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

Problem 1.2.1

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

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:

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

Problem 1.2.2

$$\prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} \sqrt{1-4\gamma_t^2}$$

The predictive function is:

$$f_pred = \frac{f_T}{\prod_{t=1}^{T} \beta_t}$$

$yf(x) \leq \theta$ means that:

$$y \sum_{t=1}^{T} \beta_t h_t(x) \leq \theta \sum_{t=1}^{T} \beta_t$$

which is true if and only if:

$$I(yf(x) \leq \theta) \leq exp(-y \sum_{t=1}^{T} \beta_t h_t(x) + \theta \sum_{t=1}^{T} \beta_t)$$

Having said that:

$$\frac{1}{N} \sum_{i=1}^{N} I(y_i f_T(x_i) \leq \theta) \leq \frac{1}{N} \sum_{i=1}^{N} exp(-y \sum_{t=1}^{T} \beta_t h_t(x) + \theta \sum_{t=1}^{T} \beta_t()$$

$$= \frac{exp(\theta \sum_{t=1}^{T} \beta_t)}{N} \sum_{i=1}^{N} exp(-y \sum_{t=1}^{T} \beta_t h_t(x) + \theta \sum_{t=1}^{T} \beta_t()$$

$$= exp(\theta \sum_{t=1}^{T} \beta_t)(\prod_{t=1}^{T} Z_t)$$

Because we are considering a simplified case, we can ignore $\sum_{t=1}^{T} \beta_t$, thus getting:

$$= exp(\theta)(\prod_{t=1}^{T} Z_t)$$

This proves that the training error at margin θ satisfies:

$$\frac{1}{N} \sum_{i=1}^{N} I(y_i f_T(x_i) \leq \theta) \leq exp(\theta)(\prod_{t=1}^{T} Z_t)$$

Problem 2

Problem 2.1

x_i^{l-1} being symmetric means that the distribution is an even function, so we have:

$$E[\sigma(x_i^{l-1})^2] = \frac{1}{2}E[(x_i^{l-1})^2]$$

$$\frac{1}{2}E[(x_i^{l-1})^2] = \frac{1}{2}(var(x_i^{l-1}) + E[x_i^{l-1}]^2)$$

Note that the mean $E[x_i^{l-1}] = 0, \forall l$, we have:

$$E[\sigma(x_i^{l-1})^2] = \frac{1}{2}var(x_i^{l-1}) \tag{2}$$

We can now proceed to prove that:

$$var(x_i^l) = \sum_{j=1}^{N_{l-1}}(var(W_{ij}^l)var\sigma(x_j^{l-1}) + var(W_{ij}^l)E[\sigma(x_j^{l-1})]^2))$$

$$= \sum_{j=1}^{N_{l-1}}(var(W_{ij}^l)(E[\sigma(x_j^{l-1})^2] - E[\sigma(x_j^{l-1})^2]) + var(W_{ij}^l)E[\sigma(x_j^{l-1})]^2))$$

$$= \sum_{j=1}^{N_{l-1}}(var(W_{ij}^l)E[\sigma(x_j^{l-1})]^2))$$

$$= N^{l-1}var(W_{ij}^l)E[\sigma(x_j^{l-1})]^2$$

Based on 2, we have:

$$var(x_i^l) = N^{l-1}var(W_{ij}^l)\frac{1}{2}var(x_i^{l-1})$$

In order for the variance of the network output to not explode or vanish, the mean and variance of the activations in layer l and $l-1$ must be "equal", which means that $var(x_i^l) = varx_j^{l-1}$. To prove this, we must achieve $\frac{1}{2}N^{-1}var(W_{ij}^l) = 1$:

$$var(W_{ij}^l) = \frac{2}{N^{l-1}}$$

$$var(x_i^l) = \frac{1}{2}\frac{2}{N^{l-1}}N^{l-1}var(x_i^{l-1})$$

Thus finally proving that:

$$var(x_i^l) = var(x_i^{l-1})$$

We know that the standard deviation is just the square root of the variance:

$$\sigma_l = \sqrt{(}\frac{2}{N^{l-1}}) = \frac{\sqrt{2}}{\sqrt{N^{l-1}}}$$

So $\alpha = \sqrt{2}$

Problem 2.2

$$\frac{\partial Loss(W)}{\partial x_i^l} = W^l \frac{\partial Loss(W)}{\partial y_i^l} \tag{3}$$

In back propagation we have $\frac{\partial Loss(W)}{\partial y_i^l} = f'(y^l)\frac{\partial Loss(W)}{x_i^{l+1}}$.
In case of the ReLu activation, $f'(y_i^l) = 0$ or 1, and because they are independent, we have:

$$E[\frac{\partial Loss(W)}{\partial y_i^l}] = \frac{1}{2}E[\frac{\partial Loss(W)}{x_i^{l+1}}] = 0$$

$$E[(\frac{\partial Loss(W)}{\partial y_i^l})^2] = \frac{1}{2}var(\frac{\partial Loss(W)}{x_i^{l+1}})$$

By applying the above equations on 3, we get:

$$var(\frac{\partial Loss(W)}{\partial x_i^l}) = \frac{1}{2}N^l var(W^l)var(\frac{\partial Loss(W)}{x_i^{l+1}})$$

A sufficient condition for the gradient to not explode or vanish is:

$$var(W_{ij}^l) = \frac{2}{N^l}$$

Thus getting:

$$var(\frac{\partial Loss(W)}{\partial x_i^l}) = \frac{2}{N^l}\frac{1}{2}N^l var(\frac{\partial Loss(W)}{\partial x_i^{l+1}})$$

$$var(\frac{\partial Loss(W)}{\partial x_i^l}) = var(\frac{\partial Loss(W)}{\partial x_i^{l+1}})$$

Problem 3

Problem 3.1

$$P(x|\mu) = \frac{n!}{\prod_i x_i!}\prod_i \mu_i^{x_i}, i = 1...d$$

where$x_i \in \mathbb{N}, \sum_i x_i = n$ and $0 < \mu_i < 1, \sum_i \mu_i = 1$.
The log-likelihood is:

$$l(P(x|\mu)) = log(\frac{n!}{\prod_i x_i!}\prod_i \mu_i^{x_i})$$

$$= logn! + \sum_{i=1} x_i log\mu_i - \sum_{i=1} logx_i!$$

Now we can proceed with the lagrange multiplier:

$$\mathcal{L}(\mu, \lambda) = l(P(x|\mu)) + \lambda(1 - \sum_{i=1} \mu_i)$$

Now we need to set the partial derivates to 0:

$$\frac{\partial}{\partial \mu_i}\mathcal{L}(\mu, \lambda) = \frac{\partial}{\partial \mu_i} l(P(x|\mu)) + \frac{\partial}{\partial \mu_i}(1 - \sum_{i=1} \mu_i) = 0$$

$$\frac{\partial}{\partial \mu_i} \sum_{i=1} x_i log \mu_i - \lambda \frac{\partial}{\partial \mu_i}(\sum_{i=1} \mu_i) = 0$$

$$\frac{x_i}{\mu_i} - \lambda = 0$$

$$\mu_i = \frac{x_i}{\lambda}$$

The value of λ is:

$$\sum_{i=1} \mu_i = \sum_{i=1} \frac{x_i}{\lambda}$$

$$1 = \frac{1}{\lambda} \sum_{i=1} x_i$$

$$\lambda = n$$

So, in the end we have:

$$\mu_i = \frac{x_i}{n}$$

Problem 3.2