National University of Singapore
School of Computing
CS2109S: Introduction to AI and Machine Learning
Semester II, 2022/2023

# Tutorial 6
# SVMs and Regularisation

These questions will be discussed during the tutorial session. Please be prepared to answer them.

## Summary of Key Concepts

In this tutorial, we will discuss and explore the following learning points from Lecture:

1. Visualising Regularisation
2. Hinge Loss in Support Vector Machines (SVM)
3. Bias and Variance

## Problems

### L1-norm vs. L2-norm.

In lecture, we discussed using regularization on linear regression using the L2-norm. This is also called Ridge Regression, and the cost function is:

$$J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_w(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=1}^{n} {w_i}^2\right]$$

It is also possible to do regularization using the L1-norm. This is called Lasso Regression, and the resulting cost function is:

$$J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_w(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=1}^{n} |w_i|\right]$$

Now, in this problem, we will investigate how regularization will work with these 2 norms in a 2D space. The figures below shows contour plots for a linear regression problem with the 2 different regularizers. The elliptic contours represent the squared error term. The diamond (Figure 1) and circle (Figure 2) contours represent the regularisation penalty term when $\lambda = 5$.

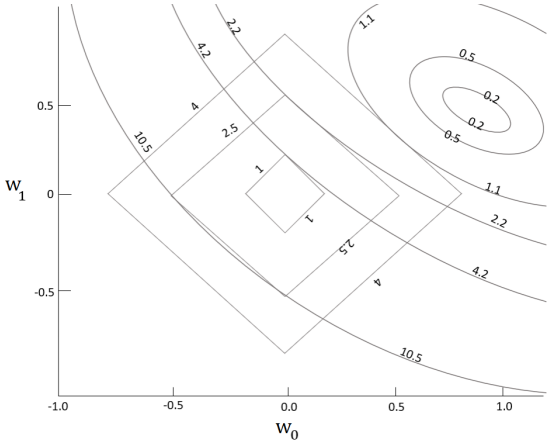


Figure 1: Contour plots for the linear regression problem with L1 regularisation.

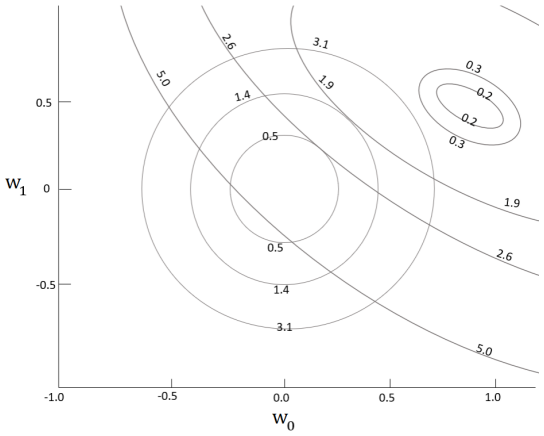


Figure 2: Contour plots for the linear regression problem with L2 regularisation.

Along a contour, the corresponding loss remains the same as $w_0$ and $w_1$ vary. Intersections between the 2 different contours represent points with possible values for $w_0$ and $w_1$. The total value of the loss function at such a point is the sum of the two contours values. For example, for the point $(w_0 = 0.0,\ w_1 = -0.5)$ in Figure 1, the total loss (regularization and error loss) is $2.5 + 10.5 = 13$.

**1.** For each of the following cases, provide an estimate of the optimal values of $w_0$ and $w_1$ using the figures as reference.

1. No regularisation.
2. L1 regularisation with $\lambda = 5$.
3. L2 regularisation with $\lambda = 5$.

Additionally, what do you think makes L2 Regularisation different from L1 Regularisation in terms of what they do to the parameters?

