CS2109S Tutorial 6

SVMs and Regularisation

(AY 24/25 Semester 2)

March 21, 2025

(Prepared by Benson)

Contents

Recap: Regularisation

Why do we need regularisation?

A. To decrease model complexity.

B. To improve training time efficiency.

C. To avoid underfitting.

D. To avoid overfitting.

Recap: Regularisation

How does regularisation address overfitting?

A. By penalizing weights for transformed features.
B. By reducing noise in the training data.
C. By penalizing large weights.
D. By reducing the number of transformed features.

Q1. L1-norm vs. L2-norm.

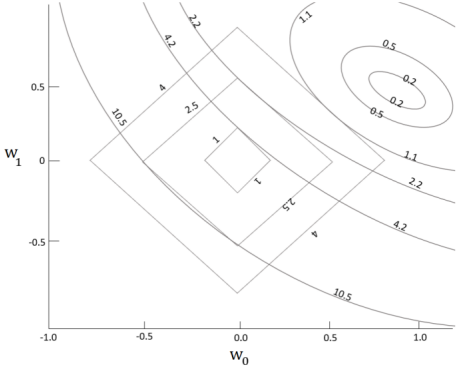
L2-norm (Ridge Regression):

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

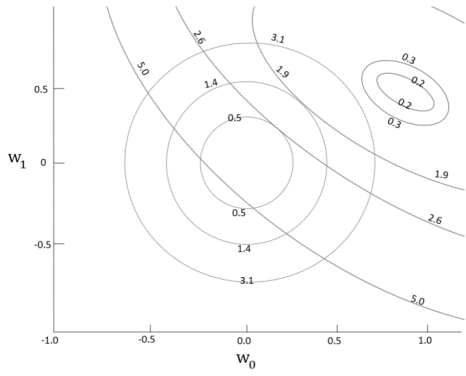
L1-norm (Lasso Regression):

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

Q1. L1-norm vs. L2-norm.



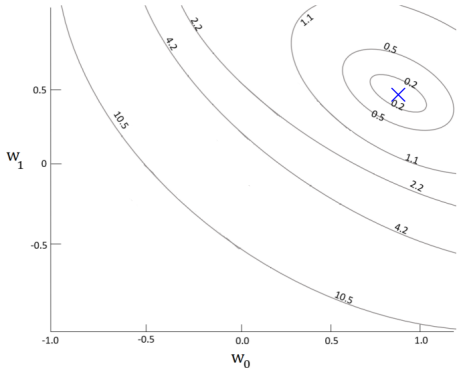
L1 regularisation



L2 regularisation

Q1. L1-norm vs. L2-norm.

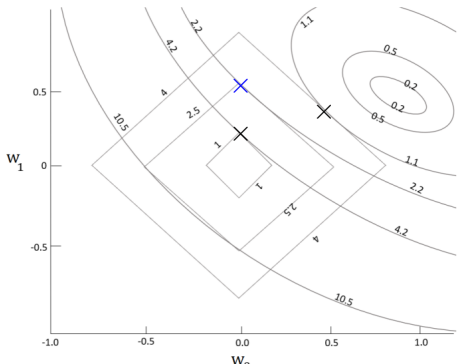
(a) No regularisation.



$w_0 = 0.9, w_1 = 0.5$
Cost ≈ 0

Q1. L1-norm vs. L2-norm.

(b) L1 regularisation with $\lambda = 5$.



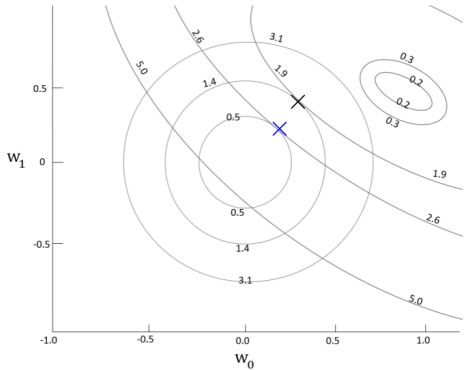
Total cost of the three points (from left to right):

- $4.2 + 1 = 5.2$
- $2.2 + 2.5 = 4.7$
- $1.1 + 4 = 5.1$

$w_0 = 0.0, w_1 = 0.5$
Cost $= 4.7$

Q1. L1-norm vs. L2-norm.

(c) L2 regularisation with $\lambda = 5$.

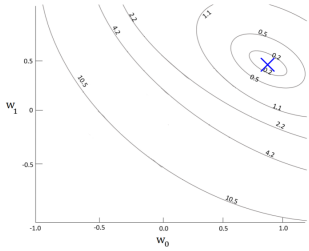


Total cost of the two points (from left to right):

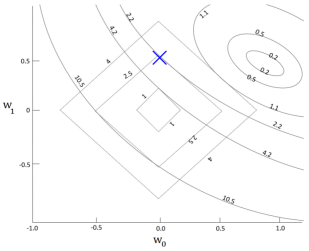
- $0.5 + 2.6 = 3.1$
- $1.9 + 1.4 = 3.3$

$w_0 = 0.2, w_1 = 0.25$
Cost $= 3.1$

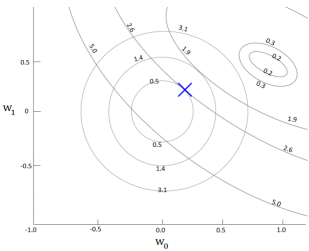
Q1. L1-norm vs. L2-norm.



