Consistent Density Estimation from Distribution Functions

Malik Magdon-Ismail*
malik@wild.work.caltech.edu
Caltech Learning Systems Group
Department of Electrical Engineering
California Institute of Technology
136-93 Pasadena, CA, 91125

Amir F. Atiya
amir@deep.caltech.edu
Caltech Learning Systems Group
Department of Electrical Engineering
California Institute of Technology
136-93 Pasadena, CA, 91125

July 24, 1998

Abstract

We approach density estimation by way of estimating the distribution function and taking the derivative. Let G be an unknown distribution function on the real line for which the derivatives up to order $K > 1$ are continuous and bounded, and let g be the associated unknown density. Let N samples $\{x_i\}_{i=1}^N$ be generated independently from G, the task being to estimate the unknown density g. The *generalized sample distribution function* is introduced, for which $O(N^{-1})$ L_2-convergence to the true distribution function and $O((N^{-1}\log\log N)^{1-1/K})$ L_2-convergence to the true density are obtained. When the support of G is compact, we additionally demonstrate $O(N^{-1})$ convergence of the generalized inverse sample distribution function to G^{-1}. Our methods are suggestive of density estimation techniques via sample distribution functions.

1 Introduction

We consider the problem of density estimation, and propose an approach whereby we first estimate the distribution function, and then take the derivative to obtain the density estimate. Let X be a random variable for which the (unknown) distribution function is $G(x)$ and the corresponding (unknown) density is $g(x)$. Let $\{x_1, \ldots, x_N\}$ be the data set drawn

*To whom correspondence should be addressed.

from that distribution. A well-known estimate for the distribution function is the sample distribution function, given by

$$G_N(x) = \frac{1}{N}\sum_{i=1}^{N}\Theta(x - x_i) \tag{1}$$

where $\Theta(x) = 1$ if $x \geq 0$ and $\Theta(x) = 0$ otherwise. The convergence properties of the sample distribution function have been well-studied [17, pg 62]. For example, it is known to converge uniformly to the true distribution function:

$$\sup_x |G(x) - G_N(x)| \xrightarrow{N\to\infty} 0 \qquad \text{with probability 1.} \tag{2}$$

Because the estimate (1) is not differentiable, researchers have not gone further to apply it to the density estimation problem. In this study, we circumvent this problem by introducing an approximate distribution function estimate, that is a smooth approximate interpolation of the sample distribution function. We call these functions *generalized sample distribution functions*. Because of their smoothness, we propose using them to estimate the density function. We obtain bounds on the estimation error for these estimates.

We break down the paper as follows. In the next section, we present a discussion of our results in the context of existing results. The formal statements of the results are given in section 3 and some practical issues are addressed in section 4 along with a simulation.

2 Discussion of Results

Traditional density estimation methods can be grouped into two broad categories (see [18], [5], and [10]). The first of these is the parametric approach. Typically $O(1/N)$ convergence is obtained provided that the true density is implementable by the parametric class. For example, [16, pg 19-40] shows that when estimating a Gaussian density and the parameter of a uniform density, O(1/N) convergence is obtained. The only case we find in the literature that obtains faster than $O(1/N)$ convergence is for the special case of the uniform density where $O(1/N^2)$ can be obtained using an estimator of the form cx_N where x_N is the n^{th} order statistic and $c \approx 1 + \log 2/N$. Rarely is it the case that the true distribution is in the parametric class, therefore the need for the non parametric and "semi-parametric" techniques.

By semi-parametric is meant fixing a parametric class but allowing the number of parameters to be adjusted. As an example, consider the mixture of Gaussians model where the number of Gaussians is a variable. By allowing the number of parameters to tend to infinity in a suitable way as $N \to \infty$, it is possible to obtain convergence to the true density. Such a semi-parametric technique is studied by Stone [23] using a log-spline model. A convergence rate of $O_P(N^{-r})$ where $r = (K-1)/(2K-1)$ is obtained in L_2. Under suitable specifications, these are the optimal rates of convergence, [20, 21, 22]. The optimal rate for the integrated squared error for density functions with bounded r-Sobolev norm (where r is the maximum degree for which $\int \left|f^{(r)}\right| < \infty$) is $O(N^{-(1-1/(2r+1))})$, [6] and [11]. Barron, [3], has obtained a similar rate of convergence for a cross entropy error function using sequences of exponential families.