***You have to check for different values of regularisation cost and squared error term to see which gives the smallest sum****. Also note that the contours (as given in the problem description,) already factor in $\lambda = 5$, so there is no need to multiply by $\lambda$ when computing the sum.*

**Solution:**

1. $w_0 = 0.9$, $w_1 = 0.5$. *Cost: approx 0 (no MSE and no regularization penalty).*
2. $w_0 = 0.0$, $w_1 = 0.5$. *Cost: 4.7 = 2.2 (MSE) + 2.5 (L1 penalty).*
3. $w_0 = 0.2$, $w_1 = 0.25$. *Cost: 3.1 = 2.6 (MSE) + 0.5 (L2 penalty).*

*In contrast to L1 norm regularization, L2 norm regularization heavily penalizes larger parameters. L2 thus attempts to pull* **all** *parameters towards small values. In contrast, L1 regularisation may set values of certain parameters, which are multiplied with less important features, to zero, while others are set to some non-zero value (that are possibly large). Observe that by setting some parameter values to zero, L1 regularization implicitly selects features to be 'excluded'. Therefore, it is a feature selection strategy.*

**Understanding SVM and Hinge Loss.**

In class, we formulated SVMs as the following maximization problem:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha^{(i)} \alpha^{(j)} \bar{y}^{(i)} \bar{y}^{(j)} K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)$$

subject to $\forall i \ \alpha^{(i)} \geq 0$ and $\sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} = 0$, where $K(\mathbf{x}, \mathbf{l}) = \mathbf{x} \cdot \mathbf{l}$ for a linear SVM classifier.

**2.** Consider the data points given in tabular and graphical form (Figure 3) below:

| $i$ | $x_1^{(i)}$ | $x_2^{(i)}$ | $\bar{y}^{(i)}$ |
|---|---|---|---|
| 1 | -2 | -2 | -1 |
| 2 | -2 | 0 | -1 |
| 3 | 0 | 2 | 1 |
| 4 | 1 | 1 | 1 |
| 5 | 3 | 0 | 1 |

The red and blue points correspond to points with $\bar{y} = -1$ and $\bar{y} = 1$ respectively. The decision boundary for a linear model on this data would be the function $h(x_1, x_2) = \sum_{i=1}^{5} \alpha^{(i)} \bar{y}^{(i)} \left(x_1 x_1^{(i)} + x_2 x_2^{(i)}\right) + b$.

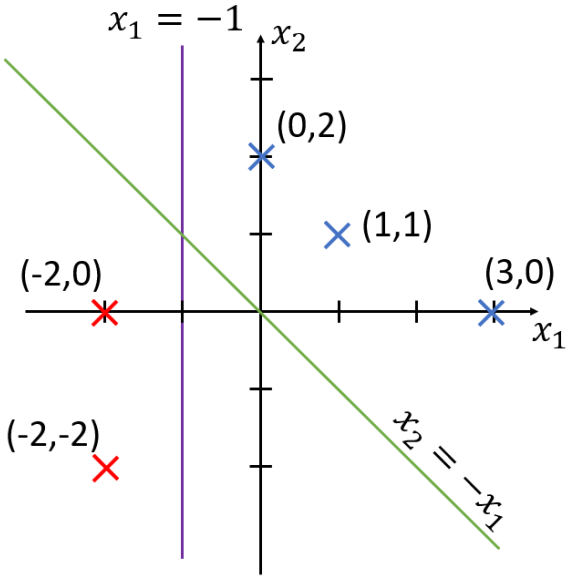


Figure 3: SVM example.

**(a)** The two lines (green and purple) represent decision boundaries of 2 different linear models. How can we parametrize the lines, i.e. what are the values for $\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}, \alpha^{(4)}, \alpha^{(5)}, b$ for the 2 lines?

**Solution:** *For the green line, the equation is* $x_2 = -x_1 \Rightarrow x_1 + x_2 + 0 = 0$. *More generally, we can write the equation as* $kx_1 + kx_2 + 0 = 0$ *for any non-zero real number* $k$. *Comparing with* $\mathbf{w}^\intercal \mathbf{x} + b$, *we have* $\mathbf{w} = [k, k]^\intercal$ *and* $b = 0$. *We can form the following system of equations using the constraints and equation for the weights.*

$$-\alpha^{(1)} \begin{bmatrix} -2 \\ -2 \end{bmatrix} - \alpha^{(2)} \begin{bmatrix} -2 \\ 0 \end{bmatrix} + \alpha^{(3)} \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \alpha^{(4)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha^{(5)} \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} k \\ k \end{bmatrix} \tag{1}$$

$$-\alpha^{(1)} - \alpha^{(2)} + \alpha^{(3)} + \alpha^{(4)} + \alpha^{(5)} = 0 \tag{2}$$

