**RISK BOUNDS FOR MIXTURE DENSITY ESTIMATION**

**ALEXANDER RAKHLIN**\(^1\) , **DMITRY PANCHENKO**\(^2\) and **SAYAN MUKHERJEE**\(^3\)

**Abstract.** In this paper we focus on the problem of estimating a bounded density using a finite combination of densities from a given class. We consider the Maximum Likelihood Estimator (MLE) and the greedy procedure described by Li and Barron (1999) under the additional assumption of boundedness of densities. We prove an \(O(\sqrt{n})\) bound on the estimation error which does not depend on the number of densities in the estimated combination. Under the boundedness assumption, this improves the bound of Li and Barron by removing the \(\log n\) factor and also generalizes it to the base classes with converging Dudley integral.

**Mathematics Subject Classification.** 62G05, 62G07, 62G20.

Received July 21, 2004.

1. **Introduction**

In the density estimation problem, we are given i.i.d. sample \(S = (x_1, \ldots, x_n)\) drawn from an unknown density \(f\). The goal is to estimate this density from the given data. We consider the Maximum Likelihood Procedure (MLE) and the greedy procedure described by Li and Barron [7,8] and prove estimation bounds for these procedures. Rates of convergence for density estimation were studied in [3,12,13,15]. For neural networks and projection pursuit, approximation and estimation bounds can be found in \([1,2,5,11]\).

**Keywords and phrases.** Mixture density estimation, maximum likelihood, Rademacher processes.

* This report describes research done at the Center for Biological & Computational Learning, which is in the Department of Brain & Cognitive Sciences at MIT and which is affiliated with the McGovern Institute of Brain Research and with the Artificial Intelligence Laboratory. This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. N00014-00-1-0907, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, and National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506. Additional support was provided by: AT&T, Central Research Institute of Electric Power Industry, Center for e-Business (MIT), DaimlerChrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., The Eugene McDermott Foundation, Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research, Inc., Sony MOUT, Sumitomo Metal Industries, Toyota Motor Corporation, WatchVision Co., Ltd, and the SLOAN Foundation.

1 Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; rakhlin@mit.edu
2 Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02143, USA.
3 Institute of Statistics and Decision Sciences, Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA.

© EDP Sciences, SMAI 2005

Article published by EDP Sciences and available at http://www.edpsciences.org/ps or http://dx.doi.org/10.1051/ps:2005011
Let \((X, F)\) be a measurable space and let \(\lambda\) be a \(\sigma\)-finite measure on \(F\). Whenever we mention below that a probability measure on \(F\) has a density we will understand that it has a Radon-Nikodym derivative with respect to \(\lambda\).

To evaluate the accuracy of the density estimate we need a notion of distance. Kullback-Leibler (KL) divergence and Hellinger distance are the most commonly used. In this paper we will work with the KL-divergence, defined for two distributions \(f\) and \(g\) as

\[
D(f \parallel g) = \int f(x) \log \frac{f(x)}{g(x)} \, d\lambda(x) = \mathbb{E}_x \log \frac{f(x)}{g(x)}
\]

Here \(x\) has distribution with density \(f\).

Consider a parametric family of probability density functions \(\mathcal{H} = \{\phi_\theta(x) : \theta \in \Theta \subset \mathbb{R}^d\}\) over \(X\). The class of \(k\)-component mixtures \(f_k\) is defined as

\[
\mathcal{C}_k = \text{conv}_k(\mathcal{H}) = \left\{ f : f(x) = \sum_{i=1}^k \lambda_i \phi_{\theta_i}(x), \sum_{i=1}^k \lambda_i = 1, \lambda_i \geq 0, \theta_i \in \Theta \right\}.
\]

Let us define the class of continuous convex combinations

\[
\mathcal{C} = \text{conv}(\mathcal{H}) = \left\{ f : f(x) = \int_\Theta \phi_\theta(x) P(d\theta), \ P \text{ is a probability measure on } \Theta \right\}.
\]