Perhaps the most common non-parametric approach is the Parzen/Kernel estimate, [15]. Under the assumption that the density has square integrable second derivative and the optimal kernel width is used, $O(N^{-4/5})$ L_2 convergence is obtained, [18, pg 40].

We present convergence results for generalized distribution functions. A generalized distribution function $H(x)$ approximately interpolates the points $\{x_i, i/(N+1)\}$ and is an estimate for the distribution function. Provided that the interpolation error at every point is $O(N^{-1/2})$, we show that the mean integrated squared error (MISE) is $O(N^{-1})$, i.e.,

$$\left\langle \int_{-\infty}^{\infty} dx\; g(x)(H(x) - G(x))^2 \right\rangle_{\{x_i, i/(N+1)\}} \leq \frac{C}{N} \tag{3}$$

where we use $\langle\cdot\rangle$ to denote expectations. Similarily, one can obtain L_1 convergence at a rate $O(N^{-1/2})$. For comparison, the L_∞ convergence is $O(\sqrt{\log\log(N)/N})$ with probability 1, [17, pg 62]. The implementation of the generalized approximate distribution functions can be accomplished using non-parametric or semi-parametric techniques as in section 4. Assuming that the support of G is compact, one can obtain $O(N^{-1})$ convergence to G^{-1} for the MISE. To obtain convergence to the true density g, we assume that the derivatives of G are bounded up to order K. Then the distribution function that minimizes the maximum value of the magnitude of the K^{th} derivative subject to the constraint that the interpolation error is at most $O(\sqrt{\log\log N/2N})$ has its MISE bounded by $O((\log\log N/N)^{1-1/K})$ with probability 1. We place no other restrictions on the interpolators thus our results are applicable to non-parametric distribution function estimation.

3 Results

Two well studied statistics of the sample distribution function are the Kolmogorov-Smirnov statistic ($\mathcal{D}_N$) and the Cramér–von Mises statistic ($\mathcal{C}_N$) defined as follows

$$\mathcal{D}_N = \sup_x |G_N(x) - G(x)| \tag{4}$$

$$\mathcal{C}_N = \int_{-\infty}^{\infty} (G_N(x) - G(x))^2 dG(x) \tag{5}$$

The following theorems are valid [17, pg 61–64]

Theorem 3.1 *With probability 1*

$$\limsup_{N\to\infty} \mathcal{D}_N \sqrt{\frac{N}{2\log\log(N)}} = \frac{1}{2} \tag{6}$$

PROOF: See [19], [7] or [8]. ■

Theorem 3.2 *With probability 1*

$$\limsup_{N\to\infty} \mathcal{C}_N \frac{N}{2\log\log(N)} = \frac{1}{\pi^2} \tag{7}$$

PROOF: See [12].

It is not hard to extend these theorems to apply to generalized sample distribution functions which we define shortly. These theorems give the probability 1 convergence behavior. We will look at $\langle \mathcal{C}_N \rangle_D$, the expected performance in the L_2 error criterion (MISE), as well as the corresponding quantity in the L_1 error criterion. The expectations are with respect to the data set. We will derive bounds on the estimation error for the distribution function estimator, as well as the density estimator. Let us start with the following definition.

Let $\mathcal{G}$ be the space of functions such that for every $X \in \mathcal{G}$, the following holds

- $X : \mathbf{R} \to [0, 1]$
- $X(t)$ is continuously differentiable and $X'(t) \geq 0$.
- $X(-\infty) = 0$ and $X(\infty) = 1$.

Thus, $\mathcal{G}$ is the space of distribution functions on the real line that possess continuous density functions, which is the class of functions that we will be interested in. We define a metric, the L_2 X-norm of f, as follows

$$\| f \|_X^2 = \int_{-\infty}^{\infty} f(t)^2 X'(t) dt \tag{8}$$

$\| f \|_X^2$ is the expectation of the squared value of f with respect to the distribution $X \in \mathcal{G}$. We can analogously define the L_1 X-norm of f as

$$|f|_X = \int_{-\infty}^{\infty} |f(t)| X'(t) dt \tag{9}$$

Let the data set (D) be $\{x_1 \leq x_2 \leq \ldots \leq x_N\}$, and corresponding to each x_i, let $y_i = i/N+1$. We are interested in the mapping $x_i \mapsto y_i$. Define an approximate sample distribution function as follows

Definition 3.3 *A* <u>*ν-approximate generalized sample distribution function*</u>*, H, satisfies the following two conditions*

- $H \in \mathcal{G}$
- $\left| H(x_i) - \frac{i}{N+1} \right| \leq \nu(N) \sqrt{\frac{\log\log(N)}{2N}}, \ \forall i$

Where $\nu(N)$ is some function of N. We will denote the set of all ν-approximate sample distribution functions for a data set, D, and a given $\nu(N)$ by $\mathcal{H}_D^\nu$.