No regularisation
$w_0 = 0.9, w_1 = 0.5$
Cost ≈ 0



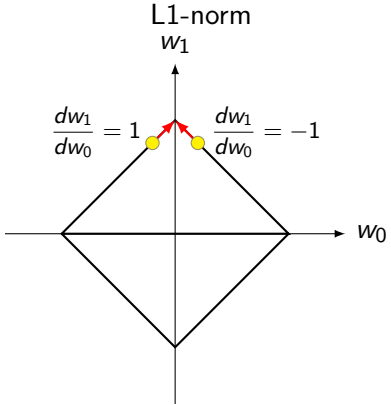
L1 regularisation
$w_0 = 0.0, w_1 = 0.5$
Cost $= 4.7$



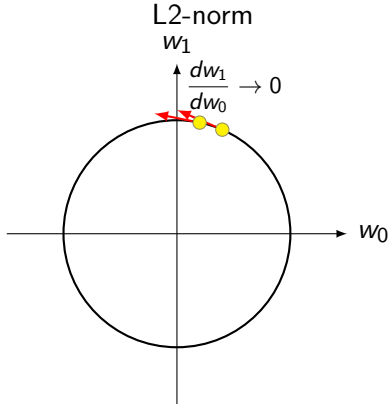
L2 regularisation
$w_0 = 0.2, w_1 = 0.25$
Cost $= 3.1$

- L2 heavily penalizes larger parameters, preferring **all** smaller values.
- L1 may set values of certain parameters to 0 (why?).

Q1. L1-norm vs. L2-norm.



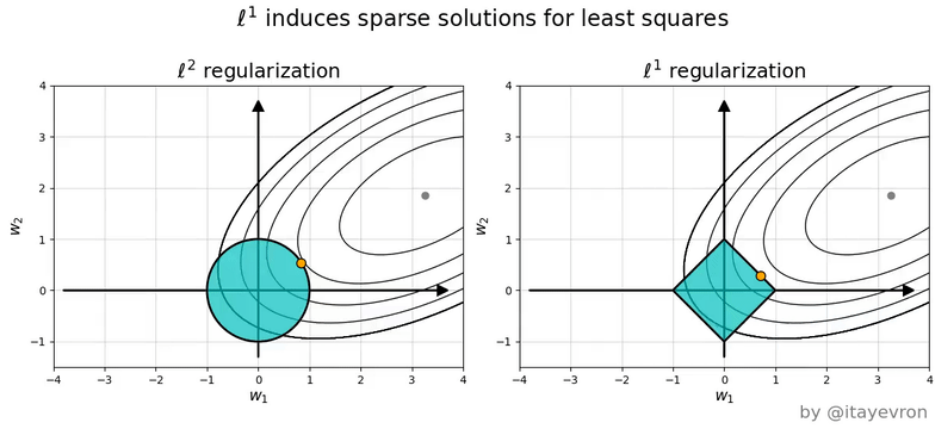
If w_1 is more important than w_0, pushing towards w_1 is free lunch ("one for one").



As $w_0 \to 0$, pushing towards w_1 has little gain ($\because$ push will be "less aggressive").

Q1. L1-norm vs. L2-norm.

Animation (See HTML slides):



Q1. L1-norm vs. L2-norm.

L2-norm (Ridge Regression):

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

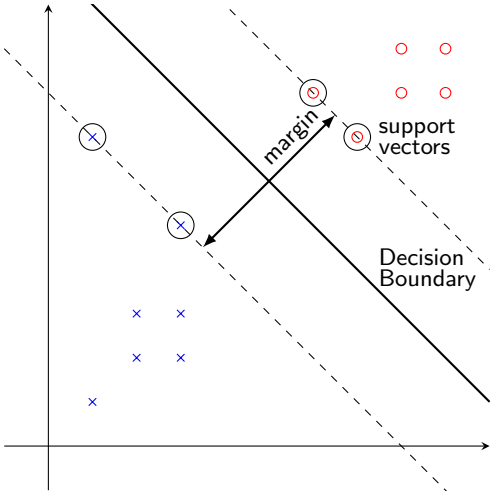
- L2 heavily penalizes larger parameters, preferring **all** smaller values.

L1-norm (Lasso Regression):

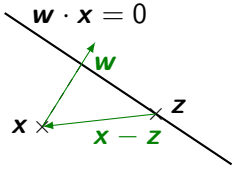
$$J(\boldsymbol{w}) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}) - y^{(i)})^2\right] + \lambda\sum_{i=1}^{n}|\boldsymbol{w}_i|$$

- L1 may set values of certain parameters to 0 $\rightarrow$ effectively "feature selection" (select the more important features, and zero out the rest).

