Question 1 :

Given that $E_{agg}(x) = E\left[\left\{\frac{1}{M}\sum_{i=1}^{M} h_i(x) - f(x)\right\}^2\right]$

$$\varepsilon_i(x) = f(x) - h_i(x)$$

hence $E_{agg}(x) = E\left\{\frac{1}{M}\sum_{i=1}^{M} \varepsilon_i(x)\right\}^2\Big] \rightarrow (1)$

→ we are also given assumptions that each of the errors have a 0 mean

$$E(\varepsilon_i(x)) = 0 \quad \text{for all } i$$

and errors are uncorrelated

$$E(\varepsilon_i(x)\,\varepsilon_j(x)) = 0 \quad \text{for all } i \neq j$$

From (1),

$$E_{agg}(x) = E\left\{\frac{1}{M}\sum_{i=1}^{M} \varepsilon_i(x)\right\}^2\Big]$$

$$= \frac{1}{M^2}\sum_{i=1}^{M}\left[E(\varepsilon_i(x))^2\right]$$

$(\because E(ax+b) = aE(x)+b)$

$$= \frac{1}{M} \times \frac{1}{M} \times \sum_{i=1}^{M}\left[E(\varepsilon_i(x))^2\right]$$

$$= \frac{1}{M} \times \left[\frac{1}{M}\sum_{i=1}^{M} E(\varepsilon_i(x))^2\right]$$

According to our question,

$$E_{avg} = \frac{1}{M} \sum_{i=1}^{M} E(\varepsilon_i(x)^2)$$

Hence, $E_{agg}(x) = \frac{1}{M} \times E_{avg}$

Question 2:-

Jensen's inequality states that for any convex function f

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i)$$

and we have to assume that each of the errors generated from different models using bootstrap samples are correlated.

We know that, if f is a convex function on (a,b) and X is a random variable, then
$f(E(x)) \leq E(f(x))$

$$E(f(x)) = \lambda_1(x_1) f(x_1) + \sum_{i=2}^{m} \lambda(x_i) f(x_i)$$

$$= \lambda(x_1) f(x_1) + (1-\lambda)(x_i) \sum_{i=2}^{m} \lambda(x_i) f(x_i) / (1-\lambda)(x_i)$$

$$f(E(x)) \leq \lambda(x_1) f(x_1) + (1-\lambda)(x_1) f\left(\sum_{i=2}^{m} \frac{\lambda(x_i)\, x_i}{(1-\lambda)(x_i)} \right)$$

$$f(E(x)) \leq f\left(\lambda(x_1) x_1 + (1-\lambda)(x_1) \left(\sum_{i=2}^{m} \lambda(x_i)(x_i) \middle/ (1-\lambda)(x_i) \right) \right)$$

$$E f(x) \leq f(E(x))$$

Hence $E_{agg} \leq E_{avg}$

Question 3:-

We are given that

Hypothesis for boolean Classification problem

$$H(x) = \text{sign}\left(\sum_{t=1}^{N} \alpha_t h_t(x) \right) \rightarrow (1)$$

The weight for point i at time step $t+1$ is

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times e^{-\alpha_t h_t(i) y(i)} \rightarrow (2)$$

Here, $D_t(i) \rightarrow$ normalized weight for point i at time t

$Z_t \rightarrow$ normalization factor at step t

$y(i) \rightarrow$ true label for point i

$\alpha_t \rightarrow$ Voting parameter for hypothesis h_t

$h_t(i) \rightarrow$ hypothesis function for i at time t

$D_1 = \frac{1}{N}$ fd all points $i \rightarrow$ (3)

The error E_t for Adaboost process can be measured with respect to D_t as

$$E_t = \sum_{i\, h_t(i) \neq y(i)} D_t(i) \rightarrow (4)$$

as the error at a particular time t is sum of weights corresponding to all points i which are mis classified i.e, $h_t(i) \neq y_i$

