Problem 1

Problem 1.1

$$\mathcal{L}(f) := \sum_{i=1}^{N} exp(-y_i f(x_i))$$

We can prove that AdaBoost is equivalent to stagewise minimization of the above exponential loss by using a greedy approach with additive models.
We can consider a binary classifier described as follows:

$$f(x) = sign(\sum_{t=1}^{N} \beta_t h_t(x)) \tag{1}$$

We need to calculate the empirical error minimization by solving the optimization problem:

$$\min_{\beta_t, h_t} \sum_{i=1}^{n} \mathcal{L}(\sum_{t=1}^{N} \beta_t h_t(x_i) y_i)$$

Because it is not realistic to optimize $\beta_1....\beta_t$ and $h_1....h_t$, we can use a greedy approach. We can add a classifier to be optimized at each stage of the algorithm (**stage wise learning of additive models**). Given equation 1, for each t we have:

$$f_t(x) = \sum_{i=1}^{t} \beta_i h_i(x)) = f_{t-1}(x) + \beta_t h_t(x)$$

With the greedy approach, at each stage t we have already learnt $t-1$ classifiers and we leave them as they are, so the problem reduces to find the optimal values of β_t and h_t. So in the end we have:

$$(\beta_t, h_t) = arg \min_{\beta >= 0, h} \sum_{i=1}^{N} exp(-y_i(f_{t_1}(x_i) + \beta h(x_i))$$

Problem 1.2

$$\frac{1}{N} \sum_{i=1}^{N} I(y_i \neq F(x_i)) \leq \prod_{t=1}^{T} \sqrt{1 - 4\gamma_t^2}$$

We know that the $0-1$ loss is upper bounded by the exponential loss, i.e.

$$I(y_i \neq F(x_i)) \leq exp(-y_i f_T(x_i))$$

We can continue with:

$$\frac{1}{N} \sum_{i=1}^{N} I(y_i \neq F(x_i)) \leq \frac{1}{N} \sum_{i=1}^{N} exp(-y_i f_T(x_i))$$

$$\frac{1}{N}\sum_{i=1}^{N} exp(-y_i f_T(x_i)) = \frac{1}{N}\sum_{i=1}^{N} exp(-y_i \sum_{t=1}^{T}\beta_t h_t(x_i))$$

$$= \frac{1}{N}\sum_{i=1}^{N}\prod_{t=1}^{T} exp(-y_i\beta_t h_t(x_i))$$

The AdaBoost algorithm states that:

$$D_{t+1} = \frac{D_t(i)exp(-\beta_t y_i h_t(x_i))}{Z_t}$$

So we have:

$$\frac{1}{N}\sum_{i=1}^{N}\prod_{t=1}^{T} exp(-y_i\beta_t h_t(x_i)) = \frac{1}{N}\sum_{i=1}^{N}\prod_{t=1}^{T} Z_t\frac{D_{t+1}(i)}{D_t(i)}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\prod_{t=1}^{T} Z_t\frac{D_{t+1}(i)}{D_1(i)}$$

We remember that $D_1(i) = \frac{1}{N}$ and due to the normalization, $\sum_{i=1}^{N} D_{t+1} = 1$, in the end we have:

$$\frac{1}{N}\sum_{i=1}^{N}\prod_{t=1}^{T} exp(-y_i\beta_t h_t(x_i)) = \prod_{t=1}^{T} Z_t$$

Let's now find the value of Z_t:

$$Z_t = \sum_{i=1}^{N} D_t(i)exp(-y_i\beta_t h_t(x_i))$$

$$= \sum_{i:y_i=h_t(x_i)} D_t(i)exp(-\beta_t) + \sum_{i:y_i\neq h_t(x_i)} D_t(i)exp(\beta_t)$$

$$= exp(-\beta_t)(1-\epsilon_t) + exp(\beta_t)\epsilon_t$$

By solving the above equation, we can find that $\beta_t = \frac{1}{2}ln\frac{1-\epsilon_t}{\epsilon_t}$, and note that $\epsilon_t = \frac{1}{2} - \gamma$ (stated in the hypothesys):

$$= exp(-(\frac{1}{2}ln\frac{1-\epsilon_t}{\epsilon_t}))(\frac{1}{2}+\gamma_t) + exp(\frac{1}{2}ln\frac{1-\epsilon_t}{\epsilon_t})(\frac{1}{2}-\gamma_t)$$

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

Based on the above results, we have proved that:

$$\frac{1}{N}\sum_{i=1}^{N} I(y_i \neq F(x_i)) \leq \prod_{t=1}^{T}\sqrt{1-4\gamma_t^2}$$