The approximation bound of Li and Barron \([7,8]\) states that for any \(f\), there exists an \(f_k \in \mathcal{C}_k\), such that

\[
D(f \parallel f_k) \leq D(f \parallel \mathcal{C}) + \frac{c_f^2 P \gamma}{k}, \tag{1}
\]

where \(c_f P\) and \(\gamma\) are constants and \(D(f \parallel \mathcal{C}) = \inf_{g \in \mathcal{C}} D(f \parallel g)\). Furthermore, \(\gamma\) is an upper bound on the log-ratio of any two functions \(\phi_\theta(x), \phi_{\theta'}(x)\) for all \(\theta, \theta', x\) and therefore

\[
\sup_{\theta, \theta', x} \log \frac{\phi_\theta(x)}{\phi_{\theta'}(x)} < \infty \tag{2}
\]

is a condition on the class \(\mathcal{H}\).

Li and Barron prove that \(k\)-mixture approximations satisfying (1) can be constructed by the following greedy procedure: Initialize \(f_1 = \phi_\theta\) to minimize \(D(f \parallel f_1)\) and at step \(k\) construct \(f_k\) from \(f_{k-1}\) by finding \(\alpha\) and \(\theta\) such that

\[
D(f \parallel f_k) \leq \min_{\alpha, \theta} D(f \parallel ((1 - \alpha)f_{k-1}(x) + \alpha \phi_\theta(x))).
\]
Furthermore, a connection between KL-divergence and Maximum Likelihood suggests the following method to estimate the density $f_k$ from the data by greedily choosing $\theta_\alpha$ at step $k$ so that

$$\sum_{i=1}^{n} \log \hat{f}_k(x_i) \geq \max_{\alpha,\theta} \sum_{i=1}^{n} \log[(1 - \alpha)\hat{f}_{k-1}(x_i) + \alpha \phi_\theta(x_i)].$$

(3)

Li and Barron proved the following theorem:

**Theorem 2.1.** Let $\hat{f}_k(x)$ be either the maximizer of the likelihood over $k$-component mixtures or more generally any sequence of density estimates satisfying (3). Assume additionally that $\Theta$ is a $d$-dimensional cube with side-length $A$, and that

$$\sup_{x \in \mathcal{X}} |\log \phi_\theta(x) - \log \phi_{\theta'}(x)| \leq B \sum_{j} |\theta_j - \theta'_j|$$

for any $\theta, \theta' \in \Theta$. Then

$$E \left[ D(f||\hat{f}_k) \right] - D(f||C) \leq \frac{c_1}{k} + \frac{c_2 k}{n} \log(nc_3),$$

where $c_1, c_2, c_3$ are constants (dependent on $A, B, d$).

The above bound combines the approximation and estimation results. Note that the first term decreases with the number of components $k$, while the second term increases. The rate of convergence for the optimal $k$ is therefore $O(\sqrt{\log n}/n)$.

2. **Main results**

We assume that $f$ and the densities in $\mathcal{H}$ are bounded above and below by some constants $a$ and $b$, respectively. This boundedness naturally extends to the convex combinations as well. We prove the following results:

**Theorem 2.1.** For any target density $f$ such that $a \leq f(x) \leq b$ for all $x \in \mathcal{X}$ and $\hat{f}_k(x)$ being either the maximizer of the likelihood over $k$-component mixtures or more generally any sequence of density estimates satisfying (3),

$$E \left[ D(f||\hat{f}_k) \right] - D(f||C) \leq \frac{c_1}{k} + \frac{c_2}{\sqrt{n}} \int_{a}^{b} \log^{1/2} D(\mathcal{H}, \epsilon, d_n) d\epsilon,$$

where $c_1, c_2$ are constants (dependent on $a, b$) and $D(\mathcal{H}, \epsilon, d_n)$ is the $\epsilon$-covering number of $\mathcal{H}$ with respect to empirical distance $d_n (d_n^2(\phi_1, \phi_2) = \frac{1}{n} \sum_{i=1}^{n} (\phi_1(x_i) - \phi_2(x_i))^2)$.