The set $\mathcal{H}_D^\nu$ contains those continuously differentiable distribution functions that approximately interpolate the data set $\{(x_i, y_i)\}_{i=1}^N$. For $\nu(N) = 0$ we have a generalized sample distribution function. Note that the conventional distribution function estimator (1) is not in this class of generalized sample distribution functions, but it is the limit of functions that are in this class.

3.1 Convergence to the Distribution Function

Let us now derive the estimation error for the distribution function estimator. Let $H \in \mathcal{H}_D^\nu$, and let $G \in \mathcal{G}$ be the true distribution function. As we mentioned, H approximates the distribution function, given a number N of observed values by interpolation. Write $H(x_i) = y_i - \eta_i$, where $|\eta_i| \leq \nu(N)\sqrt{\log\log N/2N}$. The true distribution function evaluated at x_i is generally not equal to y_i. Therefore we will write $G(x_i) = y_i + \epsilon_i$. There are two sources of estimation error: the error at each data point, $\epsilon_i + \eta_i$, and the interpolation error between the data points – an error that would persist even if the $\epsilon_i + \eta_i$'s were all zero. As a first step, let us analyze the statistics of the ϵ_i's. It is well known that $G(x)$ has a uniform distribution in [0,1], therefore, $u_i = G(x_i)$ is the i^{th} order statistic of the uniform distribution. Thus, the density of (u_i, u_j) can be obtained by standard techniques (eg. [4, pg 200]), and is given by

$$g(u_i, u_j) = \frac{N!}{(i-1)!(j-i-1)!(N-j)!} u_i^{i-1}(1-u_j)^{N-j}(u_j-u_i)^{j-i-1} \quad j \geq i \tag{10}$$

Noticing that $\epsilon_i = u_i - i/(N+1)$, we can calculate the first two moments of the ϵ_i's as

$$\langle \epsilon_i \rangle = 0 \qquad \langle \epsilon_i \epsilon_j \rangle = \frac{i(N+1-j)}{(N+1)^2(N+1)} \tag{11}$$

The following theorem provides a bound for the estimation error, which will prove consistency of the distribution estimator.

Theorem 3.4 (**L_2 convergence to the true distribution**) *Let the data set D consist of N data points, x_i drawn i.i.d. from the (true) distribution $G \in \mathcal{G}$. For every $H \in \mathcal{H}_D^\nu$ and every $F \in \mathcal{G}$, the inequality*

$$\left\langle \| H - G \|_F^2 \right\rangle_D = \left\langle \int_{-\infty}^{\infty} (H(x) - G(x))^2 F'(x) dx \right\rangle_D \leq \mathcal{Q}_2(N) \tag{12}$$

holds, where

$$\mathcal{Q}_2(N) = \frac{1}{2(N+1)} + \frac{3}{2(N+1)^2} + \nu(N)^2 \frac{\log\log N}{N} + \frac{4\nu(N)}{N+1}\sqrt{\frac{\log\log N}{2N}} \tag{13}$$

Proof: Let $x_0 = -\infty$, $x_{N+1} = \infty$.

$$\| H - G \|_F^2 = \int_{-\infty}^{\infty} (H(x) - G(x))^2 F'(x) dx = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} (H(x) - G(x))^2 F'(x) dx \tag{14}$$

Now by definition, $G(x_i) = H(x_i) + \epsilon_i + \eta_i$. Let $x \in [x_i, x_{i+1}]$. Because both H and G are monotonically increasing we have that $i/(N+1) - \eta_i = H(x_i) \leq H(x) \leq H(x_{i+1}) = (i+1)/(N+1) - \eta_{i+1}$ and $i/(N+1) + \epsilon_i = G(x_i) \leq G(x) \leq G(x_{i+1}) = (i+1)/(N+1) + \epsilon_{i+1}$.