Recap: Support Vector Machines



Want to maximize the **margin**.



1. Find a point $\boldsymbol{z}$ on $\boldsymbol{w} \cdot \boldsymbol{x} = 0$.
   (Can always take $\boldsymbol{z} = \boldsymbol{0}$)
2. Distance from $\boldsymbol{x}$ to $\boldsymbol{w} \cdot \boldsymbol{x} = 0$
   $= \dfrac{|(\boldsymbol{x} - \boldsymbol{z}) \cdot \boldsymbol{w}|}{\|\boldsymbol{w}\|}$
   (Length of projection from $\boldsymbol{x} - \boldsymbol{z}$ to $\boldsymbol{w}$)

Recap: Support Vector Machines

Consider the decision boundary $\boldsymbol{w} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$.

Given that the sample $\boldsymbol{x} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ is a support vector, what is the size of the **margin**?

A. 2.2

B. 2.8

C. 4.4

D. 5.6

E. None of the above

Solution. We have $z = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Distance from x to decision boundary

$$= \frac{1}{3^2 + 4^2} \left(\begin{bmatrix} 2 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right) \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \frac{6 \cdot 3 + 8 \cdot 4}{5} = 2.8.$$

$\therefore$ The size of the margin is $2.8 \times 2 = 5.6$.

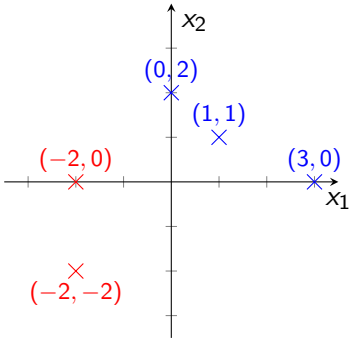
Recap: Support Vector Machines

Primal Formulation:

$$\begin{aligned} &\text{minimize}_{\boldsymbol{w}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to } y^{(i)}(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)}) \geq 1 \end{aligned}$$

- **Convex** optimization problem with linear constraints.
- Slow for **high-dimensional data** (computing $\boldsymbol{w} \cdot \boldsymbol{x}^{(i)}$ requires $O(m)$ time, $m =$ number of features).

Recap: Support Vector Machines



SVM Example

$$\text{minimize}_{\boldsymbol{w}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 \qquad \blacktriangleleft \text{ quadratic}$$

$$\text{subject to } (-1)\left(\boldsymbol{w} \cdot \begin{bmatrix}-2\\-2\end{bmatrix}\right) \geq 1 \qquad \blacktriangleleft \text{ linear}$$

$$(-1)\left(\boldsymbol{w} \cdot \begin{bmatrix}-2\\0\end{bmatrix}\right) \geq 1$$

$$(1)\left(\boldsymbol{w} \cdot \begin{bmatrix}0\\2\end{bmatrix}\right) \geq 1$$

$$(1)\left(\boldsymbol{w} \cdot \begin{bmatrix}1\\1\end{bmatrix}\right) \geq 1$$

$$(1)\left(\boldsymbol{w} \cdot \begin{bmatrix}3\\0\end{bmatrix}\right) \geq 1$$

"Throws to an off-the-shelf solver"

$$\therefore \boldsymbol{w} = \begin{bmatrix}-0.5\\-0.5\end{bmatrix}$$

Recap: Support Vector Machines

Dual Formulation: First,

$$\text{maximize}_{\boldsymbol{\alpha}} \sum_i \alpha^{(i)} - \frac{1}{2} \sum_i \sum_j \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \left(\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)} \right)$$

$$\text{subject to } \sum_i \alpha^{(i)} y^{(i)} = 0 \text{ and } \alpha^{(i)} \geq 0$$

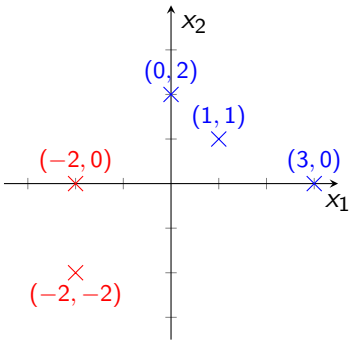
Then compute

$$\boldsymbol{w} = \sum_i \alpha^{(i)} y^{(i)} \boldsymbol{x}^{(i)}$$

- **Convex** optimization problem with linear constraints.
- $\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}$ can be precomputed – efficient when there are many features.
- For the **optimal solution**, $\alpha^{(i)}$ for non-support vectors can always be 0.