**Corollary 2.2.** Under the conditions of Theorem 1.1 (i.e. $\mathcal{H}$ satisfying condition (4) and $\Theta$ being a cube with side-length $A$) and assuming boundedness of the densities, the bound of Theorem 2.1 becomes

$$E \left[ D(f||\hat{f}_k) \right] - D(f||C) \leq \frac{c_1}{k} + \frac{c_2}{\sqrt{n}},$$

where $c_1$ and $c_2$ are constants (dependent on $a, b, A, B, d$).

**Corollary 2.3.** The bound of Corollary 2.2 holds for the class of (truncated) Gaussian densities $\mathcal{H} = \{f_{\mu, \sigma} : f_{\mu, \sigma}(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left(-\frac{(x-\mu)^2}{2\sigma^2}\right), |\mu| \leq M, \sigma_{\min} \leq \sigma \leq \sigma_{\max}\}$ over a compact domain $\mathcal{X}$ ($Z_{\mu, \sigma}$ is needed for normalization).
Remark 2.4. Theorem 2.1 hides the dependence of constants $c_1, c_2$ on $a$ and $b$ for the sake of easy comparison to Theorem 1.1. We now state the result with explicit dependence on $a$ and $b$:

$$D(f\|\hat{f}_k) - D(f\|C) \leq \frac{1}{k} \frac{8b^2}{a^2} \left(2 + \log \frac{b}{a}\right) + \frac{1}{\sqrt{n}} \left(\frac{b}{a^2} \mathbb{E} \left[ c_1 \int_0^b \log^{1/2} D(\mathcal{H}, \epsilon, d_n) d\epsilon \right] + \frac{8b}{a}\right) + \sqrt{\frac{t}{n}} \left(4\sqrt{2} \log \frac{b}{a}\right)$$

with probability at least $1 - e^{-t}$, or, by integrating,

$$\mathbb{E} \left[ D(f\|\hat{f}_k) \right] - D(f\|C) \leq \frac{1}{k} \frac{8b^2}{a^2} \left(2 + \log \frac{b}{a}\right) + \frac{1}{\sqrt{n}} \left(\frac{b}{a^2} \mathbb{E} \left[ c_1 \int_0^b \log^{1/2} D(\mathcal{H}, \epsilon, d_n) d\epsilon \right] + \frac{8b}{a}\right) + 4\sqrt{2} \log \frac{b}{a},$$

where $c_1$ is an absolute constant.

Remark 2.5. Upper and lower bounds $a$ and $b$ are determined by the class $\mathcal{H}$ and by the target density $f$. Assume there exists a sequence of truncations $\{f_i\}$ of $f$, such that $a_i \leq f_i(x) \leq b_i$ for all $x \in \mathcal{X}$, and $\{a_i\}$ is decreasing and $\{b_i\}$ increasing. Further assume that each class $\mathcal{H}_i$ contains functions bounded by $a_i$ and $b_i$. As the number of samples $n$ grows, one can choose more and more complex models $\mathcal{H}_i$. If $a_i$ is a decreasing function of $n$ and $b_i$ is an increasing function of $n$, Remark 2.4 provides the rate for learning $f_i$, the truncated version of $f$. This could be applied, for instance, to a sequence of classes $\mathcal{H}_i$ of Gaussian densities over increasing domain and increasing range of variances.

3. DISCUSSION OF THE RESULTS

The result of Theorem 2.1 is twofold. The first implication concerns dependence of the bound on $k$, the number of components. Our results show that there is an estimation bound of the order $O(\frac{1}{k})$ that does not depend on $k$. Therefore, the number of components is not a trade-off that has to be made with the approximation part (which decreases with $k$). The bound also suggests that the number of components $k$ should be chosen to be $O(\sqrt{n})$.