*The first equation comes from* $\mathbf{w} = \sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} \mathbf{x}^{(i)}$ *and the second one comes from* $\sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} = 0$*. Splitting up equation 1 and combining with equation 2 we get*

$$\begin{bmatrix} 2 & 2 & 0 & 1 & 3 \\ 2 & 0 & 2 & 1 & 0 \\ -1 & -1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha^{(1)} \\ \alpha^{(2)} \\ \alpha^{(3)} \\ \alpha^{(4)} \\ \alpha^{(5)} \end{bmatrix} = \begin{bmatrix} k \\ k \\ 0 \end{bmatrix}$$

*We form the augmented matrix and perform Gauss-Jordan Elimination to obtain the following matrix*

$$\left[\begin{array}{ccccc|c} 1 & 0 & 0 & -1 & -5/2 & 0 \\ 0 & 1 & 0 & 3/2 & 4 & k/2 \\ 0 & 0 & 1 & 3/2 & 5/2 & k/2 \end{array}\right]$$

*From the above augmented matrix, we get the solutions* $\alpha^{(4)} = c_1, \alpha^{(5)} = c_2, \alpha^{(1)} = c_1 + \frac{5c_2}{2}, \alpha^{(2)} = \frac{k}{2} - \frac{3c_1}{2} - 4c_2, \alpha^{(3)} = \frac{k}{2} - \frac{3c_1}{2} - \frac{5c_2}{2}$*, where* $c_1$ *and* $c_2$ *are any non-negative real numbers and* $k \geq 3c_1 + 8c_2$ *so that* $\alpha^{(2)} \geq 0$ *and* $\alpha^{(3)} \geq 0$ *to satisfy the constraint that* $\alpha^{(i)} \geq 0$*.*

*For the purple line, the equation is* $x_1 = -1 \Rightarrow x_1 + 0x_2 + 1 = 0$*. More generally, we can write the equation as* $kx_1 + 0x_2 + k = 0$ *for any non-zero real number* $k$*. Comparing with* $\mathbf{w}^\intercal \mathbf{x} + b$*, we have* $\mathbf{w} = [k, 0]^\intercal$ *and* $b = k$*. We can form the following system of equations using the constraints and equation for the weights.*

$$-\alpha^{(1)} \begin{bmatrix} -2 \\ -2 \end{bmatrix} - \alpha^{(2)} \begin{bmatrix} -2 \\ 0 \end{bmatrix} + \alpha^{(3)} \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \alpha^{(4)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha^{(5)} \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} k \\ 0 \end{bmatrix} \tag{3}$$

$$-\alpha^{(1)} - \alpha^{(2)} + \alpha^{(3)} + \alpha^{(4)} + \alpha^{(5)} = 0 \tag{4}$$

*The first equation comes from* $\mathbf{w} = \sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} \mathbf{x}^{(i)}$ *and the second one comes from* $\sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} = 0$*. Splitting up equation 3 and combining with equation 4 we get*

$$\begin{bmatrix} 2 & 2 & 0 & 1 & 3 \\ 2 & 0 & 2 & 1 & 0 \\ -1 & -1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha^{(1)} \\ \alpha^{(2)} \\ \alpha^{(3)} \\ \alpha^{(4)} \\ \alpha^{(5)} \end{bmatrix} = \begin{bmatrix} k \\ 0 \\ 0 \end{bmatrix}$$

*We form the augmented matrix and perform Gauss-Jordan Elimination to obtain the following matrix*

$$\left[\begin{array}{ccccc|c} 1 & 0 & 0 & -1 & -5/2 & -k/2 \\ 0 & 1 & 0 & 3/2 & 4 & k \\ 0 & 0 & 1 & 3/2 & 5/2 & k/2 \end{array}\right]$$

