In Lecture 11 we saw that in two dimensional feature spaces dyadic decision trees yielded a rate of convergence of $n^{-1/3}$, when the Bayes decision boundary was a one-dimensional curve and the marginal density of $X$ was bounded above. It turns out that a better rate of convergence is possible using a slightly different bounding procedure, which you will develop in this homework.

a. First notice that the difference between the true and empirical risk of a decision tree can be written as a sum of local errors at each leaf. If $f$ is a decision tree, then

$$R(f) - \hat{R}(f) = \sum_{A \in \text{leafs of } f} R(A) - \hat{R}(A)$$

where $R(A) = E[1_{\{X \in A, f(X) \neq Y\}}]$. Now we can bound the error at a leaf using Chernoff’s bound (Hoeffding’s inequality) as follows: for a single leaf $A$ with probability $\geq 1 - \delta$

$$R(A) - \hat{R}(A) \leq \sqrt{\log(1/\delta)/(2n)}$$

This version of Chernoff’s bound is sometimes referred to as the “additive form.” Another version exists called the “relative form.” That version states that if $S_n = \sum_{i=1}^n B_i$, where $B_i$ are i.i.d. Bernoulli random variables, then

$$P(S_n \leq (1 - \epsilon)E[S_n]) \leq e^{-E[S_n]\epsilon^2/2}$$

Use this form of Chernoff’s bound to obtain another upper bound on $R(A) - \hat{R}(A)$. This upper bound, which should also hold with probability at least $1 - \delta$, will depend on $R(A)$.

b. Now we need a bound that holds for all leaves simultaneously (i.e., we need to take a union bound over leaves). Devise a prefix encoding scheme for each leaf of a dyadic tree. The goal is to obtain a unique code for every possible dyadic leaf (each leaf corresponds to a dyadic rectangle of a certain size and at a certain location; you need to encode each rectangle of this form). A prefix code can be devised so that the codelength for each leaf is proportional to the depth of the leaf in a dyadic tree. Use this code and the relative form of Chernoff’s bound above to obtain a bound of the form: with probability at least $1 - \delta$

$$R(f) - \hat{R}(f) \leq \sum_{A \in \text{leafs of } f} C(A, \delta, R(A)) \ , \ \forall f \in \mathcal{F}^T$$

where $C(A, \delta, R(A))$ depends on the leaf $A$, $\delta$ and the risk $R(A)$. The bound holds for all dyadic decision trees, since our union bound holds for all possible dyadic leaves.

c. In order for the upper bound derived in (b) to be of use, we must eliminate the dependence on $R(A)$, which of course is unknown. In our analysis of risk bounds for histograms and trees in Lecture 11, we assumed that the density of $X$ was bounded from above like $p(x) \leq C$, for a constant $C > 0$. Using this assumption, derive an upper bound for $R(A)$ that only depends on the size of $A$ and the constant $C$. Use this upper bound to derive another upper bound on $R(f) - \hat{R}(f)$ that can be computed without additional assumptions on the underlying distribution $P_{XY}$. Hint: The bound should now have the form: with probability at least $1 - \delta$

$$R(f) - \hat{R}(f) \leq \text{constant} \times \sum_{A \in \text{leafs of } f} \sqrt{(j(A)2^{-j(A)} + \log(1/\delta))/2n} \ , \ \forall f \in \mathcal{F}^T$$

where $j(A)$ is the depth of leaf $A$ in a dyadic tree.
d. Now we are in a position to obtain a risk bound for dyadic decision trees constructed using a complexity regularization procedure based on the bound obtained in (c). Let \( \hat{f}_n \) denote the minimizer of such a complexity regularization, and upper bound the excess risk in terms of the approximation error and a bound on the estimation error. The estimation error bound should be proportional to

\[
\sum_{A \in \text{leafs of } f} \sqrt{\left(j(A)2^{-j(A)} + \log(n)\right)/2n}
\]

Let \( f_k^* \) denote the optimal dyadic decision tree with \( k \) (square) leaves intersecting the one-dimensional Bayes decision boundary (i.e., the tree obtained by beginning with \( k^2 \) cell histogram partition and pruning back all cells that do not intersect with the Bayes decision boundary), and use this tree to show that the approximation error is \( O(1/k) \). To bound the estimation error, show that \( f_k^* \) has \( O(2^{j/2}) \) leaves at depth \( j \).

e. Note that the maximum depth of \( f_k^* \) is \( O(\log n) \). From here, conclude that the excess risk is \( O(\sqrt{\log n}/n) \), which shows that the rate of convergence, ignoring logarithmic factor, is \( n^{-1/2} \).

f. Note that the penalty in this construction is additive (sum over leaves), and consequently \( \hat{f}_n \) can be obtained by starting with a histogram of very small cells (e.g., \( O(\sqrt{n}) \) cells) and pruning from the bottom-up. Briefly outline the pruning algorithm and justify why it leads to the optimal tree.