The second implication concerns the rate of convergence in terms of $n$, the number of samples. The rate of convergence (in the sense of KL-divergence) of the estimated mixture to the true density is of the order $O(1/\sqrt{n})$. As Corollary 2.2 shows, for the specific class $\mathcal{H}$ considered by Li and Barron, the Dudley integral converges and does not depend on $n$. We therefore improve the results of Li and Barron by removing the $\log n$ factor. Furthermore, the result of this paper holds for general base classes $\mathcal{H}$ with a converging entropy integral, extending the result of Li and Barron. Note that the bound of Theorem 2.1 is in terms of the metric entropy of $\mathcal{H}$, as opposed to the metric entropy of $C$. This is a strong result because the convex class $C$ can be very large [10] even for small $\mathcal{H}$.

Rates of convergence for the MLE in mixture models were studied by Sara van de Geer [12]. As the author notes, the optimality of the rates depends primarily on the optimality of the entropy calculations. Unfortunately, in the results of [12], the entropy of the convex class appears in the bounds, which is undesirable. An advantage of the approach of [12] is the use of Hellinger distance to avoid problems near zero. Li and Barron address this problem by requiring (2), which is boundedness of the log of the ratio of two densities. Birgé and Massart [3, p. 122] cite a counterexample of Bahadur (1958) which shows that even with a compact parameter space, M.L.E. can diverge when likelihood ratios are unbounded. Unfortunately, boundedness of the ratios of densities is not enough for the proofs of this paper. We assume boundedness of the densities themselves. This is critical in one step of the proof, when the contraction principle is used (for the second time). Although the boundedness condition seems as a somewhat strict requirement, note that a class of densities that satisfies (2), but not boundedness of the densities, has to contain functions which all go to zero (or infinity) in exactly the same manner. Also note that on a non-compact domain $\mathbb{R}$ even a simple class of Gaussian densities does not satisfy (2). Indeed, the log-ratio of the tails of two Gaussians with the same variance but different means...
becomes infinite. If one considers a compact domain \( \mathcal{X} \), the boundedness of densities assumption does not seem very restrictive.

The proof technique of this paper seems to be a powerful general method for bounding uniform deviations of empirical and expected quantities. The main ingredients of the proof are the Comparison inequality for Rademacher processes and the fact that Rademacher averages (as defined in Lem. A.2) of the convex hull are equivalent to those of the base class.

4. Proofs

Assume

\[ 0 < a \leq \phi_b(x) \leq b \quad \forall x \in \mathcal{X}, \forall \phi_b \in \mathcal{H}. \]

Constants which depend only on \( a \) and \( b \) will be denoted by \( c \) with various subscripts. The values of the constants might change from line to line.

**Theorem 4.1.** For any fixed density \( f \) such that \( 0 < a \leq f(x) \leq b \quad \forall x \in \mathcal{X} \) and \( S = (x_1, \ldots, x_n) \) drawn i.i.d. from \( f \), with probability at least \( 1 - e^{-t} \),

\[
\sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \leq \mathbb{E} \left[ \frac{c_1}{\sqrt{n}} \int_{0}^{b} \log^{1/2} D(\mathcal{H}, \epsilon, d_\text{e}) \, d\epsilon \right] + c_2 \sqrt{\frac{t}{n}},
\]

where \( c_1 \) and \( c_2 \) are constants that depend on \( a \) and \( b \).

**Proof.** First, we apply Lemma A.3 to the random variable \( Z(x_1, \ldots, x_n) = \sup_{h \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \).