Now, from 2,

$$D_{(t+1)}i = \frac{D_t(i)}{Z_t} \times e^{-\alpha_t h_t(i) y(i)}$$

Since $h_t(i)$ and y_i are in both in $\{-1, 1\}$

$$D_{t+1}(i) = D_1(i) \cdot \frac{e^{-\alpha_1 h_1(i) y(i)}}{Z_1} \times \frac{e^{-\alpha_2 h_2(i) y(i)}}{Z_2} \times \cdots \times \frac{e^{-\alpha_t h_t(i) y(i)}}{Z_t}$$

$$= \frac{1}{N} \frac{e^{-\sum_{j=1}^{t} \alpha_j h_j(i) y(i)}}{\prod_{j=1}^{t} z_j}$$

where $f_t(i) = -\sum_{j=1}^{t} \alpha_j h_j(i)$

Now, total training error of $h(x)$

$$T_H = \frac{1}{N} \sum_{i = h(i) \neq y(i)} 1$$

average of mis-classified points also known as accuracy loss

Now, since $H(i) = \text{sign}(f(i))$

$$T_H = \frac{1}{N} \sum_{i = y(i) f(i) \leq 0} 1$$

because, for misclassified points y_i and $f(i)$ would have opposite signs and hence $y_i f(i) \leq 0$

$$\Rightarrow T_H \leq \frac{1}{N} \sum_i e^{-y(i) f(i)}$$

$$\Rightarrow T_H \leq \left(\prod_t z_t\right) \left(\sum_i D(t+1)^i\right)$$

D_{t+1} is a probabability distribution

hence $\sum_i D(t+1)^i = 1$

$$\Rightarrow T_H \leq \prod_t z_t \rightarrow ⑥$$

$$Z_t = \sum_i D_t(i)\, e^{-\alpha_t h_t(i) y(i)}$$

$$= \sum_{i: h_t(i)=y(i)} D_t(i) e^{-\alpha_t} + \sum_{i=h_t(i)\neq y(i)} D_t(i) e^{\alpha_t}$$

because for $h_t(i) = y(i)$

$h_t(i) \times y(i) = 1$ as both would be 1

or both would be -1

If $h_t(i) \neq y(i)$, then $h_t(i) \times y(i) = -1$ as only one of them would be 1.

$$\Rightarrow Z_t = e^{-\alpha_t} \sum_{i: h_t(i)=y(i)} D_t(i) + e^{\alpha_t} \sum_{i: h_t(i)\neq y(i)} D_t(i)$$

$$Z_t = e^{-\alpha_t}(1-\varepsilon_t) + e^{\alpha_t}\varepsilon_t \quad \text{(from 4)}$$

Now, to minimize error T_H,

α_t comes out to be $\frac{1}{2} \frac{1-\varepsilon_t}{\varepsilon_t}$

putting it in above eqn,

$$Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$

$\varepsilon_t = \frac{1}{2} - \gamma_t$ gives

$$Z_t = 2\sqrt{\left(\frac{1}{2}-\gamma_t\right)\left(\frac{1}{2}+\gamma_t\right)}$$

$$= \frac{\cancel{2}\sqrt{1-4\gamma_t^2}}{\cancel{2}} = \sqrt{1-4\gamma_t^2}$$

Since, $1+x \leq e^{x} \quad \forall x \in \mathbb{R}$

$$\Rightarrow 1-4\gamma_t^2 \leq e^{-4\gamma_t^2}$$

$$z_t \leq \sqrt{e^{-4\gamma_t^2}} = e^{-2\gamma_t^2}$$

putting z_t in 6 gives

$$T_H \leq \prod_t z_t$$

$$\Rightarrow T_H \leq \prod_t e^{-2\gamma_t^2}$$

$$\Rightarrow \boxed{T_H \leq e^{-2\sum_{t=1}^{T} \gamma_t^2}}$$

Hence proved.