Convex Set :

Q5) a) A set C is convex if the line segment b/w any 2 pts in C lies in C,

i.e. $\forall x_1, x_2 \in C$, $\forall \theta \in [0,1]$

$$\theta x_1 + (1-\theta) x_2 \in C$$

Convex Function :

Let C be a convex function of R^n.
A function $F: C \to R$ is called convex, if

$$F(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y) \quad , \forall x, y \in C$$

If F is a convex function, then all its level sets $\langle x \in C \mid F(x) \leq a \rangle$, where 'a' is a scalar, are convex.

Defn: A convex function is a continuous function whose values at the midpoint of every interval in its domain doesn't exceed the arithmetic mean of its values at the ends of interval.

A function $f(x)$ is said to be strictly convex if for every (A, B) in the domain, the line segment obtained by joining these strictly lies above the curve except the two end points which would be $(A, f(A))$,

$(B, f(B))$ that are common.

Convexity of Lasso & Ridge Regression.

For LASSO the lemma has been proved that,
For any y, x and $\lambda \geq 0$ and lasso solutions $B(1)$ and $B(2)$ must satisfy $B_i^{(1)}$ and $B_i^{(2)} \geq 0$ For $i = 1 \cdots p$. In other words, any two LASSO solutions must have same signs over the common support.

Hence, LASSO penalty is not strict because if A, B have the same sign then the line segment & curve are eql. which means infinite no of points could be common.

In Ridge regression, the minimisation problem is

$$J(w) = \|y - Xw\|_2^2 + \lambda \|w\|_2^2 \quad \text{—— (1)}$$

Another defn for convex loss function is that $\hat{\beta} \in \underset{}{\operatorname{argmin}} F(X\beta) + \lambda \|\beta\|_1$ —— (2)

where the loss function $F : R^n \to R$ is differentiable & strictly convex.

By joining (1) & (2) we can say that Ridge regression is strictly convex.

Q5(b)

We prefer odd values for K because if we take an even value of K then a situation can arise when there is no tie breaker.

If we are dealing with problem of binary classification & 50% of neighbours are of its class & rest 50% of class 2. Then it would be impossible to decide which class current points belongs.

But, if we have a odd K then these kind of situations can be avoided.

so K=3 would be better than K=2.

Q5(c)

Assuming training set D consists of N pts(x_i, y_i) sampled IID,

let the classifier trained on D be H_D

Given that:

$y = H_D(x)$ is a distribution with mean 'μ' & variance σ^2

$$\text{Total loss} = \underbrace{E_{xD}\left[\underbrace{L(H_D(x), y)}_{\text{loss over a single } x}\right]}_{\text{Avg loss over all 'x'}}$$

$$\boxed{\text{let } L(h(x), y) = \frac{1}{2}\left(h(x) - y\right)^2}$$ Squared loss

$$E_{xD}\left[\left(H_D(x) - y\right)^2\right] = E_{xD}\left[\left(H_D(x) - E_D[H_D(x)] + E_D[H_D(x)] - y\right)^2\right]$$

$$E_{x,D}\left[\left(H_D(x) - E_D H_D(x)\right)\right]^2 + \left(E_D(H_D(x)) - y\right)^2 +$$

$$2\left(H_D(x) - E_D[H_D(x)]\right)\left(E_D[H_D(x) - y]\right)$$

$$\Rightarrow \underbrace{E_{x,D}\left[H_D(x) - E_D[H_D(x)]\right]^2}_{\text{Variance}} + E_x\left[\underbrace{\left(E_D[H_D(x)] - y\right)^2}_{\text{bias}}\right]$$

Suppose K models are trained on K subsets of D $\{D_i\}_{i=1}^{K}$

The bias term, given some data point (x,y) depends on $E_D[H_D(x)]$ only

$$H_D(x) = \frac{1}{K}\sum_{i=1}^{K} H_{D_i}(x)$$

$$E_D(H_D(x)) = \frac{1}{K}\sum_{i=1}^{K} E_{D_i}(H_{D_i}(x))$$

$$= \frac{1}{K} \times K\mu$$

$$= \mu$$

Thus, ensembling doesn't change the bias term.

Now,
$$H_D(x) = \frac{1}{K}\sum_{i=1}^{K} H_{D_i}(x)$$

$$\Rightarrow Var(H_D(x)) = Var\left(\frac{1}{K}\sum_{i=1}^{K} H_{D_i}(x)\right)$$

$\{Var(Kx) = K^2 Var(x)\}$
$$= \frac{1}{K^2} Var\left(\sum_{i=1}^{K} H_{D_i}(x)\right)$$

$$= \frac{1}{K^2}\left(\sum_{i=1}^{K} Var(H_{D_i}(x))\right)$$

$$= \frac{1}{K^2} \times K\sigma^2 = \frac{1}{K}\sigma^2$$

We now reduce the variance term to reduce the loss as the bias term doesn't change.

thus $\underbrace{\frac{L}{K}\cancel{\sigma}^2}_{\text{variance of ensemble models}} < \underbrace{\cancel{\sigma}^2}_{\text{variance of single model trained on Entire data}}$

$$\boxed{L < K}$$

thus, there must be more than L learners for the ensemble to perform better than the single bigger model.