We then get

$$\epsilon_i + \eta_{i+1} - \frac{1}{N+1} \leq G(x) - H(x) \leq \epsilon_{i+1} + \eta_i + \frac{1}{N+1} \tag{15}$$

Thus

$$(G(x) - H(x))^2 \leq (\epsilon_i + \eta_{i+1} - \frac{1}{N+1})^2 + (\epsilon_{i+1} + \eta_i + \frac{1}{N+1})^2 \tag{16}$$

Therefore using (11)

$$\begin{aligned}\left\langle (G(x) - H(x))^2 \right\rangle_D &= \frac{2}{(N+1)^2} + \frac{i(N+1-i) + (i+1)(N-i)}{(N+1)^2(N+2)} + \\ &\quad \eta_i^2 + \eta_{i+1}^2 + \frac{2}{N+1}(\eta_i - \eta_{i+1}) \\ &\leq \mathcal{Q}_2(N), \end{aligned} \tag{17}$$

where the last inequality follows by maximizing over i. Thus we find

$$\begin{aligned}\left\langle \| H - G \|_F^2 \right\rangle_D &= \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} \left\langle (H(x) - G(x))^2 \right\rangle_{x_i} F'(x) dx, & (18)\\ &\leq \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} \mathcal{Q}_2(N) F'(x) dx, & (19)\\ &= \mathcal{Q}_2(N) & (20)\end{aligned}$$

The last line follows because $\int_{-\infty}^{\infty} F'(x)dx = 1$ since F is a distribution function. ■

Note 1: The theorem applies to any $H \in \mathcal{H}_D^\nu$ and any $F \in \mathcal{G}$, in particular to $F(x) = G(x)$, the true distribution function. Thus the expected squared error approaches zero asymptotically at a rate $\mathcal{Q}_2(N)$.

Note 2: The same proof can be modified for the case of the integrated squared error over some bounded interval.

Note 3: Without much extra effort, this result can be extended to a larger class ${\mathcal{H}_D^\nu}^*$ of interpolators where the monotonicity condition is replaced by the less restrictive condition $H \in {\mathcal{H}_D^\nu}^*$ if

$$x \in [x_i, x_{i+1}] \quad \Rightarrow \quad H(x) \in [H(x_i), H(x_{i+1})] \tag{21}$$

with no other constraints being made on continuity or the existence of derivatives. To see this notice that the monotonicity of H is used only in obtaining (15), where (21) can be used in its place. Note that the conventional distribution function (1) belongs to this extended class $\mathcal{H}_D^{\nu *}$.

Note 4: If $\nu(N) \leq 1/\sqrt{\log\log N}$– i.e., one interpolates the distribution function with an error $O(N^{-1/2})$, convergence to the true distribution function is obtained at a rate $O(1/N)$.

The previous theorem can easily be extended to obtain the L_1 convergence.

Theorem 3.5 (L_1 convergence to the true distribution) *Let the data set D consist of N data points, x_i drawn i.i.d. from the (true) distribution $G \in \mathcal{G}$. For every $H \in \mathcal{H}_D^\nu$ and every $F \in \mathcal{G}$, the inequality*

$$\langle |H - G|_F \rangle_D = \left\langle \int_{-\infty}^{\infty} |H(x) - G(x)| F'(x) dx \right\rangle_D \leq \mathcal{Q}_1(N) \tag{22}$$

holds, where

$$\mathcal{Q}_1(N) = \frac{1}{\sqrt{N+2}} + \frac{2}{N+1} + 2\nu(N)\sqrt{\frac{\log\log N}{2N}} \tag{23}$$

Proof: The same reasoning that led up to (15) is still valid. Thus, we find for $x \in [x_i, x_{i+1}]$

$$|H(x) - G(x)| \leq |\epsilon_i| + |\epsilon_{i+1}| + \frac{2}{N+1} + |\eta_i| + |\eta_{i+1}| \tag{24}$$

By Jensen's inequality, $\langle|\epsilon_i|\rangle \leq \sqrt{\langle\epsilon_i^2\rangle}$ therefore we find that $\langle|\epsilon_i|\rangle \leq \frac{1}{2}\sqrt{1/(N+2)}$. The rest of the analysis is exactly analogous to the L_2 convergence proof, the final result being (22). ∎

Thus we have L_1 convergence at a rate $O(1/\sqrt{N})$ for $\nu(N) \leq 1/\sqrt{\log\log N}$.