Let \( t_i = \log \frac{h(x_i)}{f(x_i)} \) and \( t'_i = \log \frac{h(x'_i)}{f(x'_i)} \). The bound on the martingale difference follows:

\[
|Z(x_1, \ldots, x'_1, \ldots, x_n) - Z(x_1, \ldots, x_i, \ldots, x_n)| = \left| \sup_{h \in \mathcal{F}} \left| \mathbb{E} \log \frac{h}{f} - \frac{1}{n} \left( t_1 + \ldots + t_i + \ldots + t_n \right) \right| \right|
\]

\[
- \sup_{h \in \mathcal{F}} \left| \mathbb{E} \log \frac{h}{f} - \frac{1}{n} \left( t_1 + \ldots + t'_i + \ldots + t_n \right) \right| \right|
\]

\[
\leq \sup_{h \in \mathcal{F}} \frac{1}{n} \left| \log \frac{h(x'_i)}{f(x'_i)} - \log \frac{h(x_i)}{f(x_i)} \right| \leq \frac{1}{n} \left( \log \frac{b}{a} - \log \frac{b}{a} \right)
\]

\[
= \frac{1}{2} \log \frac{b}{a} = c_i.
\]

The above chain of inequalities holds because of triangle inequality and properties of sup. Applying McDiarmid’s inequality (see Lem. A.3),

\[
\mathbb{P}(Z - \mathbb{E}Z > u) \leq \exp \left( -\frac{u^2}{2 \sum c_i^2} \right) \leq \exp \left( -\frac{nu^2}{2 \sqrt{2} \log \frac{b}{a}} \right).
\]

Therefore,

\[
\sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \leq \mathbb{E} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| + 2\sqrt{2} \log \frac{b}{a} \sqrt{\frac{t}{n}}
\]

with probability at least \( 1 - e^{-t} \) and by Lemma A.2,

\[
\mathbb{E} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \leq 2\mathbb{E} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \log \frac{h(x_i)}{f(x_i)} \right|.
\]
Combining,
\[ \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \leq 2 \mathbb{E} \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \log \frac{h(x_i)}{f(x_i)} \right| + 2 \sqrt{2 \log \frac{b}{a} \sqrt{\frac{T}{n}}} \]
with probability at least \(1 - e^{-t}\).

Therefore, instead of bounding the difference between the “empirical” and the “expectation”, it is enough to bound the above expectation of the Rademacher average. This is a simpler task, but first we have to deal with the log and the fraction (over \(f\)) in the Rademacher sum. To eliminate these difficulties, we apply Lemma A.1 twice. Once we reduce our problem to bounding the Rademacher sum of the basis functions \(\sup_{\phi \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi(x_i) \right|\), we will be able to use the entropy of the class \(\mathcal{H}\).

Let \(p_i = \frac{h(x_i)}{f(x_i)} - 1\) and note that \(\frac{a}{b} - 1 \leq p_i \leq \frac{b}{a} - 1\). Consider \(\phi(p) = \log(1 + p)\). The largest derivative of \(\log(1 + p)\) on the interval \(p \in [\frac{a}{b} - 1, \frac{b}{a} - 1]\) is at \(p = a/b - 1\) and is equal to \(b/a\). So, \(\frac{a}{b} \log(p + 1)\) is \(1\)-Lipschitz. Also, \(\phi(0) = 0\). By Lemma A.1 applied to \(\phi(p)\),
\[ 2 \mathbb{E} \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \log \frac{h(x_i)}{f(x_i)} \right| = \mathbb{E} \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi(p_i) \right| \]
\[ \leq \frac{4b}{a} \mathbb{E} \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \frac{h(x_i)}{f(x_i)} \right| \leq 4 \sqrt{2 \log \frac{b}{a} \sqrt{\frac{T}{n}}} \]

