

# COMP0086 Summative Assignment

Alexei Pisacane and Jian Shu (James) Wu

Nov 16, 2022

## 1 PART I

### 1.1 Linear Regression

1. (a) Superimposing the four different curves corresponding to each fit over the four given data points:

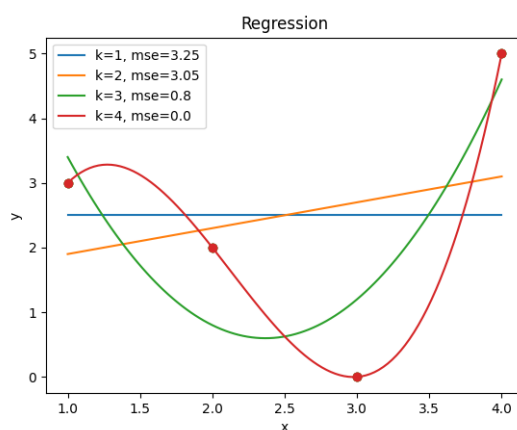


Figure 1: Data set  $\{(1, 3), (2, 2), (3, 0), (4, 5)\}$  fitted with basis  $\{1, x\}$  and basis  $\{1, x, x^2, x^3\}$

- (b) The weights for the equations:

Basis	k=1	k=2	k=3	k=4
1	0.0	0.0	0.0	-5.0
$x^1$	0.0	0.0	9.0	15.17
$x^2$	0.0	1.5	-7.1	-8.5
$x^3$	2.5	0.4	1.5	1.33

- (c) The mean squared error for each fitted curve:

Metric	k=1	k=2	k=3	k=4
MSE	3.25	3.05	0.8	0.0

2. (a) i. Plotting the function  $\sin^2(2\pi x)$  in the range  $0 \leq x \leq 1$  with the sampled points superimposed:

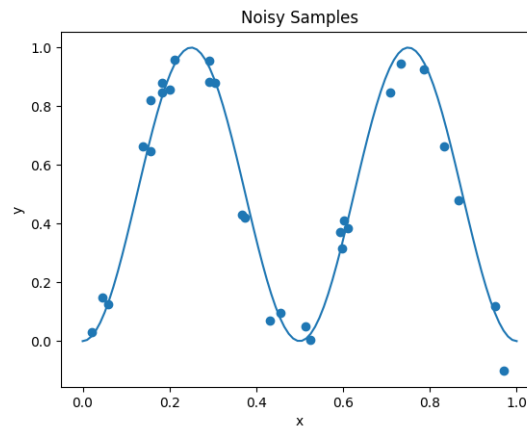


Figure 2: Noisy samples

- ii. Fitting the data set with polynomial bases of dimension  $k = 2, 5, 10, 14, 18$  and plotting each of these 5 curves with the sampled points superimposed:

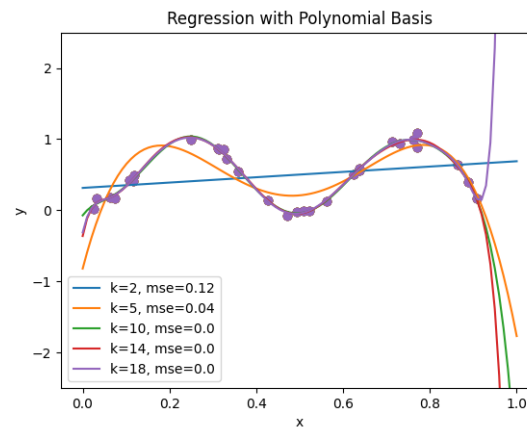


Figure 3: Polynomial fit curves

- (b & c) Plotting the  $\ln$  of the train and test error for a single trial of 1000 test points versus the polynomial dimension  $k = 1, \dots, 18$ :

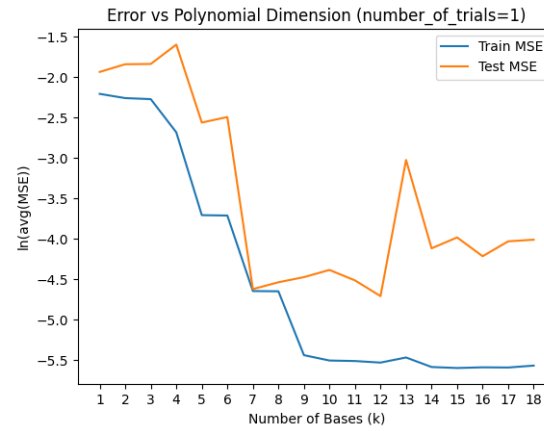


Figure 4: Training and Testing Error of Polynomial Curves (1 trial)