*From the above augmented matrix, we get the solutions* $\alpha^{(4)} = c_1, \alpha^{(5)} = c_2, \alpha^{(1)} = c_1 + \frac{5c_2}{2} - \frac{k}{2}, \alpha^{(2)} = k - \frac{3c_1}{2} - 4c_2, \alpha^{(3)} = \frac{k}{2} - \frac{3c_1}{2} - \frac{5c_2}{2}$*, where* $c_1$ *and* $c_2$ *are any non-negative real numbers and* $k \leq 2c_1 + 5c_2$*,* $k \geq \frac{3c_1}{2} + 4c_2$ *and* $k \geq 3c_1 + 5c_2$ *so that* $\alpha^{(1)} \geq 0$*,* $\alpha^{(2)} \geq 0$ *and* $\alpha^{(3)} \geq 0$ *to satisfy the constraint that* $\alpha^{(i)} \geq 0$*. Using the inequalities of* $k$*, since* $\frac{3c_1}{2} + 4c_2 \leq 3c_1 + 5c_2$ *as both* $c_1 \geq 0$ *and* $c_2 \geq 0$*, we simplify the three inequalities to get* $k \leq 2c_1 + 5c_2$ *and* $k \geq 3c_1 + 5c_2$*. Combining the two inequalities,* $3c_1 + 5c_2 \leq k \leq$

$2c_1 + 5c_2 \Rightarrow 3c_1 \leq k - 5c_2 \leq 2c_1$ *However, for this inequality to be consistent,* $c_1 = 0$ *and* $k = 5c_2$*. So the final solution is* $\alpha^{(1)} = \alpha^{(3)} = \alpha^{(4)} = 0$ *and* $\alpha^{(2)} = \alpha^{(5)} = \frac{k}{5}$ *for any positive real number* $k$*.*

Recall that when we were deriving the SVM optimization problem we came across the loss function

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \max\{0, 1 - \bar{y}^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)\}$$

The term $\max\{0, 1 - \bar{y}^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)\}$ is called the hinge loss.

Lets now compute the total losses associated with purple line using the above loss function with $C = 1$. The purple line has equation $x_1 = -1 \Rightarrow kx_1 + 0x_2 + k = 0$, for any non-zero real number $k$, so $\boldsymbol{w} = (k, 0)$ and $b = k$ to follow the form $\mathbf{w}^\intercal\mathbf{x} + b = 0$. The hinge loss is:

$$\begin{aligned}
&\max(0, 1 + (-2k + 0 + k)) + \max(0, 1 + (-2k + 0 + k)) + \max(0, 1 - (0 + 0 + k)) \\
&\qquad + \max(0, 1 - (k + 0 + k)) + \max(0, 1 - (3k + 0 + k)) \\
&= 3 \cdot \max(0, 1 - k) + \max(0, 1 - 2k) + \max(0, 1 - 4k)
\end{aligned}$$

Hence, the total loss is $3 \cdot \max(0, 1 - k) + \max(0, 1 - 2k) + \max(0, 1 - 4k) + \frac{k^2+0^2}{2}$. This is minimized when $k = 1$, which results in a total loss of $0.5$.

**(b)** Calculate the total loss for the green line in in a similar manner. Also find the parameter(s) that result in the least loss.

**Solution:** *The green line has equation* $x_1 = -x_2 \Rightarrow kx_1 + kx_2 + 0 = 0$*, for any non-zero real number* $k$*, so* $\mathbf{w} = (k, k)$ *and* $b = 0$*. Hinge loss is:*

$$\begin{aligned}
&\max(0, 1 - 4k) + \max(0, 1 - 2k) + \max(0, 1 - 2k) + \max(0, 1 - 2k) + \max(0, 1 - 3k) \\
&\qquad = 3 \cdot \max(0, 1 - 2k) + \max(0, 1 - 3k) + \max(0, 1 - 4k)
\end{aligned}$$

*and total loss is* $3 \cdot \max(0, 1 - 2k) + \max(0, 1 - 3k) + \max(0, 1 - 4k) + k^2$*. Minimized at* $0.25$ *when* $k = 0.5$*.*

**(c)** Which line is a better solution to the SVM?

**Solution:** *The green line is a better solution to the SVM as both lines completely separate the sets without mislabels, but the green line that has the larger margin of* $\sqrt{2}$ *units compared to the purple line's margin of 1 unit.*

**(d)** (Optional) Solve $\max_{\boldsymbol{\alpha}} \sum_{i=1}^n \alpha^{(i)} - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha^{(i)}\alpha^{(j)}\bar{y}^{(i)}\bar{y}^{(j)}\left(\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}\right)$ using the possible values of $\alpha^{(i)}$ found in part **(a)** for the green line. Using the equation $\mathbf{w} = \sum_{i=1}^n \alpha^{(i)}\bar{y}^{(i)}\mathbf{x}^{(i)}$, what can you conclude about the value of $\mathbf{w}$ found in part **(b)** and the one calculated here?

