1 Question 1

Since the VC-Dimension($\mathcal{H}$) $= d$, for a set S with $|S| = D$ we know that there are at most D^d ways of partitioning with function in $\mathcal{H}$. Given that our weighted voting is based on t functions of $\mathcal{H}$, 2^D is upper bounded by $(D^d)^t$ since each function can independently give a different labeling (possibly overlapping, nonetheless an upper bound). Thus, $2^D \leq D^{dt}$ and $D \leq 2dt\log(dt) = \mathrm{O}(dt\log(dt))$.

2 Question 2

We assume $h_t = h_{t+1}$ and will show $\epsilon_{t+1} \not< \frac{1}{2}$. This will contradict the initial assumption that h was a weak learner with $\epsilon < \frac{1}{2}$, thus $h_t \neq h_{t+1}$.

$$\epsilon_t = \sum_{i=1}^{m} D_t(i)\mathbb{1}_{h_t(x_i)\neq y_i} < \frac{1}{2} \tag{1}$$

$$\epsilon_{t+1} = \sum_{i=1}^{m} D_{t+1}(i)\mathbb{1}_{h_t(x_i)\neq y_i} = \sum_{i=1}^{m} \frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{Z_t}\mathbb{1}_{h_t(x_i)\neq y_i} \tag{2}$$

When $h_t(x_i) = y_i$, $\mathbb{1}_{h_t(x_i)\neq y_i} = 0$ and when $h_t(x_i) \neq y_i$, $\mathbb{1}_{h_t(x_i)\neq y_i} = 1$. So we can replace $y_i h_t(x_i)$ with -1 without changing the resultant sum, because whenever it's not valid $\mathbb{1}_{h_t(x_i)\neq y_i}$ will be zero and it will not make a difference anyways.

$$\epsilon_{t+1} = \sum_{i=1}^{m} \frac{D_t(i)e^{\alpha_t}}{Z_t}\mathbb{1}_{h_t(x_i)\neq y_i} = \frac{e^{\alpha_t}}{Z_t}\sum_{i=1}^{m} D_t(i)\mathbb{1}_{h_t(x_i)\neq y_i} \tag{3}$$

$$\Rightarrow \epsilon_{t+1} = \frac{e^{\alpha_t}}{Z_t}\epsilon_t = \frac{\left(\frac{1-\epsilon_t}{\epsilon_t}\right)^{1/2}}{2[\epsilon_t(1-\epsilon_t)]^{1/2}}\epsilon_t = \frac{1}{2} \not< \frac{1}{2} \tag{4}$$

3 Question 3

Given that $\forall x_i \neq x_j \Leftrightarrow K(x_i, x_j) = 0$ we understand that points in space defined by Φ are pairwise orthogonal. This suggests that $\dim(\Phi) \geq m$ as points are pairwise independent in that space and their span will have exactly m dimensions. Though if we assume the space X is finite we can use all the points, in which case the space defined by $P := \Phi(X)$ will be fixed. Furthermore, the space defined by $\Phi(S), S \subseteq X$ is a subspace of P.

(a) For dataset $S = \{(x_i, y_i)\}_{i=1}^{m}$ I propose $\Phi : X \to \mathbb{R}^m$ to be as follows (More generally if we have an enumeration of the finite space $X := \{(x_i, y_i)\}_{i=1}^{|X|}$):

$$\Phi(X) : [X == x_1, X == x_2, X == x_3, \ \ldots, \ X == x_m]^T \tag{5}$$

$$\text{or the general case } \Phi(X) : [X == x_1, X == x_2, X == x_3, \ \ldots, \ X == x_{|X|}]^T \tag{6}$$

Where each $X == x_i$ is either true($\equiv 1$) or false($\equiv 0$). It can be easily seen that $K(x_i, x_j) = \Phi(x_i)^T\Phi(X_j) = 1$ if and only if $x_i = x_j$. Furthermore, the kernel matrix associated to S and K is I_m, which is a symmetric positive semidefinite matrix thus satisfying the kernel condition. Therefore K is a legal kernel.

(b) Given that each point $\Phi(x_i)$ lies in a different dimension, let w be the vector whose ith coordinate is $y_i \in \{-1, +1\}$. Could the hyperplane $w \cdot \Phi(X_i)$ separate these points any easier?

$$\text{sgn}(w \cdot \Phi(X_i)) = y_i. \tag{7}$$

(c) This is no more than a table lookup! If we've already seen a point before, we can recover the label. If not, this new point will always be mapped to 0 regardless of its y(or in the general case to a new dimension through which our hyperplane does not pass). This kernel does not offer any generalization for our learning task and it doesn't make learning any easier.