The last inequality holds because
\[ \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \right| \leq \left( \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \right)^2 \right)^{1/2} = \frac{1}{\sqrt{n}} \]
Let \(h_i = h(x_i)\), \(f_i = f(x_i)\). We apply Lemma A.1 again with the contraction \(\phi_i(h_i) = a \frac{h_i}{f_i}\). Note that \(|\phi_i(h_i) - \phi_i(g_i)| = \frac{a}{f_i} |h_i - g_i| \leq |h_i - g_i|\). Therefore,
\[ \frac{4b}{a} \mathbb{E} \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \frac{h(x_i)}{f(x_i)} \right| \leq 8 \frac{b}{a^2} \mathbb{E} \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(x_i) \right| \]
Combining the inequalities, with probability at least \(1 - e^{-t}\)
\[ \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \leq \frac{8b}{a^2} \mathbb{E} \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(x_i) \right| + \sqrt{8 \log \frac{b}{a} \sqrt{\frac{T}{n}} + \frac{4b}{a} \frac{1}{\sqrt{n}}} \]
The power of using Rademacher averages to estimate complexity comes from the fact that the Rademacher averages of a class are equal to those of the convex hull. Indeed, consider \(\sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(x_i) \right|\) with \(h(x) = \int_{\mathcal{H}} \phi \phi(x) \text{d}\theta\). Since a linear functional of convex combinations achieves its maximum value at the vertices, the above supremum is equal to \(\sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi_i(x_i) \right|\), the corresponding supremum on the basis functions \(\phi\). Therefore, \(\mathbb{E} \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(x_i) \right| = \mathbb{E} \sup_{\phi \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi_i(x_i) \right|\).
Next, we use the following classical result (see [14]),
\[ E_\epsilon \sup_{\phi \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi(x_i) \right| \leq \frac{c_1}{\sqrt{n}} \int_{0}^{b} \log^{1/2} D(\mathcal{H}, \epsilon, d_n) \, d\epsilon, \]
where \( d_n \) is the empirical distance with respect to the set \( S \).

Combining the results together, the following holds with probability at least \( 1 - e^{-t} \):
\[ \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - E \log \frac{h}{f} \right| \leq \left[ \frac{c_1}{\sqrt{n}} \int_{0}^{b} \log^{1/2} D(\mathcal{H}, \epsilon, d_n) \, d\epsilon \right] + c_2 \sqrt{\frac{T}{n}}. \]

**Remark 4.2.** If \( \mathcal{H} \) is a VC-subgraph with VC dimension \( V \), the Dudley integral above is bounded by \( c \sqrt{V} \) and we get \( O(1/\sqrt{n}) \) convergence. One example of such a class is the class of (truncated) Gaussian densities over a compact domain and with bounded variance (see Cor. 2.3). Another example is the class considered in [7], and its cover is computed in the proof of Corollary 2.2. More information on the classes with converging Dudley integral and examples of VC-subgraph classes can be found in [4, 14].

We are now ready to prove Theorem 2.1:

**Proof.**
\[ D(f \| \hat{f}_k) - D(f \| f_k) = \left( E \log \frac{f}{f_k} - \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i)}{f_k(x_i)} \right) + \left( \frac{1}{n} \sum_{i=1}^{n} \log f(x_i) - E \log \frac{f}{f_k} \right) \]
\[ + \left( \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i)}{f_k(x_i)} - \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i)}{f_k(x_i)} \right) \]
\[ \leq 2 \sup_{h \in C} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - E \log \frac{h}{f} \right| \]
\[ + \left( \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i)}{f_k(x_i)} - \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i)}{f_k(x_i)} \right) \]
\[ \leq E \left[ \frac{c_1}{\sqrt{n}} \int_{0}^{b} \log^{1/2} D(\mathcal{H}, \epsilon, d_n) \, d\epsilon \right] + c_2 \sqrt{\frac{T}{n}} + \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_k(x_i)}{f(x_i)} \]
with probability at least \( 1 - e^{-t} \) (by Th. 4.1). Note that \( \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_k(x_i)}{f(x_i)} \leq 0 \) if \( \hat{f}_k \) is constructed by maximizing the likelihood over \( k \)-component mixtures. If it is constructed by the greedy algorithm described in the previous section, \( \hat{f}_k \) achieves “almost maximum likelihood” (see p. 27 of [8], or Sect. 3 of [7]) in the following sense:
\[ \forall g \in C, \quad \frac{1}{n} \sum_{i=1}^{n} \log (f_k(x_i)) \geq \frac{1}{n} \sum_{i=1}^{n} \log (g(x_i)) - \gamma \frac{c_{f_n,p}^2}{k}. \]
Here \( c_{f_n,p}^2 = (1/n) \sum_{i=1}^{n} \int \frac{f_k^2(x_i)P(dg)}{f_k(x_i)P(dg)} \leq \frac{b^2}{a^2} \) and \( \gamma = 4 \log(3\sqrt{e}) + 4 \log \frac{b}{a} \). Hence, with probability at least \( 1 - e^{-t} \),
\[ D(f \| \hat{f}_k) - D(f \| f_k) \leq E \left[ \frac{c_1}{\sqrt{n}} \int_{0}^{b} \log^{1/2} D(\mathcal{H}, \epsilon, d_n) \, d\epsilon \right] + c_2 \sqrt{\frac{T}{n}} + \gamma \frac{c_{f_n,p}^2}{k}. \]
We now write the overall error of estimating an unknown density $f$ as the sum of approximation and estimation errors. The former is bounded by (1) and the latter is bounded as above. Note again that $c_{f,P}^2$ and $\gamma$ in the approximation bound (1) are bounded above by constants which depend only on $a$ and $b$. Therefore, with probability at least $1 - e^{-t}$,

