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

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

A popular strategy for density estimation is by maximum likelihood (ML). That is, if we know that $f$ comes from some parametric class $H = \{h_\theta : \theta \in \Theta\}$, 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, $H$ 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 $H_k$ denote the class of all $k$-component mixtures over $H$:

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

Now consider performing ML estimation over $H_k$:

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

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

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

with probability at least $1 - \delta$, where

$$f_k = \arg \min_{h \in H_k} D(f \parallel h).$$

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

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

**Solution.** Note that for any \( h \in \mathcal{H}_k \)

\[
D(f\|h) = \mathbb{E} \left[ \log \frac{f(X)}{h(X)} \right] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(X_i)}{h(X_i)} \right].
\]

Let \( \hat{f}_n \) be the empirical risk minimizer. Then

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

and therefore

\[
\frac{1}{n} \sum_{i=1}^{n} \log \frac{f(X_i)}{\hat{f}_n(X_i)} \leq \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(X_i)}{f_k(X_i)}.
\]

Then we can write

\[
D(f\|\hat{f}_n) - D(f\|f_k) = D(f\|\hat{f}_n) - \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(X_i)}{\hat{f}_n(X_i)} + \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(X_i)}{\hat{f}_n(X_i)} - \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(X_i)}{f_k(X_i)} + \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(X_i)}{f_k(X_i)} - D(f\|f_k)
\]

Because of (1), \( T_2 \leq 0 \), while

\[
T_1 + T_3 \leq 2 \sup_{h \in \mathcal{H}_k} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(X_i)}{f(X_i)} - \mathbb{E} \left[ \log \frac{h(X)}{f(X)} \right] \right|
\]

Moreover, because \( \log(1/u) = -\log u \), we can write

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(X_i)}{f(X_i)} - \mathbb{E} \left[ \log \frac{h(X)}{f(X)} \right] \right|
\]

for any \( h \). Putting everything together, we get the desired bound.

(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 \( 1 - \delta \).

**Solution.** First, we will show that \( \Delta_n(X^n) \) has bounded differences. Choose an arbitrary \( n \)-tuple \( x_1, \ldots, x_n \) and an arbitrary \( x_i' \); let \( x_i'' \) denote \( x_i \) with \( x_i \) replaced by \( x_i' \). Then, because both
f and all h are bounded between $c_-$ and $c_+$, we get
\[
\left| \Delta_n(x^n) - \Delta_n(x^n_{(i)}) \right| \\
\leq \frac{1}{n} \sup_{h \in H_k} \left| \log \frac{h(x_i)}{f(x_i)} - \log \frac{h(x'_i)}{f(x_i)} \right| \\
\leq \frac{2}{n} \log \frac{c_+}{c_-}.
\]
This holds for all $i$ and all $x^n$ and $x'_i$, so $\Delta_n(X^n)$ has bounded differences with $c_1 = \ldots = c_n = (2/n) \log(c_+/c_-)$. Therefore, McDiarmid’s inequality says that for any $t > 0$
\[
P(\Delta_n(X^n) \geq \mathbb{E}\Delta_n(X^n) + t) \leq \exp \left( -\frac{nt^2}{2\log^2(c_+/c_-)} \right).
\]
Choosing $t = \log(c_+/c_-) \sqrt{\frac{2\log(1/\delta)}{n}}$ and using the fact that $\mathbb{E}\Delta_n(X^n) \leq 2\mathbb{E}R_n(\mathcal{L}_f(X^n))$, we get the desired result.

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

**Hint:** Use 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 $a = (a_1, \ldots, a_n), a' = (a'_1, \ldots, a'_n) \in \mathcal{A}$
\[
\max_{1 \leq i \leq n} \frac{\left| \phi_i(a_i) - \phi_i(a'_i) \right|}{|a_i - a'_i|} \leq L.
\]
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 $R_n(\phi \circ \mathcal{A}) \leq LR_n(\mathcal{A})$.

**Solution.** The idea is to use the contraction principle twice. First, let us consider the function $\phi(u) = \log(u)$. Its derivative on the interval $[c_-/c_+, c_+/c_-]$ is bounded in magnitude by $c_+/c_-$, so
\[
|\phi(u) - \phi(v)| \leq \frac{c_+}{c_-} |u - v|, \quad \forall u, v \in \left[ \frac{c_-}{c_+}, \frac{c_+}{c_-} \right].
\]
Hence, $\phi$ is $(c_+/c_-)$-Lipschitz on the interval $[c_-/c_+, c_+/c_-]$. Therefore, letting $\mathcal{M}_f$ denote the class of all functions of the form $h(x)/f(x)$, $h \in \mathcal{H}_k$, we have by the contraction principle that
\[
R_n(\mathcal{L}_f(X^n)) \leq \frac{c_+}{c_-} R_n(\mathcal{M}_f(X^n)).
\]
Now consider the functions $\phi_i(u) = \frac{1}{f_i(X_i)} u$, $i = 1, \ldots, n$. Then
\[
|\phi_i(u) - \phi_i(v)| = \frac{1}{f_i(X_i)} |u - v| \leq \frac{1}{c_-} |u - v|.
\]
It is easy to see that $\mathcal{M}_f(X^n) = \phi \circ \mathcal{H}_k(X^n)$, and by the generalized contraction principle we have
\[
R_n(\mathcal{M}_f(X^n)) \leq \frac{1}{c_-} R_n(\mathcal{H}_k(X^n)).
\]
Therefore,
\[
R_n(\mathcal{L}_f(X^n)) \leq \frac{c_+}{(c_-)^2} R_n(\mathcal{H}_k(X^n)).
\]
Finally, since $\mathcal{H}_k(X^n) \subset \text{conv}(\mathcal{H}(X^n))$, we have
\[
R_n(\mathcal{H}_k(X^n)) \leq R_n(\text{conv}(\mathcal{H}(X^n))) = R_n(\mathcal{H}(X^n)).
\]

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