Problem 1

1(a)

Negative log likelihood or entropy is given by,

$$\mathcal{L}(w) = -log\left(\prod_{i=1}^{n} P(Y = y_i|X = x_i)\right)$$

$$= -\sum_{i=1}^{n} logP(y_i|x_i) = -\sum_{i=1}^{n} log\left[\sigma(w^T x_i)^{y_i}[1-\sigma(w^T x_i)^{1-y_i}]\right]$$

$$\Rightarrow \mathcal{L}(w) = -\sum_{i=1}^{n}\{y_i \log \sigma(w^T x_i) + (1-y_i)\log[1-\sigma(w^T x_i)]\} \quad (1)$$

1(b)

A function $f(x)$ is convex if its Hessian is a positive definite for all x. From eqn (1):

$$\frac{\partial \mathcal{L}(w)}{\partial w} = \sum_{i=1}^{n}\{\sigma(w^T x_i) - y_i\}x_i$$

$$\frac{\partial^2 \mathcal{L}(w)}{\partial w \partial w^T} = \sum_{i=1}^{n} x_i x_i^T \sigma(w^T x_i)(1-\sigma(w^T x_i))$$

For any v,

$$v^T X^T X v = ||X^T v||_2^2 \geqslant 0$$

and

$$\sigma(w^T x_i)(1-\sigma(w^T x_i)) \geqslant 0$$

Hence, Loss function eqn (1) is a convex function.

1(c)

If the training samples are linearly separable, it means the probability likelihood of a outcome for any given input tends to 1 or 0. If probability tends to 1:

$$P(Y = c|x) = \sigma(w^T x) \to 1$$

$$\Leftrightarrow \frac{1}{1+e^{-w^T x}} \to 1$$

$$\Leftrightarrow e^{-w^T x} \to 0$$

$$\Leftrightarrow w^T x \to \infty$$

As x is given,

$$\Leftrightarrow w^T \to \infty$$

If probability tends to 0:

$$\begin{aligned} P(Y = c|x) = \sigma(w^T x) &\to 0 \\ \Leftrightarrow \frac{1}{1 + e^{-w^T x}} &\to 0 \\ \Leftrightarrow e^{-w^T x} &\to \infty \\ \Leftrightarrow w^T x &\to -\infty \end{aligned}$$

As x is given,

$$\Leftrightarrow w^T \to -\infty$$

Hence, we prove that when Logistic Regression achieves monotonous likelihood (Linearly separable), the weights tend to Infinity.

1(d)

Given,

$$\mathcal{L}(w) = -log\left(\prod_{i=1}^{n} P(Y = y_i|X = x_i)\right) + \lambda||w||_2^2$$

$$= -\sum_{i=1}^{n}\{y_i \log \sigma(w^T x_i) + (1 - y_i) \log [1 - \sigma(w^T x_i)]\} + \lambda||w||_2^2$$

We find Gradient wrt w_I as follows:

$$\frac{\partial \mathcal{L}(w)}{\partial w} = -\frac{\partial}{\partial w}\sum_{i=1}^{n}\{y_i \log \sigma(w^T x_i) + (1 - y_i) \log [1 - \sigma(w^T x_i)]\} + \frac{\partial}{\partial w}\lambda||w||_2^2$$

$$= -\sum_{i=1}^{n} y_i x_i (1 - \sigma(w^T x_i)) + \sum_{i=1}^{n}(1 - y_i)x_i \sigma(w^T x_i) + 2\lambda w$$

$$= \sum_{i=1}^{n}[-y_i x_i + y_i x_i \sigma(w^T x_i) + x_i \sigma(w^T x_i) - y_i x_i \sigma(w^T x_i)] + 2\lambda w$$

$$= \sum_{i=1}^{n}\{\sigma(w^T x_i) - y_i\}x_i + 2\lambda w$$

$$\frac{\partial \mathcal{L}(w)}{\partial w} = \sum_{i=1}^{n}\{\sigma(w^T x_i) - y_i\}x_i + 2\lambda w \tag{2}$$

1(e)

A function $f(x)$ has unique solution if it is convex function. It is convex if its Hessian is a positive definite for all x. From eqn (2):

$$\frac{\partial \mathcal{L}(w)}{\partial w} = \sum_{i=1}^{n} \{\sigma(w^T x_i) - y_i\} x_i + 2\lambda w$$

$$\frac{\partial^2 \mathcal{L}(w)}{\partial w \partial w^T} = \sum_{i=1}^{n} x_i x_i^T \sigma(w^T x_i)(1 - \sigma(w^T x_i))$$

For any v,

$$v^T X^T X v = ||X^T v||_2^2 \geqslant 0$$

and

$$\sigma(w^T x_i)(1 - \sigma(w^T x_i)) \geqslant 0$$

Hence, Loss function eqn (1) is a convex function which implies it has a unique solution.

