1. (a) A boosting algorithm can be derived using coordinate descent by first taking the gradient of the generalized loss function.

$$\nabla_\alpha L(\alpha) = \sum_{i=1}^{m} -y_i \Phi'(-y_i \alpha^T h(x_i)) h(x_i)$$

Here h is a vector of weak classifiers h_t. The individual elements of the gradient can be expressed as

$$\begin{aligned}\nabla_\alpha^T L(\alpha) e_l &= -\sum_{i=1}^{m} y_i h_l(x_i) \Phi'(-y_i \alpha^T h(x_i)) \\ &= \sum_{i:y_i \neq h_l(x_i)} \Phi'(-y_i \alpha^T h(x_i)) \; - \sum_{i:y_i = h_l(x_i)} \Phi'(-y_i \alpha^T h(x_i))\end{aligned}$$

meaning the descent coordinate can be found by minimizing the last expression.

$$k = \arg\min_l \nabla_\alpha^T L(\alpha) e_l = \arg\min_l \sum_{i:y_i \neq h_l(x_i)} \Phi'(-y_i \alpha^T h(x_i))$$

To find the step size, η, we set the partial derivative of $L(\alpha)$ to zero and solve for η.

$$\begin{aligned}L(\alpha^{t-1} + \eta_t e_k) &= \sum_{i=1}^{m} \Phi(-y_i (\alpha^{t-1} + \eta e_k)^T h(x_i)) \\ \frac{\partial L}{\partial \eta} &= \sum_{i=1}^{m} -y_i h_k(x_i) \Phi(-y_i (\alpha^{t-1} + \eta e_k)^T h(x_i)) = 0\end{aligned}$$

(b) Function (1) is not continuous, function (2) is not strictly increasing as it's a quadratic, and function (3) is not differentiable at 0. The fourth function all of the properties consistent with generalized Φ.

$$\Phi_4(-u) = \log_2(1 + e^{-u}), \quad \frac{\partial \Phi_4}{\partial(-u)} = \frac{e^{-u}}{\ln(2)(1 + e^{-u})}, \quad \frac{\partial^2 \Phi_4}{\partial(-u)^2} = \frac{\ln(2) e^{-u}}{(\ln(2)(1 + e^{-u}))^2}$$

It's clear that Φ_4 is continuous and has strictly positive first and second derivatives. Also $\Phi_4(0) = 1$ and, given the function's other properties, this means that is greater than zero for negative arguments and greater than 1 for positive arguments.

(c) For $\Phi_4(-u) = \log_2(1 + e^{-u})$ the loss function is

$$L(\alpha) = \sum_{i=1} \log_2(1 + e^{-y_i \alpha^T h})$$

The descent coordinate can be found by taking

$$k = \arg\min_l \nabla_\alpha^T L(\alpha) e_l = \arg\min_l \sum_{i:y_i \neq h_l(x_i)} \frac{e^{-y_i \alpha^T h}}{\ln(2)(1 + e^{-y_i \alpha^T h})}.$$

The step size can be calculate by solving the following equation for η.

$$\frac{\partial L}{\partial \eta} = \sum_{i=1}^{m} -y_i h_k(x_i) \frac{e^{-y_i (\alpha^{t-1} + \eta e_k)^T h(x_i)}}{\ln(2)(1 + e^{-y_i (\alpha^{t-1} + \eta e_k)^T h(x_i)})} = 0$$

(d) Now take

$$\Phi(-u) = \begin{cases} e^{-u} & u \geq 0 \\ -u+1 & u < 0 \end{cases}$$

for which the loss function is

$$L(\alpha) = \sum_{i:y_i = h_l(x_i)} e^{-u} + \sum_{i:y_i \neq h_l(x_i)} -u+1$$

The descent coordinate can be found by taking

$$k = \arg\min_l \nabla_\alpha^T L(\alpha) e_l = \arg\min_l \sum_{i:y_i \neq h_l(x_i)} \mathbb{1}.$$

The step size can be calculate by solving the following equation for η.

$$\frac{\partial L}{\partial \eta} = \sum_{i:y_i = h_k(x_i)} -y_i h_k(x_i) e^{-y_i(\alpha^{t-1} + \eta e_k)^T h(x_i)} + \sum_{i:y_i \neq h_k(x_i)} -y_i h_k(x_i)$$

2. (a) We calculate the conditional entropy from splitting on x_1 and $x_{2,3}$ respectively (due to symmetry splitting on x_2 or x_3 give the same result).

$$H[Y|X_1] = -\frac{3}{4}\Big(\frac{2}{3}\log\big(\frac{2}{3}\big) + \frac{1}{3}\log\big(\frac{1}{3}\big)\Big) - \frac{1}{4}\log\big(0\big) = .47$$
$$H[Y|X_{2,3}] = -\frac{1}{2}\log\big(\frac{1}{2}\big) = .69$$

Splitting on x_1 thus maximizes the reduction in entropy. Again, due to symmetry, it doesn't matter if the second fork is determined with x_2 or x_3. Either way, two elements of the training set will always be labeled correctly and two will be ambiguous, since they will have different labels but end up in the same leaf. Whatever labeling is chosen for that leaf, one element of the training set will always be misclassified. Thus the training error of the resulting decision tree will always have an error of 1/4.