- (d) Plotting the  $\ln$  of the train and test error for 100 trials of 1000 test points versus the polynomial dimension  $k = 1, \dots, 18$ :

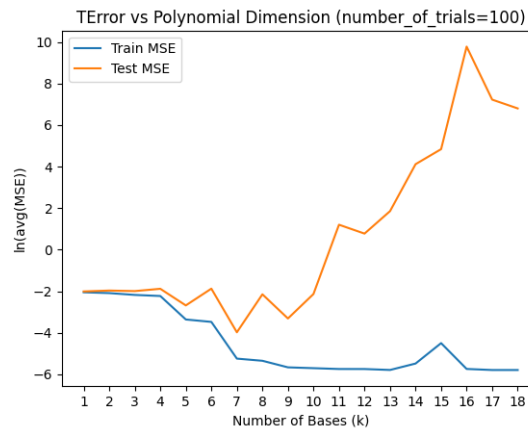


Figure 5: Training and Testing Error of Polynomial Curves (100 trials)

3. Repeating 2 (b-d) with sin basis functions:

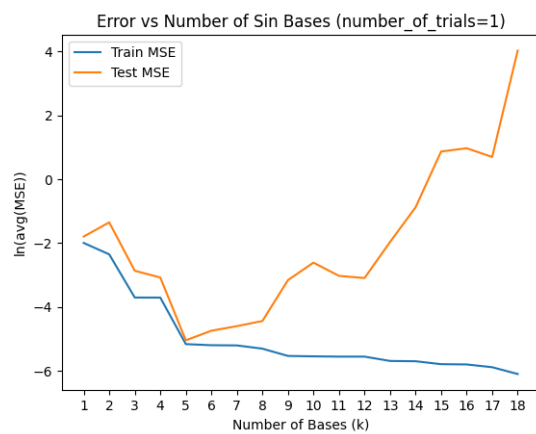


Figure 6: Training and Testing Error of sin Curves

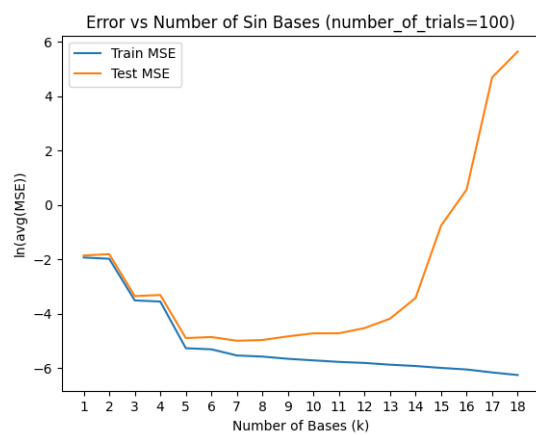


Figure 7: Training and Testing Error of sin Curves (100 trials)

## 1.2 Filtered Boston housing and kernels

4. (a) We interpret the naive linear regression model using the existing framework for least squares regressors.

Namely, we have:

$$w^* = (X^T X)^{-1} X^T y, X = (1, \dots, 1)^T$$

$$\implies w^* = m^{-1} \sum_i 1 * y_i = \bar{y}$$

$$\implies f(x_i) = 1 \times w^* = \bar{y}$$

Our model in this case is essentially always predicting the average housing price in the training data set.