**Solution:** *For the optimal solution, the non-support vectors,* $\mathbf{x}^{(1)}$ *and* $\mathbf{x}^{(5)}$*, have 0 as their corresponding* $\alpha^{(i)}$*. Using the solutions found above, since* $\alpha^{(5)} = 0$ *and* $\alpha^{(1)} = 0$*,* $c_2 = 0$ *and thus* $c_1 = 0$*. So the solution space simplifies to* $\alpha^{(2)} = \alpha^{(3)} = k$ *for any*

*non-zero real number $k$ and $\alpha^{(1)} = \alpha^{(4)} = \alpha^{(5)} = 0$.*

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha^{(i)} \alpha^{(j)} \bar{y}^{(i)} \bar{y}^{(j)} \left( \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \right) \\
&= \max_{\boldsymbol{\alpha}} \alpha^{(2)} + \alpha^{(3)} - \frac{1}{2} [\alpha^{(2)} \alpha^{(2)} \bar{y}^{(2)} \bar{y}^{(2)} \left( \mathbf{x}^{(2)} \cdot \mathbf{x}^{(2)} \right) + \alpha^{(2)} \alpha^{(3)} \bar{y}^{(2)} \bar{y}^{(3)} \left( \mathbf{x}^{(2)} \cdot \mathbf{x}^{(3)} \right) \\
&\qquad + \alpha^{(3)} \alpha^{(2)} \bar{y}^{(3)} \bar{y}^{(2)} \left( \mathbf{x}^{(3)} \cdot \mathbf{x}^{(2)} \right) + \alpha^{(3)} \alpha^{(3)} \bar{y}^{(3)} \bar{y}^{(3)} \left( \mathbf{x}^{(3)} \cdot \mathbf{x}^{(3)} \right)] \\
&= \max_{k} k + k - \frac{1}{2} [(k)(k)(-1)(-1) \begin{bmatrix} -2 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 0 \end{bmatrix} + (k)(k)(-1)(1) \begin{bmatrix} -2 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 2 \end{bmatrix} \\
&\qquad + (k)(k)(1)(-1) \begin{bmatrix} 0 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 0 \end{bmatrix} + (k)(k)(1)(1) \begin{bmatrix} 0 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 2 \end{bmatrix}] \\
&= \max_{k} 2k - \frac{1}{2} [4k^2 + 0 + 0 + 4k^2] \\
&= \max_{k} 2k - 4k^2
\end{aligned}$$

*Differentiating and setting to 0, we get $k = \frac{1}{4}$.*

$$\begin{aligned}
\mathbf{w} &= \sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} \mathbf{x}^{(i)} \\
&= \frac{1}{4}(-1) \begin{bmatrix} -2 \\ 0 \end{bmatrix} + \frac{1}{4}(1) \begin{bmatrix} 0 \\ 2 \end{bmatrix} \\
&= \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}
\end{aligned}$$

*We notice that they are exactly the same! This is the case since the above loss function was used in the derivation of the SVM maximization problem.*

**Understanding Bias & Variance.**

In lecture, we discussed how the loss varies with the degree of a polynomial (that is used as the hypothesis), and with $\lambda$, when there is regularization. In this problem, we will investigate how loss varies with the number of training samples under different conditions. Consider a dataset like the following set of points, where we know that the "correct" hypothesis is a 2nd-degree polynomial.

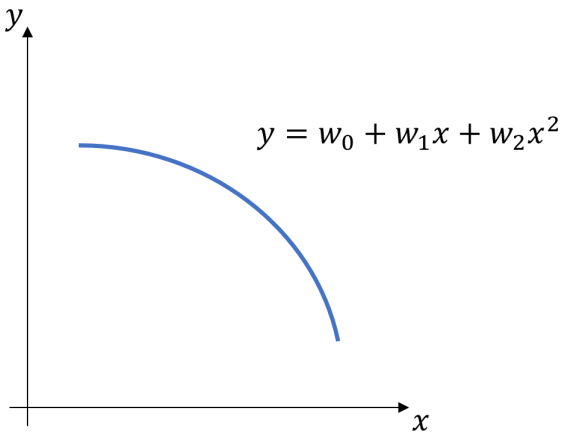


Figure 4: Sample data points.