Problem 2

2(a)

We know a function $f(x)$ is Concave if :

$$f(\lambda w_1 + (1-\lambda) w_2) \leqslant \lambda f(w_1) + (1-\lambda) f(w_2) \text{ where } 0 < \lambda < 1 \tag{3}$$

and

$$f(w) = ||w||_0 = \sum_i I_{wi \neq 0} \tag{4}$$

Lets assume the following:

$$w_1 = [0, 5, 6, 7, 8]$$
$$w_2 = [0, 1, 0]$$

We get:

$$LHS = f(\lambda w_1 + (1-\lambda) w_2)$$
$$= f([0, 5\lambda, 6\lambda, 7\lambda, 8\lambda, 0, 1(1-\lambda), 0])$$
$$= 5$$

$$\begin{aligned} RHS &= \lambda f(w_1) + (1-\lambda) f(w_2) \\ &= \lambda f([0,5,6,7,8]) + (1-\lambda) f([0,1,0]) \\ &= 4\lambda + (1-\lambda)1 \\ &= 3\lambda + 1 \end{aligned}$$

We see that, $LHS > RHS$, i.e

$$\begin{aligned} 5 &> 3\lambda + 1 \\ \Rightarrow f(\lambda w_1 + (1-\lambda)w_2) &> \lambda f(w_1) + (1-\lambda) f(w_2) \text{ where } 0 < \lambda < 1 \end{aligned}$$

Hence, we can say, $||w||_0$ is Non-convex or Concave.

2(b)

We know a function $f(x)$ is Convex if :

$$f(\lambda w_1 + (1-\lambda)w_2) \leqslant \lambda f(w_1) + (1-\lambda) f(w_2) \text{ where } 0 < \lambda < 1 \tag{5}$$

and

$$f(w) = ||w||_1 = \sum_i |w_i| \tag{6}$$

Also for every norm we have below properties:
Triangle Inequality

$$f(x+y) \leqslant f(x) + f(y) \tag{7}$$

Homogeneity

$$f(tx) = |t| f(x) \tag{8}$$

Using above properties, we get:

$$\begin{aligned} & f(\lambda w_1 + (1-\lambda)w_2) \\ &= ||\lambda w_1 + (1-\lambda)w_2|| \\ &\leqslant ||\lambda w_1|| + ||(1-\lambda)w_2|| \ (\because \text{ of Triangle Inequality eqn (7)}) \\ &= \lambda ||w_1|| + (1-\lambda)||w_2|| \ (\because \text{ of Homogeneity eqn (8)}) \end{aligned}$$

Hence we see that,

$$f(\lambda w_1 + (1-\lambda)w_2) \leqslant \lambda ||w_1|| + (1-\lambda)||w_2||$$

Which implies $||w||$ is Convex.

2(c)

Given,

$$RSS(w) = min_w \sum_n (y_n - w^T x_n)^2 + \lambda ||w||_1$$

Let $t = [t_1, t_2, \ldots, t_D]$ where $\forall\ 1 \leqslant i \leqslant D$, $t_i \geqslant w_i$.

$$\sum_i t_i \geqslant ||w||_1$$

Then, w can be replaced with t.
It can be expressed in Matrix form as follows:

$$\begin{aligned} RSS(w) &= min_w (Y^T - X^T w)^2 + \lambda ||w||_1 \\ &= min_t [(Y^T - X^T t)^T (Y^T - X^T t) + \lambda t I] \\ &= min_t [Y Y^T + t^T X X^T t - 2 Y^T X t + \lambda t I] \end{aligned}$$

If we assume:

$$\begin{aligned} 2XX^T &= Q, \\ 2Y^T X + \lambda I &= c^T, \\ t &= u \\ t &\geqslant w \end{aligned}$$