- (b-d) Here is a table of the training and test mean squared errors:

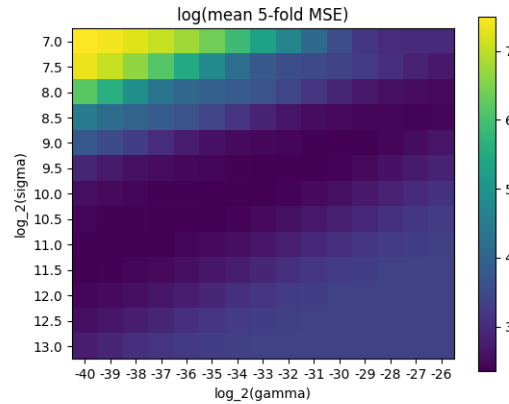
Model	Train Mean	Train St. Dev.	Test Mean	Test St. Dev.
Naive Regression	84.71	7.99	85.79	9.73
Linear Regression (feature='CRIM')	69.0	7.31	74.26	9.83
Linear Regression (feature='ZN')	74.81	7.31	72.79	8.6
Linear Regression (feature='INDUS')	65.15	4.08	63.51	6.75
Linear Regression (feature='CHAS')	82.69	4.93	83.55	6.79
Linear Regression (feature='NOX')	69.49	4.85	68.61	6.42
Linear Regression (feature='RM')	42.13	5.15	46.32	6.96
Linear Regression (feature='AGE')	75.14	6.8	69.54	6.31
Linear Regression (feature='DIS')	79.69	7.65	80.13	11.38
Linear Regression (feature='RAD')	71.44	6.28	72.9	5.98
Linear Regression (feature='TAX')	66.54	5.56	64.89	7.24
Linear Regression (feature='PTRATIO')	62.52	4.42	64.12	5.35
Linear Regression (feature='LSTAT')	38.95	2.97	38.36	3.93
Linear Regression (all features)	21.53	2.47	25.79	4.1

### 1.3 Kernelised ridge regression

5. (a) The optimal parameters:

<b>Trial</b>	$\log_2(\sigma)$	$\log_2(\gamma)$
0	-38.0	11.0

(b) Displaying the logarithm of the mean 5 fold mean squared error for different parameters for a single trial:



(c) The mean squared error for the training and test sets for the best  $\gamma$  and  $\sigma$  for a single trial:

<b>Model</b>	<b>Train Mean</b>	<b>Test Mean</b>
Kernel Ridge Regression	5.94	13.38

(d) The mean squared error for each fitted curve over 20 runs:

<b>Model</b>	<b>Train Mean</b>	<b>Train St. Dev.</b>	<b>Test Mean</b>	<b>Test St. Dev.</b>
Naive Regression	84.71	7.99	85.79	9.73
Linear Regression (feature='CRIM')	69.0	7.31	74.26	9.83
Linear Regression (feature='ZN')	74.81	7.31	72.79	8.6
Linear Regression (feature='INDUS')	65.15	4.08	63.51	6.75
Linear Regression (feature='CHAS')	82.69	4.93	83.55	6.79
Linear Regression (feature='NOX')	69.49	4.85	68.61	6.42
Linear Regression (feature='RM')	42.13	5.15	46.32	6.96
Linear Regression (feature='AGE')	75.14	6.8	69.54	6.31
Linear Regression (feature='DIS')	79.69	7.65	80.13	11.38
Linear Regression (feature='RAD')	71.44	6.28	72.9	5.98
Linear Regression (feature='TAX')	66.54	5.56	64.89	7.24
Linear Regression (feature='PTRATIO')	62.52	4.42	64.12	5.35
Linear Regression (feature='LSTAT')	38.95	2.97	38.36	3.93
Linear Regression (all features)	21.53	2.47	25.79	4.1
Kernel Ridge Regression	7.38	1.68	14.19	2.05