4 Question 4

(a) **Claim.** $w_{t+1}^T\bar{w} \geq w_t^T\bar{w} + \gamma$
Proof. Assume that $y_i = 1$.

$$w_{t+1}^T\bar{w} = (w_t + x_i)^T\bar{w} = w_t^T\bar{w} + x_i^T\bar{w} \tag{8}$$

$$\Rightarrow w_{t+1}^T\bar{w} \geq w_t^T\bar{w} + \gamma \tag{9}$$

By induction on this claim we can also see $w_{t+1}^T\bar{w} \geq (t+1)\gamma$. By Cauchy-Schwarz inequality we have:

$$w_{t+1}^T\bar{w} \leq ||w_{t+1}||.||\bar{w}|| \leq ||w_{t+1}|| \tag{10}$$

$$\Rightarrow ||w_t|| \geq t\gamma \tag{11}$$

(b) Assume $y_i = 1$.

$$||w_{t+1}||^2 = ||w_t||^2 + 2w_t^T x_i + ||x_i||^2 = ||w_t||^2 + 2w_t^T x_i + 1 \tag{12}$$

$$\Rightarrow ||w_{t+1}||^2 = ||w_t||^2(1 + \frac{2}{||w_t||}.\frac{w_t^T x_i}{||w_t||} + \frac{1}{||w_t||^2}) \tag{13}$$

$$\Rightarrow ||w_{t+1}||^2 \leq ||w_t||^2(1 + \frac{2(1-\epsilon)\gamma}{||w_t||} + \frac{1}{||w_t||^2}) \tag{14}$$

$$\Rightarrow ||w_{t+1}|| \leq ||w_t||\sqrt{(1 + \frac{2(1-\epsilon)\gamma}{||w_t||} + \frac{1}{||w_t||^2})} \leq ||w_t||(1 + \frac{(1-\epsilon)\gamma}{||w_t||} + \frac{1}{2||w_t||^2}) \tag{15}$$

$$\Rightarrow ||w_{t+1}|| \leq ||w_t|| + (1-\epsilon)\gamma + \frac{1}{2||w_t||} \tag{16}$$

At (15) we have applied the Taylor Approx. of $\sqrt{1+x}$ around $x = 0$.

(c) We consider two different cases:

- $||w_t|| < 1/(\epsilon\gamma) \Rightarrow ||w_{t+1}|| \leq ||w_t|| + ||x_i|| = ||w_t|| + 1 \Rightarrow ||w_{t+1}|| \leq 1/(\epsilon\gamma) + 1$.
- $||w_t|| \geq 1/(\epsilon\gamma) \Rightarrow ||w_{t+1}|| \leq ||w_t|| + (1-\epsilon)\gamma + \frac{1}{2||w_t||}$. Then we replace $||w_t||$ in the denominator with $1/(\epsilon\gamma)$. $||w_{t+1}|| \leq ||w_t|| + (1 - \epsilon/2)\gamma$.

We then have

$$\Rightarrow ||w_t|| < 1/(\epsilon\gamma) + 1 + t(1-\epsilon/2)\gamma \leq 2/(\epsilon\gamma) + t(1-\epsilon/2)\gamma \tag{17}$$

(d) Combining the bounds we have:

$$t\gamma \leq ||w_t|| \leq \frac{2}{\epsilon\gamma} + t\gamma(1-\epsilon/2) \tag{18}$$

$$\Rightarrow t\gamma \leq \frac{2}{\epsilon\gamma} + t\gamma(1-\epsilon/2) \Rightarrow \frac{t\epsilon\gamma}{2} \leq \frac{2}{\epsilon\gamma} \tag{19}$$

$$\Rightarrow t \leq \frac{4}{(\epsilon\gamma)^2} \tag{20}$$

(e) The idea is to perform a binary search on different values of $\gamma \in [0, 1]$. As we go further down the search tree, the possible range for γ^* decreases. More specifically, at level i the possible range for γ^* has a length of $1/2^i$ which means the real value of $\gamma*$ can't be more than $1/2^{i+1}$ away if we select the middle range value. With each γ The margin-perceptron algorithm either returns a weight vector w, or we terminate it after the number of iterations found in (d). Manual termination or a margin less than $(1-\epsilon)\gamma$ is equivalent to failing and we will need to search the left side of the test point. In the other case, we will need to search for γ^* in the right side. Either way, the possible range for γ^* is cut in half.

If we use γ which is η less than γ^* (i.e. $\gamma^* = \gamma + \eta$), we would like to have:

$$(1-2\epsilon)\gamma^* \leq (1-\epsilon)\gamma \leq (1-\epsilon)\gamma^* \tag{21}$$

$$\Rightarrow (1-\epsilon)\gamma^* - \epsilon\gamma^* \leq (1-\epsilon)(\gamma^* - \eta) \tag{22}$$

$$\Rightarrow \frac{\epsilon}{1-\epsilon}\gamma^* \geq \eta \tag{23}$$

(23) gives us an idea on how far away can we be from the γ^* which in turn helps us limit the binary search in that range.

$$\frac{1}{2^i} \leq \frac{\epsilon}{1-\epsilon}\gamma^* \tag{24}$$

Where i denotes the depth of the search. If we're searching the γ in range $[l, u]$, we can replace $\gamma*$ with the lowest possible value which is l.

$$\frac{1}{2^i} \leq \frac{\epsilon}{1-\epsilon}l \leq \frac{\epsilon}{1-\epsilon}\gamma^* \tag{25}$$

$$\Rightarrow i \geq \log(\frac{1-\epsilon}{\epsilon l}) \geq \log(\frac{1-\epsilon}{\epsilon\gamma^*}) \tag{26}$$

So if we go $\log(\frac{1-\epsilon}{\epsilon l})$ deep, we have the guarantee we need. But how many times will we be calling the Margin-Perceptron?

$$\log(\frac{1-\epsilon}{\epsilon\gamma^*}) = \log((\frac{1}{\epsilon}-1)\frac{1}{\gamma^*}) < \log(\frac{1}{\epsilon\gamma^*}) < \log(\frac{1}{\gamma^*})/\epsilon$$
$$\Rightarrow \mathrm{O}(\log(\frac{1}{\gamma^*})/\epsilon)$$