Recap: Support Vector Machines

$$y^{(1)}y^{(5)}(x^{(1)} \cdot x^{(5)}) = (-1)(1)\left(\begin{bmatrix}-2\\-2\end{bmatrix} \cdot \begin{bmatrix}3\\0\end{bmatrix}\right)$$



SVM Example

$$\begin{aligned}\text{maximize}_{\boldsymbol{\alpha}}\ & \alpha^{(1)} + \alpha^{(2)} + \alpha^{(3)} + \alpha^{(4)} + \alpha^{(5)} \\ & - \frac{1}{2}(8\alpha^{(1)2} + 4\alpha^{(1)}\alpha^{(2)} + +4\alpha^{(1)}\alpha^{(3)} + 4\alpha^{(1)}\alpha^{(4)} + 6\alpha^{(1)}\alpha^{(5)} \\ & \quad + 4\alpha^{(2)}\alpha^{(1)} + 4\alpha^{(2)2} + 4\alpha^{(2)}\alpha^{(3)} + 2\alpha^{(2)}\alpha^{(5)} \\ & \quad + 4\alpha^{(3)}\alpha^{(1)} + 4\alpha^{(3)2} + 2\alpha^{(3)}\alpha^{(4)} \\ & \quad + 4\alpha^{(4)}\alpha^{(1)} + 2\alpha^{(4)}\alpha^{(2)} + 2\alpha^{(4)}\alpha^{(3)} + 2\alpha^{(4)2} + 3\alpha^{(4)}\alpha^{(5)} \\ & \quad + 6\alpha^{(5)}\alpha^{(1)} + 6\alpha^{(5)}\alpha^{(2)} + 3\alpha^{(5)}\alpha^{(4)} + 9\alpha^{(5)2})\end{aligned}$$

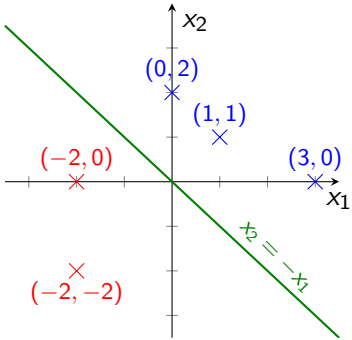
$$\text{subject to } -\alpha^{(1)} - \alpha^{(2)} + \alpha^{(3)} + \alpha^{(4)} + \alpha^{(5)} = 0 \text{ and } \alpha^{(i)} \geq 0$$

"Throws to an off-the-shelf solver"

$$\alpha^{(1)} = 0, \alpha^{(2)} = 0.25, \alpha^{(3)} = 0.25, \alpha^{(4)} = 0, \alpha^{(5)} = 0$$

$$\therefore \boldsymbol{w} = -0\begin{bmatrix}-2\\-2\end{bmatrix} - 0.25\begin{bmatrix}-2\\0\end{bmatrix} + 0.25\begin{bmatrix}0\\2\end{bmatrix} + 0\begin{bmatrix}1\\1\end{bmatrix} + 0\begin{bmatrix}3\\0\end{bmatrix} = \begin{bmatrix}0.5\\0.5\end{bmatrix}$$

Q2. SVM



SVM Example

(a) If $\boldsymbol{w} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$, which of the points are found on the SVM margins?

We have $\boldsymbol{z} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

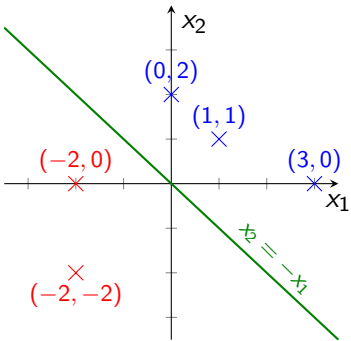
Distance from each point to decision boundary

$$= \frac{|(\boldsymbol{x} - \boldsymbol{z}) \cdot \boldsymbol{w}|}{\|\boldsymbol{w}\|} = |\frac{\boldsymbol{X}\boldsymbol{w}|}{\|\boldsymbol{w}\|}$$

$$= \frac{1}{\sqrt{0.5^2 + 0.5^2}} \left| \begin{bmatrix} -2 & -2 \\ -2 & 0 \\ 0 & 2 \\ 1 & 1 \\ 3 & 0 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \right| = \sqrt{2} \begin{bmatrix} 2 \\ 1 \\ 1 \\ 1 \\ 1.5 \end{bmatrix}$$

$\therefore$ The support vectors are $(-2, 0)$, $(0, 2)$ and $(1, 1)$.

Q2. SVM



SVM Example

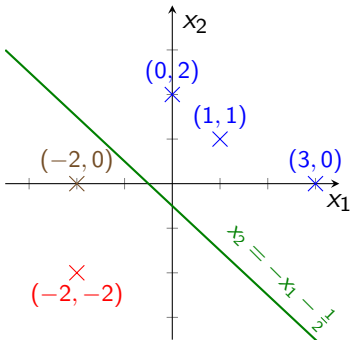
(b) Suppose we introduce another point, $\boldsymbol{x}^{(6)}$, with features $\begin{bmatrix} -5 & 1 \end{bmatrix}^\top$ and label -1, then retrain the SVM. Will the learned model change?

Distance to decision boundary

$$= \frac{|\boldsymbol{x}^{(6)} \cdot \boldsymbol{w}|}{\|\boldsymbol{w}\|} = \frac{\left| \begin{bmatrix} -5 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \right|}{\sqrt{0.5^2 + 0.5^2}} = \sqrt{2} \cdot 2.$$

Since $x^{(6)}$ is not a support vector, the learned model will not change.