**3.** Two different models were trained on the dataset many times, gradually increasing the number of samples used in training. In each iteration, the training error and test error were recorded. The model hypotheses are as below:

1. $H_w(x) = w_0 + w_1x$
2. $H_w(x) = w_0 + w_1x + w_2x^2 + \cdots + w_{10}x^{10}$

The training and test errors obtained were plotted for each model, and the resulting graphs are as below:

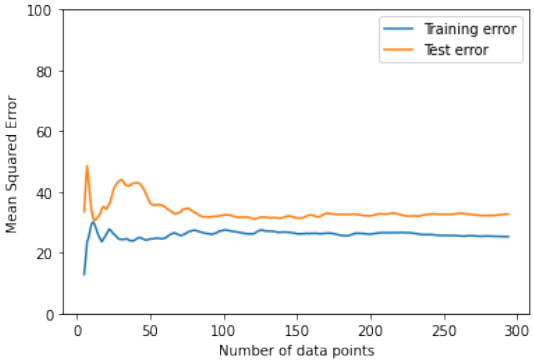


Figure 5: Learning curves for Model X.

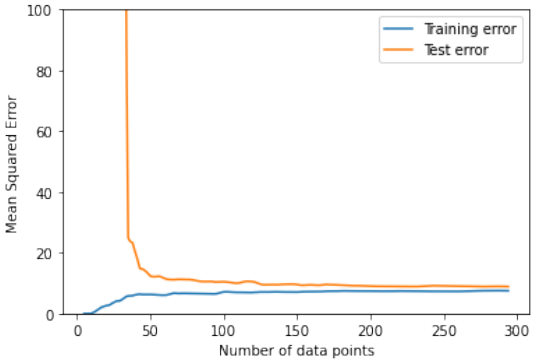


Figure 6: Learning curves for Model Y.

**(a)** Between the two graphs, which one indicates a model with a higher bias? How does bias seem to vary with the number of samples?

**Solution:** *Figure 5, Model X. Relatively higher error, even as samples increase, indicates inability to capture the true relationship sufficiently, hinting high bias. Bias does not (generally) improve with increase in number of samples.*

**(b)** Which graph indicates a model with a higher variance? How does variance seem to vary with the number of samples?

**Solution:** *Figure 6, Model Y. Lower error, but initially higher difference between the 2 error indicates high variance. Getting more data points is likely to help variance.*

**(c)** Which model do you think each graph belongs to. Explain your reasoning.

**Solution:** *Model X (high bias) is the linear model, because: linear model can't capture quadratic relationship, has high bias. Model Y (high var) is the high degree polynomial, because: overfits the points so initially high difference in errors. As number of samples increases, the "degree of overfitting" reduces approaching a roughly quadratic curve.*

**Extra:** The models above are un-regularized. How might regularization affect the graphs for each of them.

**Solution:** *Can share any of the effects regularization usually has, like: If Lasso, might improve high degree polynomial model by ignoring higher degree terms*

**Extra: Proof that the Gaussian Kernel has Infinite Dimensional Features.**

In Lecture, it was mentioned that we can have an SVM with infinite-dimensional features. The Gaussian kernel is actually one such kernel. Recall that the Gaussian kernel, $K(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$. In this proof, we will set $\sigma = 1$ to turn the denominator into 2 to simplify our expressions later. The same can be shown with any $\sigma$.

We begin by expanding $\|\mathbf{x} - \mathbf{y}\|^2$. Let $\mathbf{x} = [x_1, x_2, \ldots, x_m]^\intercal$ and $\mathbf{y} = [y_1, y_2, \ldots, y_m]^\intercal$.

$$\begin{aligned}
\|\mathbf{x} - \mathbf{y}\|^2 &= (x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_m - y_m)^2 \text{ by definition of L2-norm} \\
&= (x_1^2 + x_2^2 + \ldots + x_m^2) + (y_1^2 + y_2^2 + \ldots + y_m^2) - (2x_1y_1 - 2x_2y_2 - \ldots - 2x_my_m) \\
&= \mathbf{x} \cdot \mathbf{x} + \mathbf{y} \cdot \mathbf{y} - 2\mathbf{x} \cdot \mathbf{y}
\end{aligned}$$