By substituting values of Q, u and c^T in the above equation, we get:

$$\mathbf{RSS(u) = min_u \frac{1}{2} u^T Q u + c^T u + const} \tag{9}$$

Problem 3

Given,

$$RSS(w) = min_w \sum_n (y_n - w^T x_n)^2 + \lambda ||w||_2^2 \tag{10}$$

3(a)

Lets find the gradient of eqn (10),

$$\begin{aligned} \frac{\partial RSS(w)}{\partial w} &= -2 \sum_n (y_n - w x_n) x_n + 2\lambda w = 0 \\ &\Rightarrow w^* = (\sum_n x_i x_i^T + \lambda I_D)^{-1} X^T y \\ &\Rightarrow \mathbf{w^* = (X^T X + \lambda I_D)^{-1} X^T y} \end{aligned}$$

3(b)

If x maps to $\phi(x)$, then we get the Regularized least square as :

$$J(w) = \sum_n (y_n - w^T\phi(x))^2 + \lambda||w||_2^2$$
$$\frac{\partial J(w)}{\partial w} = 2\sum_n (y_n - w\phi(x))\phi(x) + 2\lambda w = 0$$
$$\Rightarrow w^* = (\sum_n \phi(x)\phi(x)^T + \lambda I_N)^{-1} \sum_n \phi(x)^T y$$
$$\Rightarrow w^* = (\Phi^T\Phi + \lambda I_N)^{-1}\Phi^T y$$

By using Matrix inversion Lemma, we get,

$$\Rightarrow \mathbf{w^*} = \mathbf{\Phi^T(\Phi\Phi^T + \lambda I_N)^{-1}y} \tag{11}$$

3(c)

From 3(b) we have the following:

$$w^* = (\Phi\Phi^T + \lambda I_N)^{-1}\Phi^T y$$

Let $K = \Phi\Phi^T$, where K is the gram matrix. Then,

$$w^* = (K + \lambda I_N)^{-1}\Phi^T y$$
$$= y(K + \lambda I_N)^{-1}\phi(x)^T$$

$$\hat{y} = {w^*}^T\phi(x)$$
$$= y^T(K + \lambda I_N)^{-1}\phi(x)^T\phi(x)$$
$$\Rightarrow \mathbf{\hat{y} = y^T(K + \lambda I_N)^{-1}k(x)}$$

Where $k(x) = \phi(x)^T\phi(x)$

3(d)

Lets consider N samples with D features each. So, For Linear Ridge Regression, we perform X^TX which gives a complexity of $O(D^3 + N \times D^2)$ for Training.

In Kernel Ridge Regression, we perform Matrix multiplication and Inversion. Hence, it gives a complexity of $O(N^3)$ for Training.

Problem 4

We know that: For any function $F(.)$, if k is a valid Kernel, IF we have:

$$\int_{x,x'} f(x)f(x')k(x,x')dxdx' \geqslant 0 \tag{12}$$

4(a)

Lets check if k_3 is a Kernel using property (12)

$$\begin{aligned}\int_{x,x'} f(x)f(x')k_3(x,x')dxdx' &= \int_{x,x'} f(x)f(x')[a_1k_1(x,x') + a_2k_2(x,x')]dxdx' \\ &a_1\int_{x,x'} f(x)f(x')k_1(x,x')dxdx' + a_2\int_{x,x'} f(x)f(x')k_2(x,x')dxdx' \\ &\geqslant 0 + 0 \geqslant 0\end{aligned}$$

($\because$ k_1 and k_2 are valid Kernels and $a_1, a_2 \geqslant 0$)
Hence, k_3 is a Kernel where,

$$\mathbf{k_3}(\mathbf{x},\mathbf{x}') = \mathbf{a_1}\mathbf{k_1}(\mathbf{x},\mathbf{x}') + \mathbf{a_2}\mathbf{k_2}(\mathbf{x},\mathbf{x}') \tag{13}$$

4(b)

k_4 can be written as $N \times N$ matrix for points $f(x_1) \ldots f(x_n)$ Which is symmetric.
Hence, k_4 matrix can be written in the form of $k_4 = FF^T$. For any Vector V, we have,

$$V^Tk_4V = V^TFF^TV = (F^TV)^T(F^TV) = ||F^TV||_2^2 \geqslant 0$$

Hence, k_4 is a Kernel where,

$$\mathbf{k_4}(\mathbf{x},\mathbf{x}') = \mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x}') \tag{14}$$