Q2. SVM

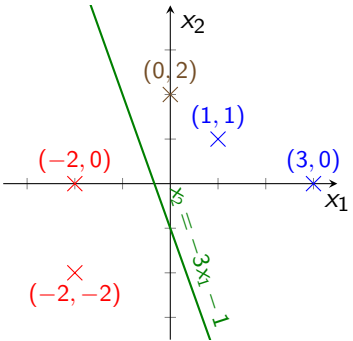


SVM Example

(c) Let's remove data point $\boldsymbol{x}^{(2)}$ and retrain. What will happen to the model?

- $\boldsymbol{x}^{(2)}$ is a support vector, so the decision boundary is probably affected.
- $\boldsymbol{x}^{(1)} = (-2, -2)$ becomes the new support vector.
- The decision boundary should be shifted towards $(-2, -2)$.

Q2. SVM



SVM Example

(d) How would the results differ when we remove $\boldsymbol{x}^{(3)}$ instead of $\boldsymbol{x}^{(2)}$ and retrain the model?

- $\boldsymbol{x}^{(3)}$ is a support vector, so the decision boundary is probably affected.
- The support vectors would be $(-2, 0)$ and $(1, 1)$.
- The decision boundary would be the perpendicular bisector of the line segment connecting them.

Q3. Primal Formulation

(a) Write down the expression for the smallest distance of all points to a hyperplane defined by some $\boldsymbol{w}$.

- Take $\boldsymbol{z} = \mathbf{0}$.
- Distance of all points to a hyperplane $= \dfrac{|(\boldsymbol{x} - \boldsymbol{z}) \cdot \boldsymbol{w}}{\|\boldsymbol{w}\|} = \dfrac{|\boldsymbol{x} \cdot \boldsymbol{w}|}{\|\boldsymbol{w}\|}$.
- $\therefore$ The expression is $\min\limits_{i} \dfrac{|\boldsymbol{x} \cdot \boldsymbol{w}|}{\|\boldsymbol{w}\|}$.

(b) Write down the expression for maximising the smallest distance of all points to a hyperplane defined by some $\boldsymbol{w}$.

- The expression is $\max\limits_{\boldsymbol{w}} \min\limits_{i} \dfrac{|\boldsymbol{x} \cdot \boldsymbol{w}|}{\|\boldsymbol{w}\|}$.

Q3. Primal Formulation

(c) Does this expression satisfy the correct classification constraints of the SVM, i.e., do the points lie on the correct side of the hyperplane?

- No!

Q3. Primal Formulation

Our current optimization problem:

$$\begin{aligned} &\text{maximize}_{\boldsymbol{w}} \ \min_i \frac{|\boldsymbol{x} \cdot \boldsymbol{w}|}{\|\boldsymbol{w}\|} \\ &\text{subject to } y^{(i)}(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)}) > 0 \end{aligned}$$

- Observation 1: $\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} = \pm c$ for support vectors.
- Observation 2: We can scale $\boldsymbol{w}$ freely.
- $\Rightarrow$ Fix $\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} = \pm 1$ by scaling $\boldsymbol{w}$.

Q3. Primal Formulation

New optimization problem:

$$\text{maximize}_{\boldsymbol{w}} \ \min_i \frac{|\boldsymbol{x} \cdot \boldsymbol{w}|}{\|\boldsymbol{w}\|} = \frac{1}{\|\boldsymbol{w}\|}$$
$$\text{subject to } y^{(i)}(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)}) \geq 1$$

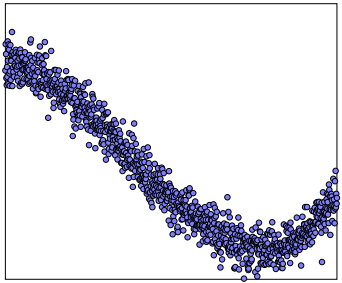
- Maximize $\frac{1}{\|\boldsymbol{w}\|} \Rightarrow$ Minimize $\|\boldsymbol{w}\| \Rightarrow$ Minimize $\frac{1}{2}\|\boldsymbol{w}\|^2$

Primal formulation:

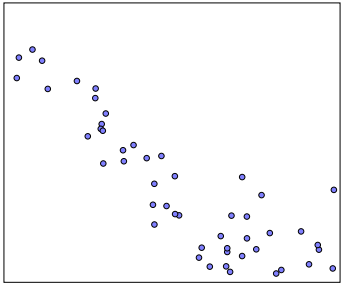
$$\text{minimize}_{\boldsymbol{w}} \ \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{subject to } y^{(i)}(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)}) \geq 1$$