Substituting into $K(\mathbf{x}, \mathbf{y})$ we get,

$$\begin{aligned}
K(\mathbf{x}, \mathbf{y}) &= e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2}} \\
&= e^{-\frac{\mathbf{x}\cdot\mathbf{x}+\mathbf{y}\cdot\mathbf{y}-2\mathbf{x}\cdot\mathbf{y}}{2}} \\
&= e^{-\frac{\mathbf{x}\cdot\mathbf{x}}{2}}\, e^{\mathbf{x}\cdot\mathbf{y}}\, e^{-\frac{\mathbf{y}\cdot\mathbf{y}}{2}}
\end{aligned}$$

Using the power series for $e^x$,

$$e^{\mathbf{x}\cdot\mathbf{y}} = 1 + \mathbf{x} \cdot \mathbf{y} + \frac{1}{2!}(\mathbf{x} \cdot \mathbf{y})^2 + \frac{1}{3!}(\mathbf{x} \cdot \mathbf{y})^3 + \frac{1}{4!}(\mathbf{x} \cdot \mathbf{y})^4 + \ldots$$

which is a weighted sum of polynomial kernels. Notice that when adding kernels together, the feature map of both kernels are concatenated, e.g., $(\mathbf{u}\cdot\mathbf{v})+(\mathbf{u}\cdot\mathbf{v})^2 = [u_1, u_2]^\intercal \cdot [v_1, v_2]^\intercal + [u_1^2, \sqrt{2}u_1u_2, u_2^2]^\intercal \cdot [v_1^2, \sqrt{2}v_1v_2, v_2^2]^\intercal = [u_1, u_2, u_1^2, \sqrt{2}u_1u_2, u_2^2]^\intercal \cdot [v_1, v_2, v_1^2, \sqrt{2}v_1v_2, v_2^2]^\intercal$ using the example in the lecture slides.

Let $\phi_i(\mathbf{x})$ be the feature map of the degree $i$ polynomial kernel and $\mathbf{x} \oplus \mathbf{y}$ be the concatenation operation. We get

$$\begin{aligned}
e^{\mathbf{x}\cdot\mathbf{y}} = &\left[[1] \oplus \phi_1(\mathbf{x}) \oplus \frac{1}{2!}\phi_2(\mathbf{x}) \oplus \frac{1}{3}\phi_3(\mathbf{x}) \oplus \frac{1}{4!}\phi_4(\mathbf{x}) \oplus \ldots\right] \\
&\cdot \left[[1] \oplus \phi_1(\mathbf{y}) \oplus \frac{1}{2!}\phi_2(\mathbf{y}) \oplus \frac{1}{3}\phi_3(\mathbf{y}) \oplus \frac{1}{4!}\phi_4(\mathbf{y}) \oplus \ldots\right]
\end{aligned}$$

Substituting back into $K(\mathbf{x}, \mathbf{y})$, we have

$$\begin{aligned}
K(\mathbf{x}, \mathbf{y}) = e^{-\frac{\mathbf{x}\cdot\mathbf{x}}{2}} &\left[[1] \oplus \phi_1(\mathbf{x}) \oplus \frac{1}{2!}\phi_2(\mathbf{x}) \oplus \frac{1}{3}\phi_3(\mathbf{x}) \oplus \frac{1}{4!}\phi_4(\mathbf{x}) \oplus \ldots\right] \\
&\cdot e^{-\frac{\mathbf{y}\cdot\mathbf{y}}{2}} \left[[1] \oplus \phi_1(\mathbf{y}) \oplus \frac{1}{2!}\phi_2(\mathbf{y}) \oplus \frac{1}{3}\phi_3(\mathbf{y}) \oplus \frac{1}{4!}\phi_4(\mathbf{y}) \oplus \ldots\right]
\end{aligned}$$

which is in the form of a dot product of feature maps, $\phi(\mathbf{x}) \cdot \phi(\mathbf{y})$ where

$$\phi(\mathbf{v}) = e^{-\frac{\mathbf{v}\cdot\mathbf{v}}{2}}([1] \oplus \phi_1(\mathbf{v}) \oplus (\frac{1}{2!}\phi_2(\mathbf{v})) \oplus (\frac{1}{3!}\phi_3(\mathbf{v})) \oplus (\frac{1}{4!}\phi_4(\mathbf{v})) \oplus \ldots)$$

We can see that the feature map for the Gaussian kernel contains all the features of every degree polynomial kernel's feature map.