## 2 PART II

### 2.1 $k$ -Nearest Neighbors

6. Generating a random  $h$ :

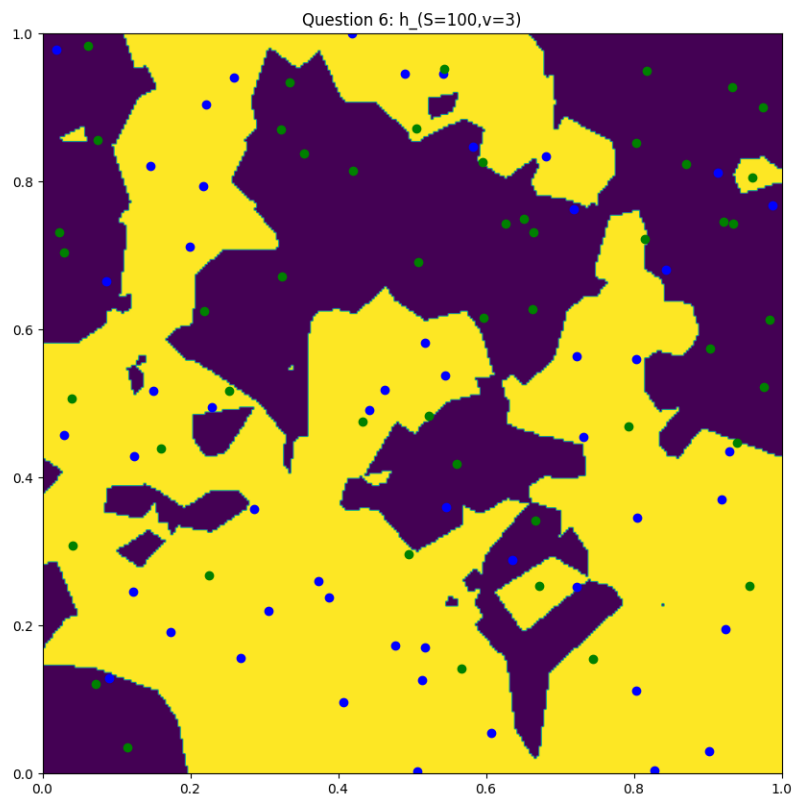


Figure 8: A hypothesis  $h_{S,v}$  visualised with  $|S| = 100$  and  $v = 3$

7. (a) Producing a visualisation using Protocol A:

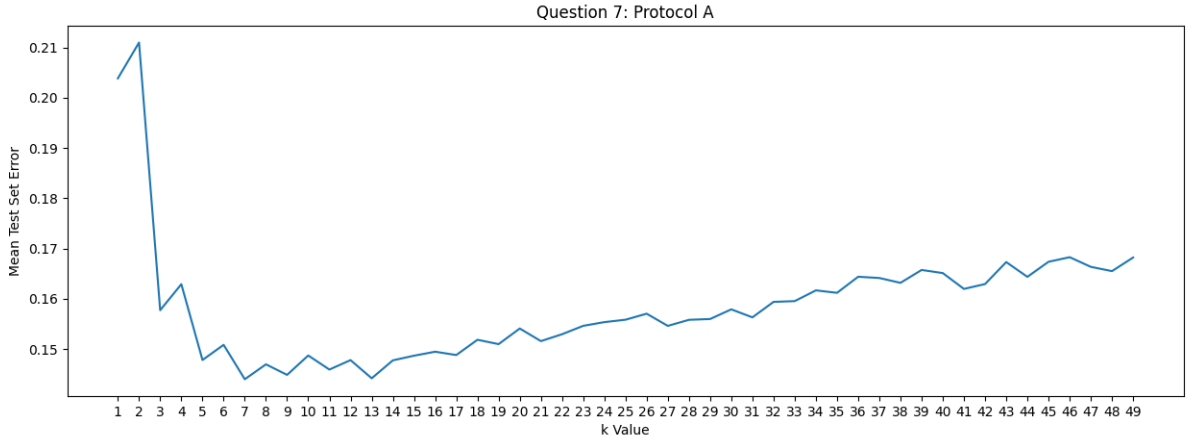


Figure 9: Protocol A

- (b) We know that the data generating process uses  $\nu = 3$ , so we would have initially expected the smallest test error to be located at the matching k-value,  $k = 3$ . However, the data generating process for the training data set labels *and* the testing data set labels were injected with noise, so k-values greater than 3 have a “de-noising” effect allowing for better performance at around  $k = 7$ . As the k-value increases beyond  $k = 7$ , this smoothing effect is too strong, causing the gradual increase in test error as the k value increases. For  $k < 7$  and in particular  $k < \nu = 3$  we see the error growing because the model has small  $k$  and is able to fit to the noise injected into the data generating process, corrupting the model. By averaging over 100 runs, the plot is relatively smooth allowing us to make generalisations about the relationship between k-value and mean test set error.



