1. (a) LFD 4.8 $\quad \Gamma = I, \ \lambda > 0$

$$\underline{\vec{w}(t+1) \leftarrow \vec{w}(t) - \eta \nabla E_{avg}(\vec{w}(t))}$$

$$E_{avg}(\vec{w}) = E_{in}(\vec{w}) + \lambda \vec{w}^T \underbrace{\Gamma^T \Gamma w}$$ identity

$$\nabla E_{avg}(\vec{w}) = \nabla E_{in}(\vec{w}) + 2\lambda \vec{w}$$

$$\vec{w}(t+1) \leftarrow \vec{w}(t) - \eta\left(\nabla E_{in}(\vec{w}) + 2\lambda \vec{w}\right)$$

$$\leftarrow \vec{w}(t) - \eta \nabla E_{in}(\vec{w}) - \eta 2\lambda \vec{w}$$

$$\leftarrow \vec{w}(t) - \eta 2\lambda \vec{w}(t) - \eta \nabla E_{in}(\vec{w})$$

$$\vec{w}(t+1) \leftarrow \boxed{\vec{w}(t)(1 - 2\eta\lambda) - \eta \nabla E_{in}(\vec{w})}$$ $\rightarrow$ same

(b) with L_1 regulizer

$$E_{avg}(\vec{w}) = E_{in}(\vec{w}) + \lambda \|\vec{w}\|_1$$

$$\text{sign}_1(\vec{x}_i) = \begin{cases} +1 & \text{if } x_i > 0 \\ 0 & \text{if } x_i = 0 \\ -1 & \text{if } x_i < 0 \end{cases}$$

$$\nabla E_{avg}(\vec{w}) = \nabla E_{in}(\vec{w}) + \lambda \frac{\partial}{\partial w_i} \|\vec{w}\|_1$$

$$\boxed{\nabla E_{avg}(\vec{w}) = \nabla E_{in}(\vec{w}) + \lambda\, \text{sign}_1(\vec{w})}$$

$$\vec{w}(t+1) \leftarrow \vec{w}(t) - \eta\left[\nabla E_{in}(\vec{w}) + \lambda\, \text{sign}_1(\vec{w})\right]$$

$$\leftarrow \underline{\vec{w}(t) - \eta \nabla E_{in}(\vec{w})} - \eta\lambda\, \text{sign}_1(\vec{w})$$

$$\boxed{\vec{w}(t+1) \leftarrow w'(t+1) - \eta\lambda\, \text{sign}_1(\vec{w})}$$

2. LFD Exercise 4.5

$$\vec{w}^T \vec{\Gamma}^T \vec{\Gamma} \vec{w} \le C$$

(a) $\sum_{q=0}^{Q} \vec{w}_q^2 \le C$

This was the case in 4.8a, $\boxed{\vec{\Gamma} = I}$

(b) $\left(\sum_{q=0}^{Q} \vec{w}_q\right)^2 \le C$

Must sum up $\vec{w}$ first so $\boxed{\vec{\Gamma} = [1, 1, 1, \dots 1]}$

with length $\vec{\Gamma}$ = length $\vec{w}$

3. LFD 4.25 a-c

(a) D_{train} size $D-k$

D_{val} size K

This changes K in the second term.

$$E_{out}(g^-_{m^*}) \le E_{val}(g^-_{m^*}) + O\left(\sqrt{\frac{\ln M}{2k}}\right)$$

(a) Should you select learner with minimum validation error?

- No. Each learner has a different number of training points and valitation points. Further, they are not validated on the same dataset so performance could vary greatly.

(b) If all models are validated on the same set - why ok?

- This is ok because all models were trained on the same dataset of same size and validated on the same dataset. So the results are comparable. Also, K is the same size, so the second term will be the same.

(c) Show $\mathbb{P}[E_{out}(m^*) > E_{val}(m^*) + \varepsilon] \le M e^{-2\varepsilon^2 K(\varepsilon)}$

Hoeffding bound: $\mathbb{P}[E_{out}(m^*) - E_{val}(m^*) > \varepsilon] \le 2e^{-2\varepsilon^2 N}$

Add up errors of every M

$$\mathbb{P}[E_{out}(m^*) - E_{val}(m^*) > \varepsilon] \le \sum_{m=1}^{M} 2e^{-2\varepsilon^2 k_m}$$

Subsistute $K(\varepsilon) = \frac{1}{2\varepsilon^2} \ln\left(\frac{1}{M}\sum_{m=1}^{M} e^{-2\varepsilon^2 k_m}\right)$

" " $\le \boxed{2M e^{-2\varepsilon^2 K(\varepsilon)}}$ → average k

4. LFD 5.4

(a)(i) There are a few things wrong with this. Mainly, the M should not be 500. We are currently looking at 500 stocks that have lasted for 50 years. At the start of this time, there were stocks that ended up going bankrupt that aren't in the index now. The fact that we looked at only the 500 biggest companies today is sampling bias.

(ii) $M = 50{,}000$ since any of the 50,000 stocks can make it into the S&P. This makes the bound 100x bigger.

$$100 \times 0.045 = 4.5$$

(b)(i) Same problem as (a). You can't base your decision to buy a stock based on results on a subset of 500 stocks. Further, these are 500 stocks that have survived to today and grown to one of the 500 biggest companies. Sampling bias and data snooping

(ii) We know the retrospective performance of the current S&P stocks is an overestimate of buy and hold performance. If we included all 50,000 stocks and accounted for stocks stopping training we could then say something

(1c)

	L1			L2		
Lambda	E_{in}	E_{out}	Zeros	E_{in}	E_{out}	Zeros
0	0.079	0.103	8	0.079	0.103	8
0.0001	0.079	0.098	8	0.079	0.103	8
0.001	0.077	0.093	13	0.078	0.093	8
0.005	0.081	0.089	16	0.076	0.098	8
0.01	0.09	0.079	18	0.084	0.098	8
0.05	0.14	0.103	32	0.112	0.117	8
0.1	0.188	0.136	37	0.124	0.121	8

For the L1 regularizer, as the weight of the regularizer increased, the E_{in} and E_{out} initially decreased and then increased. The regularizer initially reduced overfitting in the training set which in turn reduced error in the testing set. This can be seen as the E_{in} increases while the E_{out} decreases for Lambda values of 0.001, 0.005, and 0.01. As the Lambda weight increases, the number of zeros in the resulting weight vector also increased.

The L2 regularizer also decreased E_{in} and E_{out} as the weight of lambda increased. The number of zeros remained the same for all weights of lambda, as the L2 does not place a weight on the number zeros. This regularizer did not decrease E_{out} as much as the L1, but did increase E_{in} more, indicating that this regularizer did not reduce overfitting as much.