4(c)

Given, Polynomial function:

$$g(k_1(x,x')) = 1 + c_1k_1(x,x') + c_2k_1(x,x')^2 + \ldots + c_nk_1(x,x')^n \geqslant 0$$

as $k_1(x,x') \geqslant 0$.
Hence, k_5 is a Kernel where,

$$\mathbf{k_5}(\mathbf{x},\mathbf{x}') = \mathbf{g}(\mathbf{k_1}(\mathbf{x},\mathbf{x}')) \tag{15}$$

4(d)

$$\begin{aligned} k_6(x,x') &= k_1(x,x')k_2(x,x') \\ &=< \Phi_1(x), \Phi_1(x') >< \Phi_2(x), \Phi_2(x') > \\ &= \left(\sum_i \Phi_1(x_i)\Phi_1(x_i')\right)\left(\sum_j \Phi_2(x_j)\Phi_2(x_j')\right) \\ &= \sum_{i,j} \Phi_1(x_i)\Phi_2(x_j)\Phi_1(x_i')\Phi_2(x_j') \\ &= \sum_{i,j} \Phi(x_{ij})\Phi(x_{ij}') \\ &=< \Phi(x), \Phi(x') > \end{aligned}$$

Hence, k_6 is a Kernel where,

$$\mathbf{k_6}(\mathbf{x},\mathbf{x}') = \mathbf{k_1}(\mathbf{x},\mathbf{x}')\mathbf{k_2}(\mathbf{x},\mathbf{x}') \tag{16}$$

4(e)

We can express Exponential in the form of polynomial

$$\begin{aligned} k_7(x,x') &= e^{k_1(x,x')} \\ &= \sum_{k=0}^{\infty} \frac{k_1(x,x')^k}{k!} (\text{ By Taylor Expansion }) \end{aligned}$$

From Proof (15), we say the above function k_7 is a Kernel. Hence, k_7 is a Kernel where,

$$\mathbf{k_7}(\mathbf{x},\mathbf{x}') = \mathbf{e}^{\mathbf{k_1}(\mathbf{x},\mathbf{x}')} \tag{17}$$

Problem 5

5(a)

Function	**Bias2**	**Variance**
$g_1(x)$	0.4739	0
$g_2(x)$	0.32972	0.038464
$g_3(x)$	0.28201	0.11317
$g_4(x)$	0.034332	0.38877
$g_5(x)$	0.044424	0.39886
$g_6(x)$	0.05529	0.40973