Bias and Variance

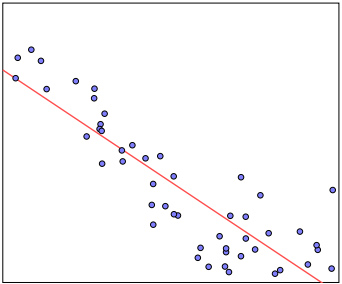
Population



Sample



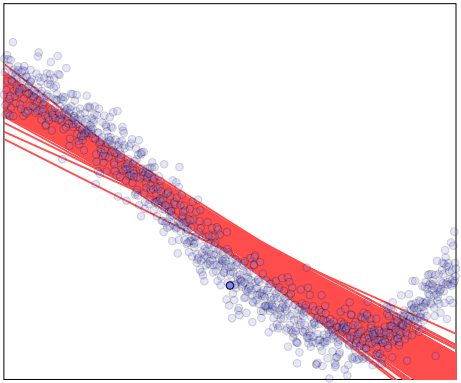
Trained Model

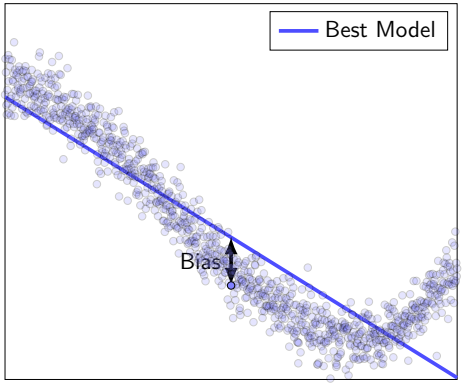


Hypothesis class:
$h_w(x) = w_0 x + b$

Bias and Variance

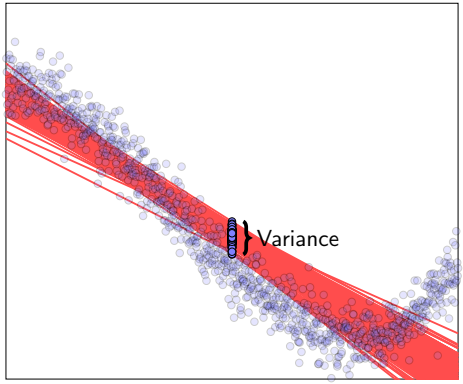
Bias measures how much the **expected** *predicted value* differs from the *actual value*. Intuition: The error of the **best model** when you are given **infinite amount of training data**.





Bias and Variance

Variance measures the consistency of predictions due to sampling.



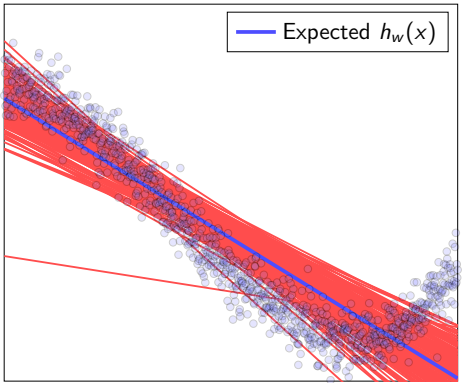
Bias and Variance

Poll:

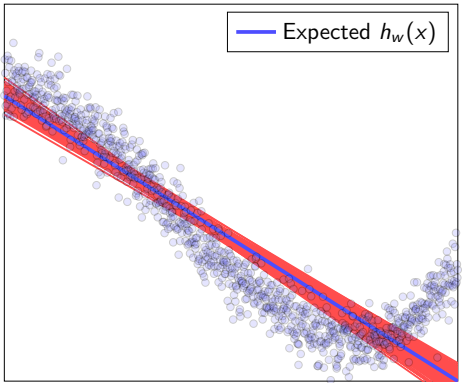
1. If we increase the amount of training samples:
   The bias _____remains unchanged_____ and the variance _____decreases_____.
2. If we increase the degree of the polynomial in our hypothesis class:
   The bias _____decreases_____ and the variance _____increases_____.

Bias and Variance

Less samples (Higher variance)

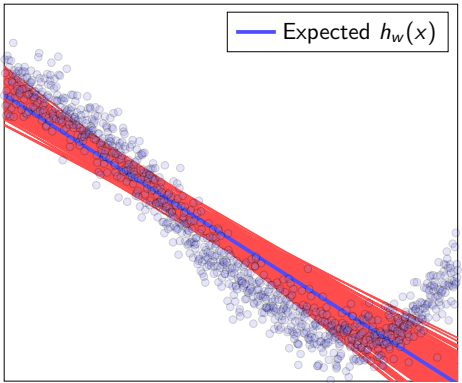


More samples (Lower variance)

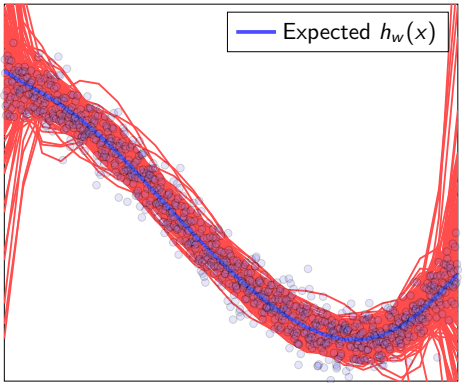


Bias and Variance

Lower degree (High bias; Low variance)

