1. Parts (a) and (b) are basically Problem 1 from Homework 2.

For part (c), the update becomes

\[ w_{t+1,i} = w_{t,i} \exp(-\eta L_{abs}(y_t, E_{i,t})) \]

in other words we just substitute the absolute loss as the loss function and leave the value of \( \eta \) open for now. We want to choose \( \hat{y}_t \) such that

\[ L_{abs}(y_t, \hat{y}_t) \leq P_t - P_{t+1} \]

holds both for \( y_t = 0 \) and \( y_t = 1 \), where \( P_t = c \ln W_t \) and \( W_t = \sum_{j=1}^{n} w_{t,j} \) as before. For \( p \in \{0, 1\} \), let \( \Delta_p \) denote the value of \( P_t - P_{t+1} \) if \( y_t = p \) and we update the weights as above. The above condition can be split into two cases, for \( y_t = 0 \) and \( y_t = 1 \), as follows:

\[ \hat{y}_t \leq \Delta_0 \]
\[ 1 - \hat{y}_t \leq \Delta_1. \]

In other words, we need

\[ 1 - \Delta_1 \leq \hat{y}_t \leq \Delta_0. \]

A solution \( \hat{t}_t \) exists if and only if \( \Delta_0 + \Delta_1 \geq 1 \). The main technical challenge in the proof is showing that this indeed is the case, for suitable value of \( c \) and \( \eta \).

It turns out that for any \( \eta > 0 \) such a constant \( c \) exists. If we denote the best (smallest) possible \( c \) for a given \( \eta \) as \( c(\eta) \), the bound becomes

\[ L_{abs}(S, AA) \leq \eta c(\eta) \min_i L(S, E_i) + c(\eta) \ln n. \]

If we know an upper bound \( K \geq \min_i L(S, E_i) \), we can use that to choose \( \eta \) such that the above becomes

\[ L_{abs}(S, AA) - \min_i L(S, E_i) \leq a \sqrt{K \ln n} + b \ln n \]

for some constants \( a, b > 0 \).

2. (a) We assume there is a fixed but unknown probability distribution \( P \) over \( X \times Y \) where \( Y = \{0, 1\} \).

The true risk of a hypothesis \( h : X \to Y \) with the 0-1 loss is

\[ R(h) = E_{(x,y) \sim P}[L_{0-1}(y, h(x))]. \]

Given a sequence of \( m \) examples \( (x_i, y_i) \in X \times Y \), we define the empirical risk as

\[ \hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} L_{0-1}(y_i, h(x_i)). \]
Fix now $\varepsilon, \delta > 0$. If we draw the $m$ examples independently from distribution $P$ where
\[
m \geq \frac{1}{2\varepsilon} \ln \frac{2|H|}{\delta},
\]
then with probability at least $1 - \delta$ all hypotheses $h \in H$ satisfy
\[
\left| R(h) - \hat{R}(h) \right| \leq \varepsilon.
\]
(This is Theorem 1.11, page 54 of lecture notes.)

(b) The proof of Theorem 1.11 in on page 55 of lecture notes.

3. As suggested, let
\[
\xi'_i = \max \{-w \cdot x_i + y_i - \varepsilon, 0\}, \quad \xi''_i = \max \{w \cdot x_i - y_i - \varepsilon, 0\}.
\]
Using this notation, we can write the $\varepsilon$-insensitive loss as
\[
L_\varepsilon(w \cdot x_i - y_i) = \xi'_i + \xi''_i.
\]
Therefore the objective function becomes
\[
\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} (\xi'_i + \xi''_i).
\]
We treat $\xi'_i$ and $\xi''_i$ as new variables and use constraints to give them the intended meaning. We get the optimisation problem