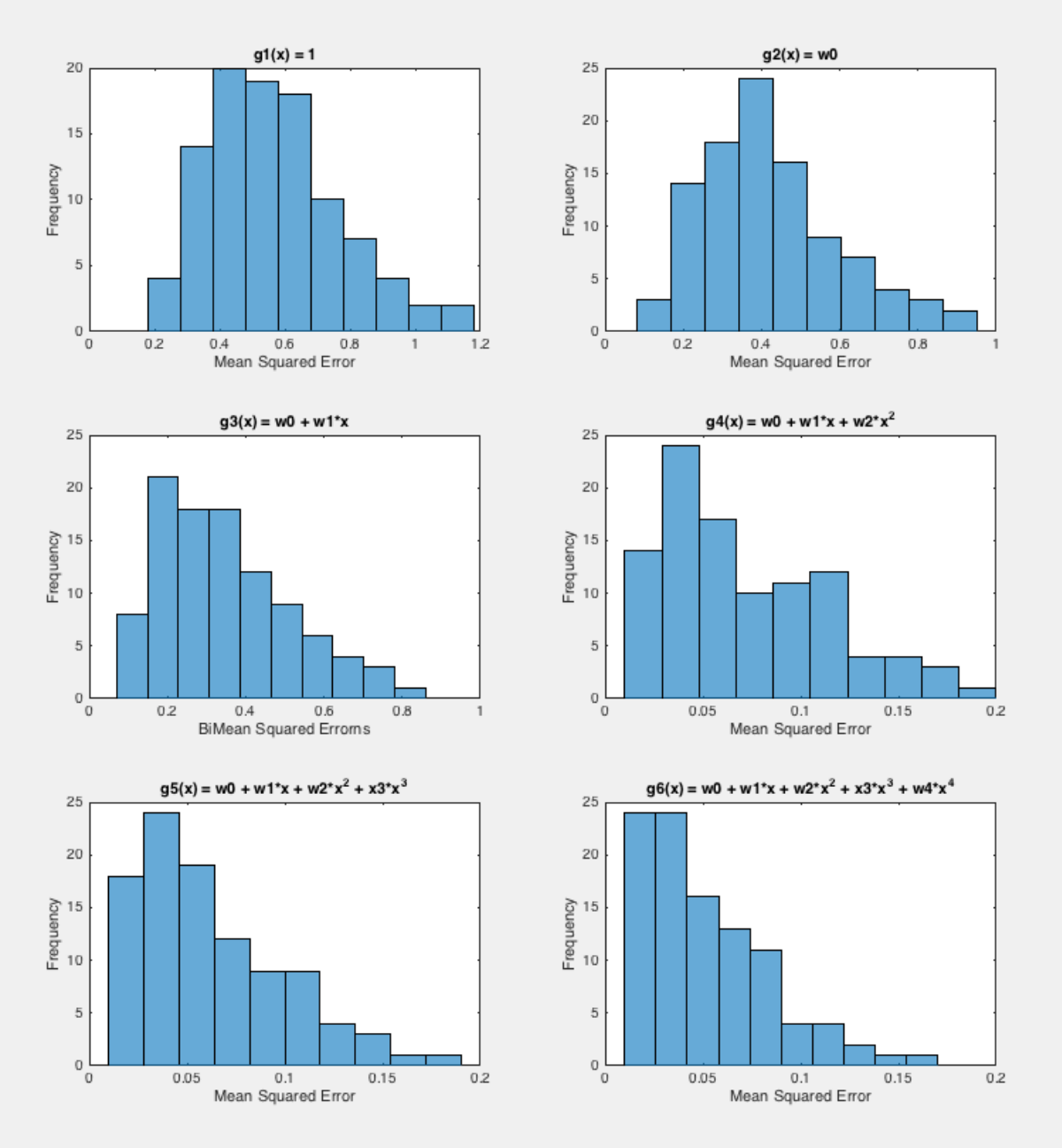

Figure 1: 5.(a) Sample size 10

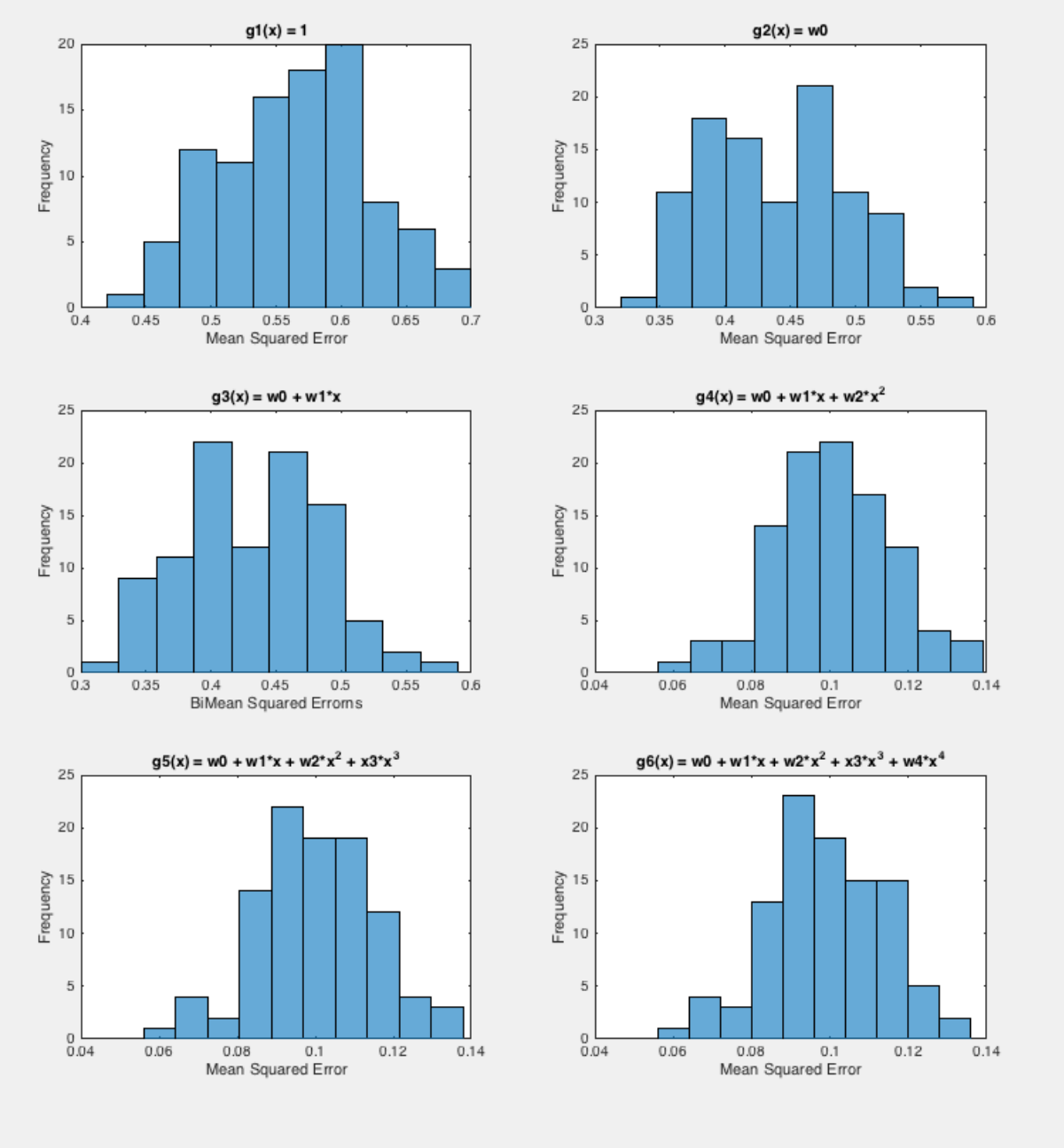

Figure 2: 5.(b) Sample size 100

5(b)

Function	Bias2	Variance
$g_1(x)$	0.46179	0
$g_2(x)$	0.34122	0.0044352
$g_3(x)$	0.33625	0.012662
$g_4(x)$	0.003084	0.34461
$g_5(x)$	0.0037683	0.34529
$g_6(x)$	0.0048795	0.3464

5(c)

Model Complexity

Model Complexity has a strong impact on both Bias and Variance. We find that Bias is very high and Variance is very low for Simple models. As Model Complexity increases, Bias decreases but Variance starts increasing as model Complexity increases. We can conclude that we always have to trade-off one for another. We cannot have the best Bias and Variance at the same time.

Sample Size

We have the readings for sample size of 10 and 100. There is hardly any impact of Sample Size on Bias and Variance. We can conclude that Bias and Variance doesnt depend on Sample Size at all.

5(d)

Lambda	Bias2	Variance
0.01	0.0051536	0.354
0.1	0.17008	0.35426
1	15.7846	0.50362
10	1113.1198	11.5829

