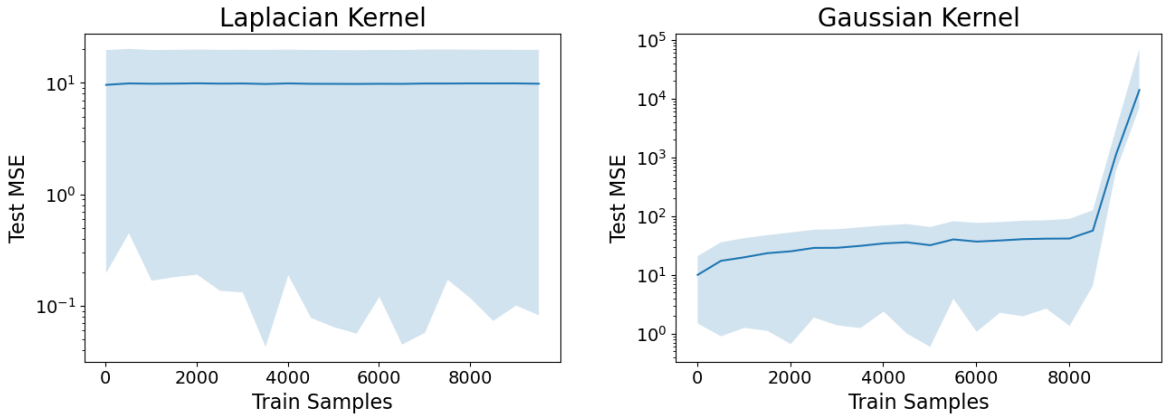
[Mallinar et al., 2022] Mallinar, Simon, Abedsoltan, Pandit, Belkin, and Nakkiran (2022). Benign, tempered, or catastrophic: A taxonomy of overfitting. In *NeurIPS*.

A Paper Reproduction of Results Experimental Details

For test error calculations, we used the median of 16 iterations and plotted the error ridge of 25 to 75 percent. We note that as detailed in Appendix C and in Figure 4 in the paper, the authors performed each experiment more iterations than we did and for 3 different values of d, yet we feel confident that our reproduction is suitable and our experiments have been repeated enough times to show confident results. As for different dimensions, we see that for each value of d, the asymptotic behavior is the same for a given kernel, which means that using one selected value of d will show the asymptotic behavior and allow us to observe the fitting taxonomy. We used 20 different values for train data size from 10 to 10,000. As a side note, we mention that we performed the experiments using a CPU, and each one took about an hour to complete.

B Experiment 2 Results For Ridgeless Kernels

As stated in this experiment's description, the results of trying to learn different non-linear functions for ridgeless Laplacian and Gaussian kernels yielded similar plots to the ones where we aimed to learn the constant 0 function. For completeness, we present those plots here:



(a) Ridgeless Laplacian Kernel (Tempered)

(b) Ridgeless Gaussian Kernel (Catastrophic)

Figure 6: Kernel regression with ridgeless kernels, non linear function

C Experiment 5 Results

On this experiment, we tested the option that using different ridge values will results the asymptotic fitting to decay on our non linear target function. We see here that this in fact (and not surprisingly) does not happened.

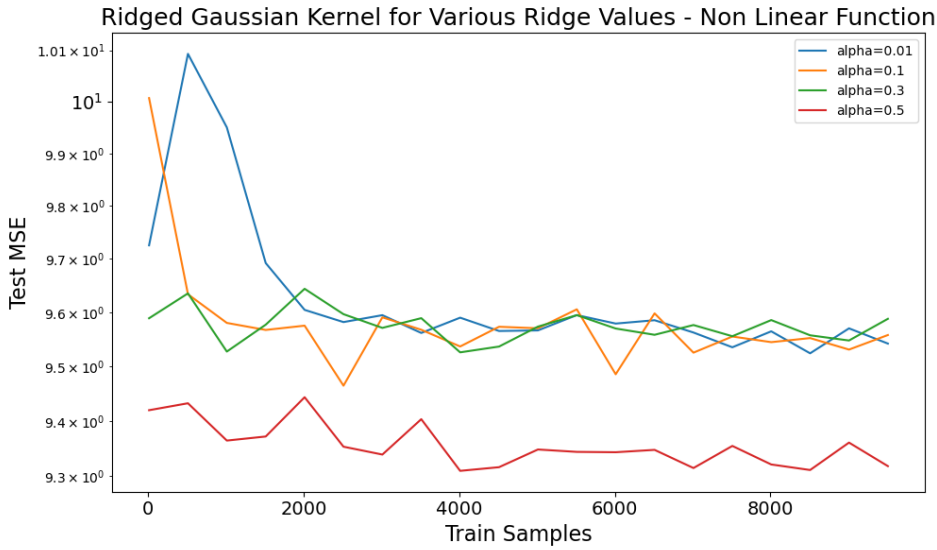


Figure 7: Test MSE for various ridge regression values with Gaussian kernel for non-linear function