The problem of density estimation is posed as follows. We obtain an i.i.d. sample $X_1, \ldots, X_n$ of $\mathbb{R}^d$-valued random variables whose common distribution is unknown. Assuming that it has a well-defined probability density function (pdf) $f$, we would like to estimate it. We will measure the accuracy of an estimate $\hat{f}$ by the Kullback–Leibler divergence

$$D(f \| \hat{f}) = \int_{\mathbb{R}^d} f(x) \log \frac{f(x)}{\hat{f}(x)} \, dx.$$ 

A popular strategy for density estimation is by maximum likelihood (ML). That is, if we know that $f$ comes from some parametric class $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$, then we let $\hat{f}_n = h_{\hat{\theta}_n}$, where $\hat{\theta}_n = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log \frac{1}{h_\theta(X_i)}$.

However, $\mathcal{H}$ may not be rich enough to accurately approximate the unknown $f$. For this reason, we may consider more complicated classes of densities. For example, given a positive integer $k$, let $\mathcal{H}_k$ denote the class of all $k$-component mixtures over $\mathcal{H}$:

$$\mathcal{H}_k = \left\{ h = \sum_{j=1}^k c_j h_{\theta_j} : c_1, \ldots, c_k \geq 0, \sum_{j=1}^k c_j = 1; \theta_1, \ldots, \theta_k \in \Theta \right\}.$$ 

Now consider performing ML estimation over $\mathcal{H}_k$:

$$\hat{f}_n = \arg \min_{h \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^n \log \frac{1}{h(X_i)}.$$ 

In this homework assignment, you will be asked to prove the following result:

**Theorem 1.** Suppose that both the unknown density $f$ and the base densities $h_\theta$, $\theta \in \Theta$, are bounded between some strictly positive constants $0 < c_- < c_+ < \infty$:

$$c_- \leq f(x) \leq c_+; \quad c_- \leq h_\theta(x) \leq c_+, \forall \theta \in \Theta.$$ 

Then

$$D(f \| \hat{f}_n) \leq D(f \| f_k) + \frac{4c_+}{(c_-)^2} \mathbb{E} R_n(\mathcal{H}(X^n)) + 2 \log \left( \frac{c_+}{c_-} \right) \sqrt{\frac{2 \log(1/\delta)}{n}}$$ 

with probability at least $1 - \delta$, where $f_k = \arg \min_{h \in \mathcal{H}_k} D(f \| h)$.

You will prove this theorem in several steps.
(1) **Uniform deviation bound.** Prove that

\[ D(f\|\hat{f}_n) - D(f\|f_k) \leq 2 \sup_{h \in \mathcal{H}_k} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(X_i)}{f(X_i)} - \mathbb{E} \left[ \log \frac{h(X)}{f(X)} \right] \right|. \]

(2) **From uniform deviations to Rademacher averages.** Let

\[ \Delta_n(X^n) \triangleq \sup_{h \in \mathcal{H}_k} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(X_i)}{f(X_i)} - \mathbb{E} \left[ \log \frac{h(X)}{f(X)} \right] \right|. \]

Let \( \mathcal{L}_f \) be the class of all functions of the form \( \log[h(x)/f(x)], h \in \mathcal{H}_k \). Prove that

\[ \Delta_n(X^n) \leq 2\mathbb{E}R_n(\mathcal{L}_f(X^n)) + \log \left( \frac{c_+}{c_-} \right) \sqrt{\frac{2 \log(1/\delta)}{n}} \]

with probability at least \( 1 - \delta \).

(3) Prove that

\[ R_n(\mathcal{L}_f(X^n)) \leq \frac{c_+}{(c_-)^2} R_n(\mathcal{H}(X^n)). \]

**Hint:** You will need the following generalization of the contraction principle: Let \( \mathcal{A} \subset \mathbb{R}^n \) be a bounded set. Let \( \phi_1, \ldots, \phi_n : \mathbb{R} \to \mathbb{R} \) be \( n \) functions, such that there exists some positive constant \( L \) so that for any two distinct \( a = (a_1, \ldots, a_n), a' = (a'_1, \ldots, a'_n) \in \mathcal{A} \)

\[ \max_{1 \leq i \leq n} \frac{|\phi_i(a_i) - \phi_i(a'_i)|}{|a_i - a'_i|} \leq L. \]

Define the set

\[ \phi \circ \mathcal{A} \triangleq \{ (\phi_1(a_1), \ldots, \phi_n(a_n)) : a = (a_1, \ldots, a_n) \in \mathcal{A} \}. \]

Then \( R_n(\phi \circ \mathcal{A}) \leq LR_n(\mathcal{A}) \).

(4) Combine the results of the previous problems to finish the proof.