1 These questions have both appeared in past examinations.

(a) If $\{X_i | i = 1,2,\ldots,n\}$ is a set of correlated random variables such that

$$\mathbb{E}\left[X_i\right] = \mu \qquad \mathbb{E}\left[(X_i-\mu)(X_j-\mu)\right] = \begin{cases} \sigma^2 & \text{if } i = j \\ \rho\sigma^2 & \text{if } i \neq j \end{cases}$$

show

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] = \rho\sigma^2 + \frac{(1-\rho)\sigma^2}{n}$$

[10 marks]

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}(X_i-\mu)\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\mu)(X_j-\mu)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[(X_i-\mu)(X_j-\mu)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 + \frac{1}{n^2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n}\rho\sigma^2 = \frac{1}{n^2}n\sigma^2 + \frac{1}{n^2}n(n-1)\rho\sigma^2 \\
&= \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2
\end{aligned}$$

(b) We consider a regression problem where the data $(\boldsymbol{x},y)$ is distributed according to $\gamma(\boldsymbol{x},y)$. We consider a learning machine that makes a prediction $\hat{f}(\boldsymbol{x}|\boldsymbol{\theta})$, where the parameters, $\boldsymbol{\theta}$ are trained using a stochastic algorithm that returns parameters distributed according to a probability density $\rho(\boldsymbol{\theta})$. We can define the mean machine as $\hat{m}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\hat{f}(\boldsymbol{x}|\boldsymbol{\theta})\right]$. We assume that

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\gamma}\left[\left(\hat{m}(\boldsymbol{x}) - y\right)^2\right] = B, \qquad \mathbb{E}_{(\boldsymbol{x},y)\sim\gamma}\left[\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\left(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}) - \hat{m}(\boldsymbol{x})\right)^2\right]\right] = V.$$

That is, we can define a bias B and variance V. We now consider ensembling n machines

$$\hat{f}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i$ are drawn independently from $\rho(\boldsymbol{\theta})$. Compute the expected generalisation error of $\hat{f}_n(\boldsymbol{x})$. (Note this is different from the usual bias-variance calculation because we are averaging the performance of n machines). [10 marks]

We note that $\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{f}_n(\boldsymbol{x})\right] = \hat{m}(\boldsymbol{x})$ so the bias variance dilemma calculation still holds

$$\begin{aligned}
\mathcal{E}_n &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left[\left(\hat{f}_n(\boldsymbol{x}) - y\right)^2\right]\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left[\left((\hat{f}_n(\boldsymbol{x}) - \hat{m}(\boldsymbol{x})) + (\hat{m}(\boldsymbol{x}) - y)\right)^2\right]\right] \\
&= \mathbb{E}_{(\boldsymbol{x},y)}\left[\left(\hat{m}(\boldsymbol{x}) - y\right)^2\right] + \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left[\left(\hat{f}_n(\boldsymbol{x}) - \hat{m}(\boldsymbol{x}))\right)^2\right]\right]
\end{aligned}$$

where the cross term vanishes as

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\left(\hat{f}_n(\boldsymbol{x}) - \hat{m}(\boldsymbol{x})\right)\left(y - \hat{m}(\boldsymbol{x})\right)\right] = \left(y - \hat{m}(\boldsymbol{x})\right)\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{f}_n(\boldsymbol{x}) - \hat{m}(\boldsymbol{x})\right] = 0.$$

Thus $\mathcal{E}_n$ consists of a bias B and a new variance V_n where

$$\begin{aligned}
V_n &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left[\left(\hat{f}_n(\boldsymbol{x}) - \hat{m}(\boldsymbol{x}))\right)^2\right]\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \hat{m}(\boldsymbol{x})\right)\right)^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \hat{m}(\boldsymbol{x})\right)^2 + \frac{1}{n^2}\sum_{i,j=1 j\neq i}^{n}\left(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \hat{m}(\boldsymbol{x})\right)\left(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_j) - \hat{m}(\boldsymbol{x})\right)\right]\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \hat{m}(\boldsymbol{x})\right)^2\right]\right] = \frac{V}{n}
\end{aligned}$$

since

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\left(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \hat{m}(\boldsymbol{x})\right)\left(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_j) - \hat{m}(\boldsymbol{x})\right)\right] = 0$$

as $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are independent. Thus the expected generalisation

performance is equal to

$$\mathcal{E}_n = B + \frac{V}{n}.$$

That is, by averaging over n different machines we reduce the variance by n. Note that we assume the predictions of the machine where independent so

$$\mathbb{E}_{\boldsymbol{\theta}}\Big[\Big(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \hat{m}(\boldsymbol{x})\Big)\Big(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_j) - \hat{m}(\boldsymbol{x})\Big)\Big] = 0$$

If the predictions of the machines are correlated then we don't do so well.

End of question 1

END OF PAPER