3.2 Convergence to the True Density Function

The density estimator is the derivative of the distribution function estimator and thus we need to evaluate $\left\langle \| H' - G' \|_F^2 \right\rangle_{x_i}$. Obtaining a tight convergence rate for the density estimation error is a tougher job. The reason is that the derivative operation accentuates the noise. It is well known that the density estimation problem is an ill-posed problem. Without any *a priori* assumptions on the true density function, it will be hard to judge the suitability of an estimator. Even the trivial estimator consisting of the summation of delta functions centered at the data points could be as valid as other estimators. In fact, [2] and [9] show that no density estimator is consistent for certain types of error measure unless one makes some *a priori* assumptions about the distribution function. Typical *a priori* assumptions used in the density estimation literature are smoothness constraints on the class of considered densities. These are realistic assumptions that are usually obeyed by densities that are typically encountered in real world applications. We will consider such constraints in our analysis. First, we need the following results relating the higher derivatives of a function to lower order derivatives.

Lemma 3.6 *Let a twice differentiable function $f(x)$ be defined on the interval (a, ∞) where the lower limit (a) could be $-\infty$. Suppose $\sup_x |f| = M_0$, $\sup_x |f'| = M_1$, and $\sup_x |f''| = M_2$, where the $\sup_x$ is taken over $x \in (a, \infty)$. Then,*

$$M_1^2 \leq 4M_0M_2 \tag{25}$$

PROOF: Let $h > 0$. Then for $x \in (a, \infty)$, by Taylor's theorem, we have

$$f(x + 2h) = f(x) + 2hf'(x) + (2h)^2 \frac{f''(\zeta)}{2} \tag{26}$$

for some $\zeta \in (x, x + 2h)$. Therefore,

$$|f'(x)| \leq \frac{M_0}{h} + hM_2 \tag{27}$$

This inequality holds for every $h > 0$. In particular for the h that minimizes the right hand side. This minimum is attained at $h = \sqrt{M_0/M_2}$ and so we have the bound

$$|f'(x)| \leq 2\sqrt{M_0M_2} \tag{28}$$

Thus $\sup_x |f'(x)| \leq 2\sqrt{M_0M_2}$. ■

Corollary 3.7 *Let $G(x)$ be differentiable m times on the interval (a, ∞). Let $\sup_x \left|G^{(i)}\right| = M_i$ for $i = 1 \ldots m$. $G^{(i)}$ denotes the i^{th} derivative. Then*

$$M_i^2 \leq 4M_{i-1}M_{i+1} \tag{29}$$

for $i = 1 \ldots (m-1)$.

PROOF: Let $f(x)$ in lemma 3.6 be $G^{(i-1)}(x)$. ■

Lemma 3.8 *Let $G(x)$ be differentiable m times on the interval (a, ∞). Let $\sup_x \left|G^{(i)}\right| = M_i$ for $i = 1 \ldots m$. Then*

$$M_1 \leq 2^{m-1} M_0^{1-\frac{1}{m}} M_m^{\frac{1}{m}} \tag{30}$$

PROOF: We prove the lemma by induction on m. For $m = 2$, the lemma is equivalent to lemma 3.6. Suppose (30) holds for $m = P$. We show that it holds for $m = P + 1$. From lemma 3.6 we have

$$M_1 \leq 2 M_0^{\frac{1}{2}} M_2^{\frac{1}{2}} \tag{31}$$

By applying the induction hypothesis to G' we have the following equation

$$M_2 \leq 2^{P-1} M_1^{1-\frac{1}{P}} M_{P+1}^{\frac{1}{P}} \tag{32}$$

Combining these two equations we find that

$$M_1 \leq \left(2^{\frac{P+1}{2}} M_0^{\frac{1}{2}} M_{P+1}^{\frac{1}{2P}}\right)^{\frac{2P}{P+1}}, \tag{33}$$

and thus (30) holds for $m = P + 1$. This concludes the proof. ■

Corollary 3.9 *Let $G(x) \in \mathcal{C}^\infty$ on the interval (a, ∞). Let $\sup_x \left|G^{(i)}\right| = M_i$. Define the sequence $\{s_k\}_{k=0}^{\infty}$ by $s_k = 2^{k-1} M_k^{1/k}$. Suppose that $\lim_{k\to\infty} s_k = \beta < \infty$. Then*

$$M_1 \leq \beta M_0 \tag{34}$$

PROOF: Taking the limit $m \to \infty$ in lemma 3.8 and using the fact that the term by term product of two convergent sequences converges to the product of the limits, the corollary follows. ■