8. (a) Producing a visualisation using Protocol B:

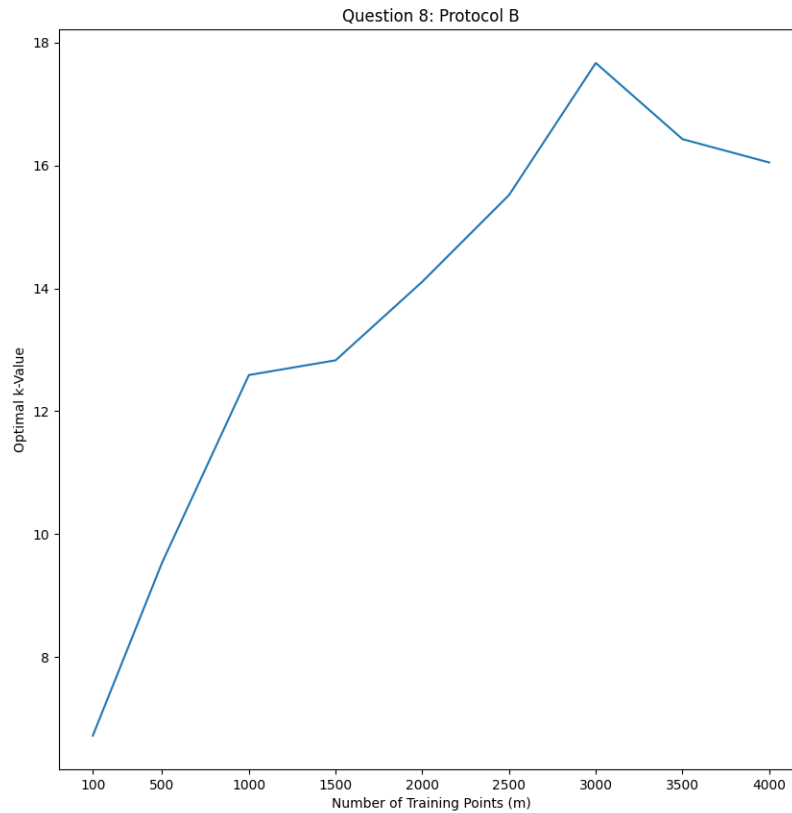


Figure 10: Protocol B

- (b) k-nearest neighbours is a non-parametric model, so when queried with a test point for prediction, the number of training points within a given fixed radius of that test point increases as the number of training points increases. If we consider this fixed radius ‘sufficiently close’ to make an informed prediction, the number of points within this radius will increase with more training points. So as the number of training points increases, the optimal k would increase to include more of the training points within this same fixed radius. We can see in the plot above that we see this general trend as expected, however there is considerable noise in the trend and so a smoother plot would likely require more runs.

### 3 PART III

#### 3.1 Questions

9. (a) let  $K_c(\mathbf{x}, \mathbf{z}) = c + \sum_{i=1}^n x_i z_i = c + \mathbf{x}^T \mathbf{z}$

Our function  $K_c(\cdot, \cdot)$  is psd iff:

$$\forall \{u_i\}_{1:m} \in \mathcal{R}, \{\mathbf{x}_j\}_{1:m} \in \mathcal{R}^n:$$

$$\sum_{k,l} u_k u_l K_c(\mathbf{x}_u, \mathbf{x}_l) \geq 0$$

We expand our kernel function to give:

$$\sum_{k,l} u_k u_l K_c(\mathbf{x}_u, \mathbf{x}_l) = \sum_{n,m} c u_n u_m + \sum_{n,m} \langle \mathbf{x}_n, \mathbf{x}_m \rangle u_n u_m$$

Here, we note that  $\langle \cdot, \cdot \rangle$  is an inner product over the Hilbert space  $\mathcal{R}^n$  and hence is psd by the representer theorem:

$$\implies \sum_{n,m} \langle \mathbf{x}_n, \mathbf{x}_m \rangle u_n u_m \geq 0 \quad (1)$$

**Proposition**  $K_c(\cdot, \cdot)$  is psd iff  $c \geq 0$

**Proof:**

( $\implies$ )

Suppose  $c \geq 0$ .

$$(1) \implies \sum_{k,l} u_k u_l K_c(\mathbf{x}_u, \mathbf{x}_l) \geq c \sum_{n,m} u_n u_m = c (\sum_n u_n)^2 \geq 0$$