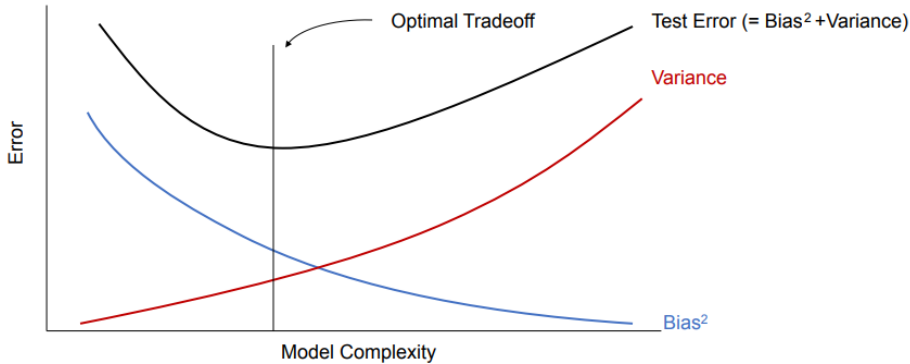


Higher degree (Low bias; High variance)



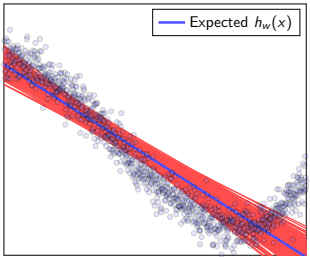
Bias-Variance Tradeoff

$$\mathrm{Err}(x) = \left(\underbrace{y - \mathbb{E}[h_w(x)]}_{\text{bias}} \right)^2 + \underbrace{\mathbb{E}\left[(\mathbb{E}[h_w(x)] - h_w(x))^2 \right]}_{\text{variance}}$$

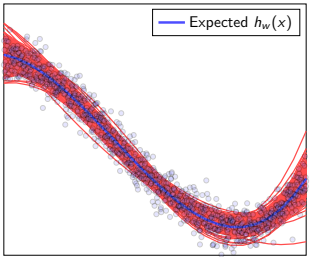


Bias-Variance Tradeoff

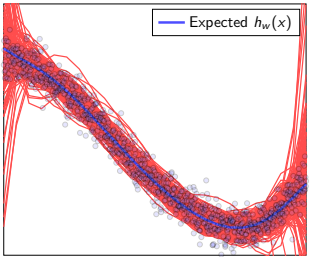
Underfitting



Just Right



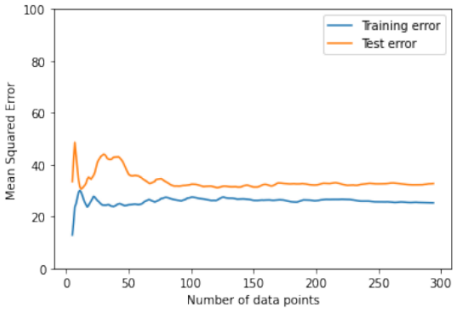
Overfitting

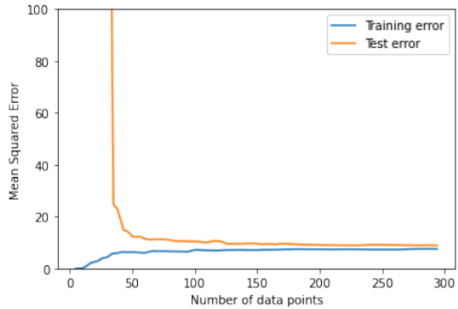


Q4. Bias and Variance

The model hypotheses are as below:

1. $H_w(x) = w_0 + w_1 x$
2. $H_w(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_{10} x^{10}$





Q4. Bias and Variance

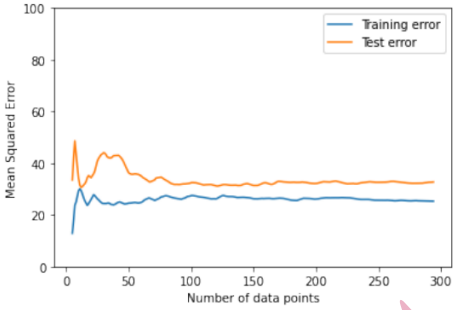
Main Idea:

- Both bias and variance contribute to the mean squared error.
- Bias does not decrease with the number of training samples. Bias causes error in **both** the training set and the test set.
- Variance decreases with the number of training samples. Variance causes a difference between the error in the training set and that in the test set.

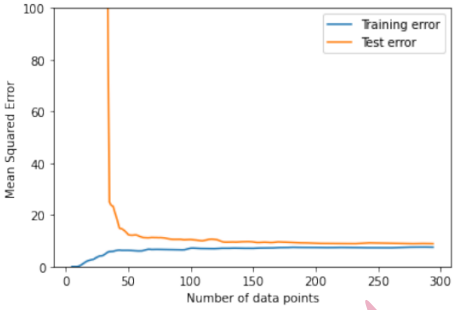
Q4. Bias and Variance

The model hypotheses are as below:

1. $H_w(x) = w_0 + w_1 x$ High bias
2. $H_w(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_{10} x^{10}$ High variance



High bias



High variance

Recap: Kernel Trick

Using the dual formulation:

$$\text{maximize}_{\boldsymbol{\alpha}} \sum_i \alpha^{(i)} - \frac{1}{2} \sum_i \sum_j \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \left(\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)} \right)$$

$$\text{subject to } \sum_i \alpha^{(i)} y^{(i)} = 0 \text{ and } \alpha^{(i)} \geq 0$$

We transform the features in $\boldsymbol{x}^{(i)}$ to $\phi(\boldsymbol{x}^{(i)})$. How does that impact the training?