After these preliminary results, we now come to our density estimator. We will assume that the true distribution function has bounded derivatives up to order K. Then, in the asymptotic limit $N \to \infty$, one expects that with probability 1, an interpolator should exist with bounded derivatives with the same bounds, because in the asymptotic limit, $G(x)$ itself becomes an interpolator. To proceed in a more formal manner, let $A_i = \sup_x \left|G^{(i)}\right|$, $i = 1 \ldots K$ and define $B_i^\nu(D)$ by

$$B_i^\nu(D) = \inf_{H \in \mathcal{H}_D^\nu} \sup_x \left|H^i\right| \tag{35}$$

for fixed $\nu > 0$. Note that by definition, for all $\epsilon > 0$, $\exists\ H \in \mathcal{H}_D^\nu$ such that $\sup_x \left|H^{(i)}(x)\right| \leq B_i^\nu + \epsilon$. $B_i^\nu(D)$ is the lowest possible bound on the i^{th} derivative for the ν-approximate sample distribution functions given a particular data set. In a sense, the "smoothest" approximating sample distribution function with respect to the i^{th} derivative has an i^{th} derivative bounded by $B_i^\nu(D)$ One expects that $B_i^\nu \leq A_i$, at least in the limit $N \to \infty$. This is the content of the next lemma. Let $B_i^\nu = A_i + \eta_i^\nu(D)$.

Lemma 3.10 *Let $\nu(N) > 1$. Then, with probability 1,*

$$\eta_i^\nu(D) \stackrel{N\to\infty}{\leq} 0 \tag{36}$$

By this is meant that for all $\delta > 0$

$$\lim_{N\to\infty} P[\eta_i^\nu(D) < \delta] = 1 \tag{37}$$

Proof: This is a straightforward application of theorem 3.1 as follows. Suppose the lemma were false. Then there exists $\delta > 0$ such that $\lim_{N\to\infty} P[\eta_i^\nu(D) > \delta] \geq \gamma > 0$. Therefore with probability$\geq \gamma$, $G \notin \mathcal{H}_D^\nu$ as $N \to \infty$, as if $G \in \mathcal{H}_D^\nu$ then $\eta \leq 0$. Thus with probability $\geq \gamma$,

$$\mathcal{D}_N = \sup_x |G_N(x) - G(x)| > \nu\sqrt{\frac{\log\log(N)}{2N}} \tag{38}$$

as $N \to \infty$. Thus,

$$\lim_{N\to\infty} D_N\sqrt{\frac{N}{2\log\log(N)}} \geq \lim_{N\to\infty} \frac{\nu}{2} > \frac{1}{2}$$

with probability$\geq \gamma > 0$, implying that

$$\limsup_{N\to\infty} D_N\sqrt{\frac{N}{2\log\log(N)}} > \frac{1}{2} \tag{39}$$

with probability$\geq \gamma > 0$, contradicting theorem 3.1. ■

In the next theorem, we present the main result of the paper, namely a bound on the estimation error for the density estimator obtained by using the approximate sample distribution functions. Its essential content is that if the true distribution function has bounded derivatives to order K then picking the approximate distribution function with "minimum" value for B_K, where B_K is a bound on the magnitude of the K^{th} derivative, we obtain a convergence rate for the estimation error of $O((\log\log(N)/N)^{1-1/K})$.

Theorem 3.11 (L_2 convergence to the true density) *Let N data points, x_i be drawn i.i.d. from the distribution $G \in \mathcal{G}$. Let $\sup_x \left|G^{(i)}\right| = A_i$ for $i = 0 \ldots K$, where $K \geq 2$. Let $\nu(N) > 1$ and fix $\epsilon > 0$. Let $B_K^\nu(D) = \inf_{H \in \mathcal{H}_D^\nu} \sup_x \left|H^{(K)}\right|$. Let $H \in \mathcal{H}_D^\nu$ be a ν-approximate distribution function with $B_K = \sup_x \left|H^{(K)}\right| \leq B_K^\nu + \epsilon$ (by the definition of B_K^ν, such a ν-approximate sample distribution function must exist). Then, for any $F \in \mathcal{G}$, as $N \to \infty$, the inequality*

$$\left\langle \| H' - G' \|_F^2 \right\rangle_D \leq 2^{2(K-1)}(2A_K + \epsilon)^{\frac{2}{K}} \mathcal{F}(N) \tag{40}$$

where

$$\mathcal{F}(N) = \left[(1 + \nu(N)) \left(\frac{2 \log\log(N)}{N}\right)^{\frac{1}{2}} + \frac{2}{N+1}\right]^{2-\frac{2}{K}} \tag{41}$$

holds with probability 1, as $N \to \infty$.