$$D(f \| \hat{f}_k) - D(f \| C) = (D(f \| f_{k_0}) - D(f \| C)) + (D(f \| \hat{f}_k) - D(f \| f_{k_0}))$$

$$\leq \frac{c}{k} + \left[ \frac{c_1}{\sqrt{n}} \int_0^b \log^{1/2} D(H, \epsilon, d_n) d\epsilon \right] + c_2 \sqrt{\frac{t}{n}}. \quad (6)$$

Finally, we rewrite the above probabilistic statement as a statement in terms of expectations. Let $\zeta = \frac{\xi}{t} + E \left[ \frac{c_1}{\sqrt{n}} \int_0^b \log^{1/2} D(H, \epsilon, d_n) d\epsilon \right]$ and $\xi = D(f \| \hat{f}_k) - D(f \| C)$. We have shown that $P \left( \xi \geq \zeta + c_2 \sqrt{\frac{t}{n}} \right) \leq e^{-t}$. Since $\xi \geq 0$,

$$E[\xi] = \int_0^\zeta P(\xi > u) du + \int_\zeta^{\infty} P(\xi > u) du \leq \xi + \int_0^{\infty} P(\xi > u + \zeta) du.$$ 

Now set $u = c_2 \sqrt{\frac{t}{n}}$. Then $t = c_3 nu^2$ and $E[\xi] \leq \xi + \int_0^{\infty} e^{-c_3 nu^2} du \leq \xi + \frac{c_4}{\sqrt{n}}$. Hence,

$$E \left[ D(f \| \hat{f}_k) \right] - D(f \| C) \leq \frac{c_1}{k} + E \left[ \frac{c_2}{\sqrt{n}} \int_0^b \log^{1/2} D(H, \epsilon, d_n) d\epsilon \right]. \quad \square$$

Remark 4.3. Inequality (6) is much stronger than the result of Theorem 2.1 because it reveals the tail behavior of $D(f \| \hat{f}_k) - D(f \| C)$. Nevertheless, to be able to compare our results to those of Li and Barron, we present our results in terms of expectations.

Remark 4.4. In the actual proof of the bounds, Li and Barron [7, 8] use a specific sequence of $\alpha_i$ for the finite combinations. The authors take $\alpha_1 = 1$, $\alpha_2 = \frac{1}{2}$, and $\alpha_k = \frac{2}{k}$ for $k \geq 2$. It can be shown that in this case

$$f_k = \frac{2}{k(k-1)} \left( \frac{1}{2} \phi_1 + \frac{1}{2} \phi_2 + \sum_{m=3}^{k} (m-1) \phi_m \right),$$

so the later choices have more weight.