(b) One can construct a tree by splitting on x_2 and x_3, in either order, which has zero training error.

3. **Claim:** $R \leq 2R^*(1-R^*)$

Proof: First we must establish that given a series of i.i.d samples $\{x_i\} \subset \mathcal{X}$ and some fixed point $x \in \mathcal{X}$, x and its nearest neighbor x_π are similar. Specifically $\eta(x_{\pi_n}) \to \eta(x)$ as $n \to \infty$.

If we assume that $\mathcal{X}$ is a separable metric space, then we know that it has a countable dense subset $\mathcal{A} \subset \mathcal{X}$. We can assume that with probability one, $P(B_\epsilon(x)) > 0$ for any $x \in \mathcal{X}$ and $\epsilon > 0$. To confirm this, let $\tilde{\mathcal{X}}$ reprent all the points for which our assumption does not hold. For each $\tilde{x} \in \tilde{\mathcal{X}}$ there exists some ϵ such that $P(B_\epsilon(\tilde{x})) = 0$. But then $B_{\epsilon/2}(\tilde{x})$ must contain some element $a \in \mathcal{A}$ such that $\tilde{x} \in B_{\epsilon/2}(a) \subset B_\epsilon(x)$. The set of balls $B_{\epsilon/2}(a)$ generated this way form a countable covering of $\tilde{\mathcal{X}}$ having measure zero.

Now we can claim that $x_{\pi_n} \to x$ where $\{x_{\pi_n}\}$ is the sequence of nearest neighboors to x

generated by taking successive samples from $\mathcal{X}$. Otherwise, there exist an $\epsilon > 0$ such that $x_{\pi_n} \notin B_\epsilon(x)$ for all $n \in \mathbb{N}$. Let $P(B_\epsilon(x)) = \gamma)$. Then

$$P(x_{\pi_n} \notin B_\epsilon(x)) = P(x_0 \notin B_\epsilon(x))^n = (1 - P(B_\epsilon(x)))^n = (1-\gamma)^n \to 0.$$

The Lipschitz property of η subsequently implies

$$|\eta(x_{\pi_n}) - \eta(x)| \leq c\|x_{\pi_n} - x\| \to 0.$$

Moving on, we define the expected risk given n samples as

$$\begin{aligned} R_n(x) &= \mathbb{E}_{y,y_{\pi_n}}[L(y, y_{\pi_n})] = P[y \neq y_{\pi_n}|x, x_{\pi_n}] \\ &= P[y=1|x]P[y_{\pi_n} = 0|x_{\pi_n}] + P[y=0|x]P[y_{\pi_n}=1|x_{\pi_n}] \\ &= \eta(x)(1-\eta(x_{\pi_n})) + \eta(x_{\pi_n})(1-\eta(x)) \end{aligned}$$

The fact that $\eta(x_{\pi_n}) \to \eta(x)$ lets us write

$$R(x) = \lim_{n\to\infty} R_n(x) = 2\eta(x)(1-\eta(x))$$

Since the conditional Bayes risk for a given x is

$$R^*(x) = \mathbb{E}_y[L(y^*, y)] = 1 - p(y = y^*|x) = \min\{\eta(x), 1-\eta(x)\},$$

we can write

$$R(x) = 2R^*(x)(1-R^*(x))$$

giving us the following expression for the total risk of our classifier $n \to \infty$.

$$R = \lim_{n\to\infty} R_n = \lim_{n\to\infty} \mathbb{E}_x[R_n(x)]$$

Since $R_n(x) \to 2R_n^*(x)(1-R_n^*(x))$ from below at each point $x \in \mathcal{X}$ and thus $R_n(x)p(x) \to 2R_n^*(x)(1-R_n^*(x))p(x)$ from below, we can apply the dominated convergence theorem to the above equation and exchange limits. Thus

$$\begin{aligned} R &= \mathbb{E}_x[\lim_{n\to\infty} R_n(x)] \\ &= \mathbb{E}_x[2R^*(x)(1-R^*(x))] \\ &= 2\mathbb{E}_x[R^*(x)] - 2\mathbb{E}_x[R^*(x)] \\ &= 2\mathbb{E}_x[R^*(x)] - 2\mathbb{E}_x[R^*(x)^2] \\ &= 2\mathbb{E}_x[R^*(x)] - 2\mathbb{E}_x[R^*(x)]^2 - 2\sigma(R^*(x)) \\ &\leq 2R^*(1-R^*) \end{aligned}$$