( $\impliedby$ )

let  $c < 0$ . Suppose our vectors  $x_n$  are identically equal to  $(\sqrt{a}, 0)^T$ , for some  $a < |c|/n \implies x_k^T x_l = a$   
 $\forall k, l$ .

$$\implies \sum_{k,l} u_k u_l K_c(\mathbf{x}_u, \mathbf{x}_l) = (a + c) \sum_{n,m} u_n u_m = (a + c) (\sum_n u_n)^2 < 0$$

Hence for any  $c < 0$ ,  $\exists \{u_i\}_{1:m} \in \mathcal{R}, \{\mathbf{x}_j\}_{1:m} \in \mathcal{R}^n$  s.t:

$$\sum_{k,l} u_k u_l K_c(\mathbf{x}_u, \mathbf{x}_l) < 0 \square$$

- (b) Using this kernel in our ridge regression, we note that as the value of  $c$  increases, the value of the kernel converges to a single value,  $c$ . More concretely, the kernel between any 2 datapoints becomes relatively uniform. We may infer from this that adding a constant in our kernel function has a regularising effect, as the model.

Expanding on this we demonstrate that we converge to naive linear regression in the limit:

Note as  $c \rightarrow \infty$ ,  $K = XX^T + C1_{n \times n} \rightarrow C1_{n \times n} = C1_n 1_n^T$ , where  $1_{l,m}$  is an  $l \times m$  dimensional vector of all ones. Note the similarity here with naive linear regression, where  $X = 1_n$ .

$$\alpha = c^{-1} \text{least squares}(1_{n \times n}, y_{train}) \text{ (we no longer assume invertibility)}$$

$$\implies \hat{y}_{test} = c\alpha 1_{n_{train}, n_{test}} = 1_{n_{train}, n_{test}} \text{least squares}(1_{n \times n}, y_{train})$$

Hence for our predictions, the constants  $C$  cancel out and we are left with the exact dual form of naive linear regression:

$$\alpha' = \text{least squares}(1_{n \times n}, y_{\text{train}}) = \bar{y}_{\text{train}}$$

$$y_{\text{test}} = \alpha' 1_{n_{\text{train}}, n_{\text{test}}} = [\bar{y}_{\text{train}}]_{1:n_{\text{train}}} \quad \text{Hence, Hence, adding } c \text{ to our kernel acts like a regulariser.}$$

10. Define  $g(t) := \text{sgn}(f(t)) = \frac{f(t)}{|f(t)|}$

**Proposition:**

as  $\beta \rightarrow \infty$ ,  $g(t) \rightarrow 1 - NN$

**Proof:**

$$f(t) = \sum_i \alpha_i K(x_i, t), \text{ where } \alpha = (K + \gamma I_l)^{-1} y$$

since  $K(x, x) = \exp(0) = 1$  for all  $x$ , we decompose  $K$  as follows:

$$K = I_l + \tilde{K}, \text{ where } \tilde{K}_{ij} = K(x_i, x_j), i \neq j. \text{ We note here that as } \beta \rightarrow \infty, \tilde{K} \rightarrow 0.$$

We take a taylor expansion in  $\tilde{K}$ , to arrive at the following result:

$$\alpha = ((\gamma + 1)I_l + \tilde{K})^{-1} y = \frac{1}{\gamma + 1} (I_l - O(\tilde{K})) y \text{ Hence as } \beta \rightarrow \infty, \alpha \rightarrow \frac{1}{\gamma + 1} y$$

Hence our predictor function becomes:

$$f(t) \rightarrow \frac{1}{\gamma + 1} \sum_i y_i K(x_i, t) \implies g(t) \rightarrow \frac{\sum_i y_i K(x_i, t)}{|\sum_i y_i K(x_i, t)|}$$

Let's assume that there exists a unique point  $x$  in our training set of minimal distance to our test point.

