1. (a) If $p(x) > 0$, then $\log_2(\frac{1}{p(x)}) \geq 0$, and the product $p(x)\log_2(\frac{1}{p(x)})$ is nonnegative. If $p(x) = 0$, then we have that:

$$\lim_{p(x)\to 0} p(x)\log_2 \frac{1}{p(x)} = 0.$$

Therefore the entropy is nonnegative, being a sum over nonnegative terms.

(b) The proof below uses Jensen's inequality:

$$\begin{aligned}
\mathrm{KL}(p||q) &= \sum_x p(x)\log_2 \frac{p(x)}{q(x)} && (1)\\
&= -\sum_x p(x)\log_2 \frac{q(x)}{p(x)} && (2)\\
&= \mathbb{E}_p\left[-\log_2 \frac{q(x)}{p(x)}\right] && (3)\\
&\geq -\log_2 \mathbb{E}_p\left[\frac{q(x)}{p(x)}\right] && (4)\\
&= -\log_2 \sum_x p(x)\frac{q(x)}{p(x)} \;=\; 0. && (5)
\end{aligned}$$

In line 2, we reformulate the KL-divergence in terms of the negative-log, which is convex. We then apply Jensen's inequality (line 4) to show that the KL-divergence is nonnegative.

(c)

$$\begin{aligned}
&\mathrm{KL}[p(x,y)||p(x),p(y)] && (6)\\
&= -\sum_{x,y} p(x,y)\log\left(\frac{p(x)p(y)}{p(x,y)}\right) \;=\; -\sum_{x,y} p(x,y)\log\left(\frac{p(x)p(y)}{p(x)p(y|x)}\right) && (7)\\
&= -\sum_{x,y} p(x,y)\big(\log p(y) - \log p(y|x)\big) && (8)\\
&= -\sum_y \log p(y)\sum_x p(x,y) \quad + \quad \sum_{x,y} p(x,y)\log p(y|x) && (9)\\
&= -\sum_y p(y)\log p(y) \quad + \quad \sum_x p(x)\sum_y p(y|x)\log p(y|x) && (10)\\
&= H(Y) \;-\; \mathbb{E}_{p(x)}\big[H(Y|X=x)\big] && (11)\\
&= H(Y) \;-\; H(Y|X). && (12)
\end{aligned}$$

2. To use the hint provided in the appendix of the homework, for a fixed t, we define $\phi_t(y) = \frac{1}{2}(y-t)^2$. First, we prove that this function is convex by using the definition of a convex function. Consider $u_1, u_2 \in \mathbb{R}$. For simplicity in the proof, denote $v_1 = u_1 + t, v_2 = u_2 + t$.

For any $0 \leq \lambda \leq 1$ we have,

$$\phi_t(\lambda u_1 + (1-\lambda)u_2) = \frac{1}{2}((\lambda u_1 + (1-\lambda)u_2) - t)^2 \tag{13}$$
$$= \frac{1}{2}(\lambda v_1 + (1-\lambda)v_2)^2 \tag{14}$$
$$\leq \frac{1}{2}(\lambda v_1 + (1-\lambda)v_2)^2 + \frac{1}{2}\lambda(1-\lambda)(v_1 - v_2)^2 \tag{15}$$
$$= \frac{1}{2}\lambda v_1^2 + \frac{1}{2}(1-\lambda)v_2^2 \tag{16}$$
$$= \frac{1}{2}\lambda(u_1 - t)^2 + \frac{1}{2}(1-\lambda)(u_2 - t)^2 \tag{17}$$
$$= \lambda\phi_t(u_1) + (1-\lambda)\phi_t(u_2) \tag{18}$$

The inequality (15) holds since the red term is always non-negative. The red term is the difference between $\lambda\phi_t(u_1) + (1-\lambda)\phi_t(u_2)$ and $\phi_t(\lambda u_1 + (1-\lambda)u_2)$. This shows that ϕ_t is a convex function (Another way to show the convexity is to look at the second derivatives of the function).

Consider the set $X = \{h_1(x), ..., h_m(x)\}$, and suppose a uniform distribution for the elements of X. In other words, for each $h_i(x)$ consider we have a probability $p(h_i(x)) = \frac{1}{m}$. Note that $\mathbb{E}[h(x)] = \bar{h}(x)$. Using the Jensen inequality for ϕ_t and the defined probability distribution,

$$L(\bar{h}(x), t) = \phi_t(\bar{h}(x)) = \phi_t(E[h(x)]) \leq E[\phi_t(h(x))] = \sum_{i=1}^{m} \frac{1}{m}\phi_t(h_i(x)) = \frac{1}{m}\sum_{i=1}^{m} L(h_i(x), t) \tag{19}$$

Note: This question can also be solved without using Jenson inequality, by rewriting the difference between the right and the left hand side of the statement as a sum of positive terms.

3. We write the err'_t using the previous weights as,

$$err'_t = \frac{\sum_{i\in E} w'_i}{\sum_{i\in E} w'_i + \sum_{i\in E^c} w'_i} = \frac{\sum_{i\in E} w_i e^{\alpha_t}}{\sum_{i\in E} w_i e^{\alpha_t} + \sum_{i\in E^c} w_i e^{-\alpha_t}} \tag{20}$$

Multiplying by $\frac{e^{\alpha_t}}{e^{\alpha_t}}$,

$$= \frac{\sum_{i\in E} w_i e^{2\alpha_t}}{\sum_{i\in E} w_i e^{2\alpha_t} + \sum_{i\in E^c} w_i} \tag{21}$$

We also know that $e^{2\alpha_t} = \frac{1-err_t}{err_t}$. Moreover, using the fact in the second tip,

$$\sum_{i\in E} w_i e^{2\alpha_t} = e^{2\alpha_t}\sum_{i\in E} w_i = \frac{1-err_t}{err_t}\sum_{i\in E} w_i \tag{22}$$
$$= \frac{\sum_{i\in E^c} w_i}{\sum_{i\in E} w_i}\sum_{i\in E} w_i = \sum_{i\in E^c} w_i \tag{23}$$

Using this equality in eq (21) yields the result.

This fact means that if the weak learner of the t-th iteration is used for the $t+1$-th iteration, there will be no improvement in our exponential classification loss. Roughly speaking, α_t is the *best* possible ratio at the iteration t for the t-th weak learner. (Similar interpretations might also be correct)