**minimize** with respect to $w \in \mathbb{R}^d$, $\xi' \in \mathbb{R}^m$, $\xi'' \in \mathbb{R}^m$

\[
f(w, \xi', \xi'') = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} (\xi'_i + \xi''_i)
\]

subject to
\[
-\xi'_i - w \cdot x_i + y_i - \varepsilon \leq 0 \quad i = 1, \ldots, m
\]
\[
-\xi''_i + w \cdot x_i - y_i - \varepsilon \leq 0 \quad i = 1, \ldots, m
\]
\[
-\xi'_i \leq 0 \quad i = 1, \ldots, m
\]
\[
-\xi''_i \leq 0 \quad i = 1, \ldots, m.
\]
We get the Lagrangian

\[ L(w, \xi_i', \xi_i'', \alpha', \alpha'', \beta', \beta'') = \frac{1}{2} w \cdot w + C \sum_{i=1}^{m} (\xi_i' + \xi_i'') \]

\[ + \sum_{i=1}^{m} \alpha'_i (-\xi_i' - w \cdot x_i + y_i - \varepsilon) \]

\[ + \sum_{i=1}^{m} \alpha''_i (-\xi_i'' + w \cdot x_i - y_i - \varepsilon) \]

\[ - \sum_{i=1}^{m} \beta'_i \xi_i' - \sum_{i=1}^{m} \beta''_i \xi_i''. \]

Taking derivatives with respect to the primal variables, we get

\[ \frac{\partial L}{\partial w_j} = w_j - \sum_{i=1}^{m} (\alpha'_i - \alpha''_i) x_{ij} \]

\[ \frac{\partial L}{\partial \xi_i'} = C - \alpha'_i - \beta'_i \]

\[ \frac{\partial L}{\partial \xi_i''} = C - \alpha''_i - \beta''_i. \]

In particular, \( w = \sum_{i=1}^{m} (\alpha'_i - \alpha''_i) x_i \). Substituting this and \( (\partial L)/(\partial \xi_i') = (\partial L)/(\partial \xi_i'') = 0 \) back to \( L \), we get the dual function

\[ g(\alpha', \alpha'', \beta', \beta'') = -\frac{1}{2} \sum_{i,j=1}^{m} (\alpha'_i - \alpha''_i)(\alpha'_j - \alpha''_j) x_i \cdot x_j + \sum_{i=1}^{m} (\alpha'_i - \alpha''_i) y_i - \sum_{i=1}^{m} (\alpha'_i + \alpha''_i) \varepsilon \].

Since all \( \alpha \) and \( \beta \) variables are constrained to be non-negative, the conditions \( C - \alpha'_i - \beta'_i = C - \alpha''_i - \beta''_i \) imply \( 0 \leq \alpha'_i \leq C \) and \( 0 \leq \alpha''_i \leq C \). Conversely, if \( 0 \leq \alpha'_i \leq C \) and \( 0 \leq \alpha''_i \leq C \), it is always possible to choose \( \beta'_i \geq 0 \) and \( \beta''_i \geq 0 \) such that \( C - \alpha'_i - \beta'_i = C - \alpha''_i - \beta''_i \). Since \( \beta \) variables have otherwise no effect, we can write the dual problem as

maximise

\[ g_2(\alpha', \alpha'') = -\frac{1}{2} \sum_{i,j=1}^{m} (\alpha'_i - \alpha''_i)(\alpha'_j - \alpha''_j) x_i \cdot x_j \]

\[ + \sum_{i=1}^{m} (\alpha'_i - \alpha''_i) y_i - \sum_{i=1}^{m} (\alpha'_i + \alpha''_i) \varepsilon \]
subject to

\[ \begin{align*}
\alpha'_i & \leq C & i = 1, \ldots, m \\
\alpha''_i & \leq C & i = 1, \ldots, m \\
-\alpha'_i & \leq 0 & i = 1, \ldots, m \\
-\alpha''_i & \leq 0 & i = 1, \ldots, m.
\end{align*} \]

In the kernelised version we replace \( x_i \) by \( \psi(z_i) \) where \( \psi(z') \cdot \psi(z'') = k(z', z'') \).

Therefore we can just replace \( g_2 \) above with

\[ -\frac{1}{2} \sum_{i,j=1}^{m} (\alpha'_i - \alpha''_i)(\alpha'_j - \alpha''_j)k(z_i, z_j). \]

In the non-kernel version, we eventually obtain \( w \), which is interpreted to represent the linear mapping \( \mathbb{R}^d \to \mathbb{R}, x \mapsto w \cdot x \). In the kernel version, we obtain the mapping

\[ z \mapsto w \cdot \psi(z) = \sum_{i=1}^{m} (\alpha'_i - \alpha''_i)k(z_i, z). \]