- Not much, we have to change $\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}$ to $\phi(\boldsymbol{x}^{(i)}) \cdot \phi(\boldsymbol{x}^{(j)})$.
- We define a **kernel function** $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \phi(\boldsymbol{x}^{(i)}) \cdot \phi(\boldsymbol{x}^{(j)})$ to compute this efficiently.

Important Property of a Kernel Function

$$K(\boldsymbol{u}, \boldsymbol{v}) = \phi(\boldsymbol{u}) \cdot \phi(\boldsymbol{v}) \text{ for some function } \phi$$

Recap: Kernel Trick

Example: $K(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u} \cdot \boldsymbol{v})^2$, where $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$.

$$
\begin{aligned}
K(\boldsymbol{u}, \boldsymbol{v}) &= (\boldsymbol{u} \cdot \boldsymbol{v})^2 \\
&= \left(\sum_i u_i v_i \right)^2 \\
&= u_1^2 v_1^2 + u_2^2 v_2^2 + u_3^2 v_3^2 + 2u_1 u_2 v_1 v_2 + 2u_1 u_3 v_1 v_3 + 2u_2 u_3 v_2 v_3 \\
&= \begin{bmatrix} u_1^2 \\ u_2^2 \\ u_3^2 \\ \sqrt{2} u_1 u_2 \\ \sqrt{2} u_1 u_3 \\ \sqrt{2} u_2 u_3 \end{bmatrix} \cdot \begin{bmatrix} v_1^2 \\ v_2^2 \\ v_3^2 \\ \sqrt{2} v_1 v_2 \\ \sqrt{2} v_1 v_3 \\ \sqrt{2} v_2 v_3 \end{bmatrix} \\
&= \phi(\boldsymbol{u}) \cdot \phi(\boldsymbol{v})
\end{aligned}
$$

Recap: Kernel Trick

Dual formulation **with kernel trick**:

$$\text{maximize} \sum_i \alpha^{(i)} - \frac{1}{2} \sum_i \sum_j \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \times K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$$

$$\text{subject to} \sum_i \alpha^{(i)} y^{(i)} = 0 \text{ and } \alpha^{(i)} \geq 0$$

Does not slow down training, but "implicitly" considers all transformed features!

Bonus. Gaussian Kernel

Extra Slide

Prove that the Gaussian Kernel, $K(\boldsymbol{u}, \boldsymbol{v}) = e^{-\frac{\|\boldsymbol{u}-\boldsymbol{v}\|^2}{2\sigma^2}}$, has infinite dimensional features.
(We assume $\sigma^2 = 1$ for simplicity.)

$$
\begin{aligned}
K(\boldsymbol{u}, \boldsymbol{v}) &= e^{-\frac{\|\boldsymbol{u}-\boldsymbol{v}\|^2}{2}} \\
&= e^{-\frac{\boldsymbol{u}\cdot\boldsymbol{u}-2\boldsymbol{u}\cdot\boldsymbol{v}+\boldsymbol{v}\cdot\boldsymbol{v}}{2}} \\
&= e^{-\frac{\boldsymbol{u}\cdot\boldsymbol{u}}{2}} e^{\boldsymbol{u}\cdot\boldsymbol{v}} e^{-\frac{\boldsymbol{v}\cdot\boldsymbol{v}}{2}} \\
&= e^{-\frac{\boldsymbol{u}\cdot\boldsymbol{u}}{2}} \left(1 + \boldsymbol{u}\cdot\boldsymbol{v} + \frac{1}{2!}(\boldsymbol{u}\cdot\boldsymbol{v})^2 + \frac{1}{3!}(\boldsymbol{u}\cdot\boldsymbol{v})^3 + \frac{1}{4!}(\boldsymbol{u}\cdot\boldsymbol{v})^4 + \cdots\right) e^{-\frac{\boldsymbol{v}\cdot\boldsymbol{v}}{2}} \\
&= e^{-\frac{\boldsymbol{u}\cdot\boldsymbol{u}}{2}} \left(\phi_0(\boldsymbol{u}) \cdot \phi_0(\boldsymbol{v})\right) e^{-\frac{\boldsymbol{v}\cdot\boldsymbol{v}}{2}} \\
&= \left(e^{-\frac{\boldsymbol{u}\cdot\boldsymbol{u}}{2}} \phi_0(\boldsymbol{u})\right) \cdot \left(e^{-\frac{\boldsymbol{v}\cdot\boldsymbol{v}}{2}} \phi_0(\boldsymbol{v})\right)
\end{aligned}
$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

Sum of polynomial kernels (in lecture)
$\Rightarrow$ Also a kernel (by bonus Q1)

scalars (multiplied to the transformed features)