We now prove Corollary 2.2:

**Proof.** Since we consider bounded densities $a \leq \phi_0 \leq b$, condition (4) implies that

$$\forall x, \log \left( \frac{\phi_0(x) - \phi_{0'}(x)}{b} + 1 \right) \leq ||\theta - \theta'||_{L_1}.$$ 

This allows us to bound $L_\infty$ distances between functions in $H$ in terms of the $L_1$ distances between the corresponding parameters. Since $\Theta$ is a $d$-dimensional cube of side-length $A$, we can cover $\Theta$ by $\left( \frac{A}{2} \right)^d$ “balls” of $L_1$-radius $d_1^0 \frac{\theta}{2}$. This cover induces a cover of $H$. For any $f_0$ there exists an element of the cover $f_0'$, so that

$$d_n(f_0, f_0') \leq ||f_0 - f_0'||_{\infty} \leq b \left( \frac{B_1}{2} \right) - b = \epsilon.$$
Therefore, \( \delta = \frac{2 \log (\frac{1}{\beta} + 1)}{Bd} \) and the cardinality of the cover is \( (\frac{A}{\delta})^d = \left( \frac{ABd}{2 \log (\frac{1}{\beta} + 1)} \right)^d \). Hence,

\[
\int_0^b \log^{1/2} D(H, \epsilon, d_n) d\epsilon = \int_0^b \sqrt{d} \log \frac{ABd}{2 \log (\frac{1}{\beta} + 1)} d\epsilon.
\]

A straightforward calculation shows that the integral above converges. \( \square \)

By creating a simple net over the class \( F \) in Corollary 2.3, one can easily show that \( F \) has a finite cover \( D(F, \epsilon, d_n) = \frac{K}{\epsilon^2} \), for some constant \( K \). Corollary 2.3 follows.

5. Appendix A

We will denote \( f_i = f(x_i) \). Let \( \epsilon_1, \ldots, \epsilon_n \) be i.i.d. Rademacher random variables, i.e. \( \Pr(\epsilon_i = -1) = \Pr(\epsilon_i = +1) = 1/2 \). The following inequality can be found in [6], Theorem 4.12.

**Lemma A.1** [6] Comparison inequality for Rademacher processes. If \( \phi_i : \mathbb{R} \to \mathbb{R} \) \((i = 1, \ldots, n)\) are contractions \((\phi_i(0) = 0 \text{ and } |\phi_i(s) - \phi_i(t)| \leq |s - t| )\), then

\[
E_\epsilon \sup_{f \in F} \left| \sum_{i=1}^n \epsilon_i \phi_i(f_i) \right| \leq 2E_\epsilon \sup_{f \in F} \left| \sum_{i=1}^n \epsilon_i f_i \right|.
\]

**Lemma A.2** [14] Symmetrization. Consider the following processes:

\[
Z(x) = \sup_{f \in F} \left| \mathbb{E} f - \frac{1}{n} \sum_{i=1}^n f(x_i) \right|, \quad R(x) = \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|.
\]

Then

\[
EZ(x) \leq 2ER(x).
\]

The quantity \( ER(x) \) is called the Rademacher average of \( F \).

**Lemma A.3** [9] McDiarmid’s inequality. Let \( x_1, \ldots, x_n, x'_1, \ldots, x'_n \in \Omega \) be i.i.d. random variables and let \( Z : \Omega^n \to \mathbb{R} \) such that

\[
\forall x_1, \ldots, x_n, x'_1, \ldots, x'_n \quad |Z(x_1, \ldots, x_n) - Z(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, x_n)| \leq c_i,
\]

then

\[
\mathbb{P}(Z - EZ > \epsilon) \leq \exp \left( -\frac{\epsilon^2}{2 \sum_{i=1}^n c_i^2} \right).
\]

**Acknowledgements.** We would like to thank Shahar Mendelson, Joel Zinn, Andrew Barron, and Eric Kolaczyk for useful discussions.

**References**