Define  $d_i := \|x_{test} - x_i\|^2$ , and  $d_* := \min_i d_i$

$$g(t) = \frac{y_* \exp(-\beta d_*)}{|\sum_i y_i \exp(-\beta d_i)|} + \sum_{i \neq *} \frac{y_i \exp(-\beta d_i)}{|\sum_i y_i \exp(-\beta d_i)|} = \frac{y_*}{|y_* + \sum_{i \neq *} y_i \exp(-\beta(d_i - d_*))|} + \sum_{i \neq *} \frac{y_i \exp(-\beta(d_i - d_*))}{|y_* + \sum_{i \neq *} y_i \exp(-\beta(d_i - d_*))|}.$$

Note for  $d_i \neq d_*$ ,  $\exp(-\beta(d_i - d_*)) \rightarrow 0$  as  $\beta \rightarrow \infty$ . Hence:

$$g(t) \rightarrow \frac{y_*}{|y_* + 0|} + \frac{0}{|y_* + 0|} = y_* = y_i : x_i = \min_i \|x_{test} - x_i\| \text{ which is the required 1-NN predictor.}$$

11. Given a whack-a-mole grid of size  $n$ , it is solvable if we can apply a sequence from our  $n^2$  permissible moves that results in a blank screen. We first index all permissible moves by  $T_{i,j}$ , which has the effect of flipping the state of squares  $(i, j), (i - 1, j), (i + 1, j), (i, j - 1), (i, j + 1)$ , as long as each coordinate is still on the grid.

We may consider each grid as an element of the space of square matrices over  $\mathcal{F}_2$ , the finite field of order 2: under this construction, we see that the application of a move  $T_{i,j}$  corresponds to the sum between the grid and a matrix  $T_{ij}$ , where the entries of  $T_{ij}$  are zero everywhere except at  $(i, j), (i - 1, j), (i + 1, j), (i, j - 1), (i, j + 1)$ , if these values are present on the grid. For example, if  $n = 4$ , the transformation  $T_{2,2}$  corresponds to the addition of the matrix  $T_{22}$ :

$$T_{22} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Since our operations on a grid are now simply the sum of matrices over a  $\mathcal{F}_2$ , it is now immediately obvious that our operations are commutative, by the commutativity of addition.

Further, since applying the same operation twice would be the same as applying the matrix twice in order, we can see that any operation applied twice would have the effect of applying no operation at all (since  $T_{ij} + T_{ij} = 0$ ).

Hence, we conclude for  $1 \leq i, j \leq n$ , each operation is applied at most once. From this point, we note that there is no reason for our operations to be square matrices, as the only required properties of the space are those of a vector space (we make no use here of matrix multiplication). Hence, we may vectorise both our input grid and our operation matrices to get a set of vectors over  $F_2^{n^2}$ . We now index our operation vectors  $t_i : i = 1, \dots, n^2$ .

Let  $x_i = \mathcal{I}[t_i \text{ is used}]$  be an indicator function representing whether  $t_i$  is added to the input to get our blank grid.

We observe that our problem is solvable for a given input grid  $G$  (now vectorised as a vector  $\mathbf{g}$ ) iff there exists a vector  $\mathbf{x} \in F_2^{n^2}$  st:

$$\sum_i t_i x_i + \mathbf{g} = \mathbf{0}.$$

If we stack the vectors  $t_i$  horizontally as the columns of a matrix  $T$ , we get:

$$T\mathbf{x} + \mathbf{g} = \mathbf{0}$$

Since every element is its own additive inverse over  $\mathcal{F}_2$ , we add  $\mathbf{g}$  to both sides and arrive at the linear system  $T\mathbf{x} = \mathbf{g}$ .

Since the matrix  $T$  may not be invertible in general, we perform gaussian elimination over  $\mathcal{F}_2$ , i.e create an augmented matrix  $[T|\mathbf{g}]$  and reduce to row echelon form. If the problem is solvable, we will end up with a consistent (but perhaps underdetermined) reduced system of equations. If the input has no solution, the reduced system of equations will have some inconsistency in its variables: there will be some row of the reduced row echelon augmented matrix with all zeros to the left of the augmentation line, and a non-zero entry to the right of the augmentation line.

Finally, we make a note of the time complexity of this algorithm: the creation of our matrix  $T$  can be performed in  $O(n^4)$  since we have  $n^2$  operation grids of size  $n^2$ . Further, gaussian elimination can be performed in  $O(n^3)$  operations in any field, and hence our total running time is  $O(n^3)$ .

Once we have determined whether a grid configuration is solvable, it remains to find a valid solution.

We note that if a solution exists, then a least squares solver should find it. This operation should take  $O(n^3)$  operations.