λ controls the size of parameters and the amount of regularization. We see that as $\lambda \downarrow 0$, we obtain the Least squares solution, i.e we have best possible Bias and Variance combination. As λ increases, both Bias and Variance seem to increase. Bias has polynomial increase rate with λ, where as Variance seem to increase exponentially.

Problem 6

Linear Ridge Regression

Iteration	Lambda
1	0.1
2	0.01
3	0.01
4	0.1
5	0.1
6	0.1
7	0
8	0
9	0.1
10	0.1

Average Test Error: 0.0167

Kernel Ridge Regression

(a)

Iteration	Lambda
1	0.1
2	0.1
3	0.01
4	0.1
5	0.0001
6	0.0001
7	0.0001
8	0.1
9	0.01
10	0.1

Average Test Error: 0.0165

(b)

Used 3 iterations and 50% partition.

Iteration	Lambda	a	b
1	10	1	3
2	10	1	3
3	10	1	3

Average Test Error: 0.0117

(c)

Used 3 iterations and 50% partition.

Iteration	**Lambda**	**Sigma2**
1	0.01	8
2	0.01	8
3	0.01	8

Average Test Error: 0.0107 Though Selection of λ is different for Kernel Ridge Linear Regression and Linear Ridge Regression, the average test error is almost same for both the models.

Among 3 kernels, RBF kernel performs best with **Average Test Error at 0.0107**