1. Let $z = z_N \times z_{N-1} \dots z_1 = \prod_{n=1}^{N} z_n$. Please derive the formula for z
   Given that
   $$z_t = \sum_{(x_i,y_i)\in D_{train}} w_t(x_i)\exp(-y_i\alpha_t h_t(x_i)) \tag{1}$$
   and
   $$w_{t+1}(x_i) = \frac{1}{z_t} w_t(x_i)\exp(-y_i\alpha_t h_t(x_i)) \tag{2}$$
   then we have another solution for z_t to equal
   $$z_t = \frac{w_t(x_i)}{w_{t+1}(x_i)}\exp(-y_i\alpha_t h_t(x_i)),\ \forall(x_i,y_i)\in D_{train} \tag{3}$$
   meaning that
   $$z = \prod_{n=1}^{N} z_n = \prod_{n=1}^{N} \frac{w_n(x_i)}{w_{n+1}(x_i)}\exp(-y_i\alpha_n h_n(x_i)),\ \forall(x_i,y_i)\in D_{train} \tag{4}$$
   which can be simplified down to
   $$z = \frac{w_1(x_i)}{w_{N+1}(x_i)}\prod_{n=1}^{N}\exp(-y_i\alpha_n h_n(x_i)),\ \forall(x_i,y_i)\in D_{train} \tag{5}$$
   we know that w is uniformly initialized as
   $$w_1(x_i) = \frac{1}{N},\ x_i\in D_{train} \tag{6}$$
   so we can rewrite (5) and further simplify to get
   $$z = \frac{1}{Nw_{N+1}(x_i)}\exp\left(-y_i\sum_{n=1}^{N}\alpha_n h_n(x_i)\right),\ \forall(x_i,y_i)\in D_{train} \tag{7}$$

2. For $H_t = sign\left(\sum_{n=1}^{N}\alpha_n h_n(x)\right)$, show that $Err(H_t) \le z$
   We know that
   $$Err(H_t) = \frac{1}{N}\sum_{(x_i,y_i)\in D_{train}} 1_{H_t(x_i)\neq y_i} \tag{8}$$
   if we assume that every single classification is correct, then we have
   $$Err(H_t) = \frac{0}{N} = 0 \tag{9}$$
   and if they are all incorrect, we have
   $$Err(H_t) = \frac{N}{N} = 1 \tag{10}$$
   now we can take (7) and make the same assumptions. If we have a perfect classifier, then we have
   $$z = \frac{1}{Nw_{N+1}(x)}\exp\left(-\sum_{n=1}^{N}\alpha_n\right) \tag{11}$$

which is a non-zero positive integer greater than $Err(H_t)$ in (9). And then we can assume that we have the worst classifier (where $H_t(x_i) \neq y_i \forall (x_i, y_i) \in D_{train}$)

$$z = \frac{1}{N w_{N+1}(x)} \exp\left(+\sum_{n=1}^{N} \alpha_n\right) \tag{12}$$

which I am assuming will be greater than or equal to 1, so greater than or equal to $Err(H_t)$ in (10).

3. Show why $\alpha_t = \frac{1}{2}\log\left[\frac{1-\epsilon_t(h_t)}{\epsilon_t(h_t)}\right]$ in step 2 of the AdaBoost algorithm

The α_t term is used to see how much each individual weak classifier influences the overall strong classifier. We can go ahead and visualize how this value works by making some observations of the function. First, we notice that if we have an error of exactly 0.5, our α_t value becomes

$$\alpha_t = \frac{1}{2}\log\left[\frac{1-0.5}{0.5}\right] = \frac{1}{2}\log 1 = 0 \tag{13}$$

This means that if we have a error in the middle (as $\epsilon_t \in (0,1)$), it will have no additional influence on the overall classifier. We can make further observations by sweeping the error value across this domain to see how it influences $alpha_t$. We can do something similar by making a function of the error, $\epsilon_t(h)$

$$\alpha_t = f(x = \epsilon_t(h)) = \frac{1}{2}\log\left[\frac{1-x}{x}\right] \tag{14}$$

by taking the inverse, we find

$$\epsilon_t(h) = f^{-1}(y = \alpha_t) = \frac{1}{1+\exp(2y)} \tag{15}$$

from here, we can take the limit of the α_t values to understand its properties

$$\lim_{\alpha_t \to \infty} \frac{1}{1+\exp(2\alpha_t)} = 0 \tag{16}$$

and

$$\lim_{\alpha_t \to -\infty} \frac{1}{1+\exp(2\alpha_t)} = 1 \tag{17}$$

(16) tells us that a small error of 0 will result in a large positive value for α_t, meaning it will have more influence on the model. This makes sense, because we want more input from a low-error classifier in the overall model. (17) tells us that a large error of 1 will result in a large negative value for α_t, meaning it will have less influence on the model. This makes sense, because we want significantly less input from a high-error classifier in the overall model.