Proof: Let $S(x) = G(x) - H(x)$. Then $M_K = \sup_x \left|S^{(K)}\right| \leq A_K + B_K$, and

$$\begin{aligned} \| H' - G' \|_F^2 &= \int_{-\infty}^{\infty} S'(x)^2 F'(x) dx & (42) \\ &\leq \sup_x S'(x)^2 & (43) \end{aligned}$$

By lemma 3.8

$$S'(x)^2 \leq 2^{2(K-1)}(A_K + B_K)^{\frac{2}{K}} \underbrace{\sup_x S(x)^{2-\frac{2}{K}}}_{I}. \tag{44}$$

where $B_K = \sup_x \left|H^{(K)}\right|$. By construction, $B_K \leq B_K^\nu(D) + \epsilon$. By lemma 3.10, as $N \to \infty$, $B_K^\nu(D) \leq A_K$ with probability 1, so,

$$(A_K + B_K)^{\frac{2}{K}} \leq (2A_K + \epsilon)^{\frac{2}{K}} \tag{45}$$

holds with probability 1 in the asymptotic limit. We are interested in the expected value of (44) with respect to D. Therefore, we need the expected value of I,

$$\langle I\rangle = \left\langle (\sup_x |S(x)|)^{2-\frac{2}{K}} \right\rangle_{\epsilon_i} \tag{46}$$

Jensen's inequality gives for $0 < a < 1$ that $\langle|\phi|^a\rangle \leq \langle|\phi|\rangle^a$. Applying this to $\phi = S^2$ yields

$$\langle I\rangle \leq \left\langle \sup_x S^2(x) \right\rangle_{\epsilon_i}^{1-\frac{1}{K}} \tag{47}$$

Because H is a ν-approximate sample distribution function, $H(x_i) = i/(N+1) + \eta_i$ where $|\eta_i| \leq \nu(N)\sqrt{\log\log(N)/2N}$. Similar reasoning to that which led to (15) can now be applied to show that for $x \in [x_i, x_{i+1}]$,

$$\epsilon_i + \eta_{i+1} - \frac{1}{N+1} \leq G(x) - H(x) \leq \epsilon_{i+1} + \eta_i + \frac{1}{N+1} \tag{48}$$

therefore

$$|S(x)| \leq |\epsilon_i| + |\epsilon_{i+1}| + \nu(N)\left(\frac{2\log\log(N)}{N}\right)^{\frac{1}{2}} + \frac{2}{N+1} \tag{49}$$

and using theorem 3.1, we see that with probability 1, as $N \to \infty$,

$$\sup_x S^2(x) \leq \underbrace{\left[(1+\nu(N))\left(\frac{2\log\log(N)}{N}\right)^{\frac{1}{2}} + \frac{2}{N+1}\right]^2}_{\mathcal{F}(N)} \tag{50}$$

This inequality holds with probability 1 in the asymptotic limit, therefore, its expectation, $\langle I\rangle_D$, is also bounded by the same quantity. To see this, suppose that $\lim_{N\to\infty}\langle I\rangle_D > \mathcal{F}(N)$, then with some probability $\gamma > 0$, $I > \mathcal{F}(N)$, otherwise $\langle I\rangle_D \leq \mathcal{F}(N)$. Therefore $\sup_x S^2(x) > \mathcal{F}(N)$ with probability $\gamma > 0$ in the asymptotic limit, contradicting (50) . Using (45) in (44), and the fact that $\langle I\rangle$ is bounded by (50), the theorem is proved. ■

Note 5: This theorem holds for any $\epsilon > 0$ and $\nu(N) > 1$. Therefore we see that for smooth density functions, with bounded higher derivatives, the convergence rate approaches $O(\log\log(N)/N)$ which is faster than $1/N^{1-\epsilon}$ convergence for any $\epsilon > 0$. Thus, we have faster convergence than the standard kernel density estimator (for which the rate is $O(N^{-4/5})$).

Note 6: No smoothing parameter needs to be determined unlike the other traditional techniques.

Note 7: From the theorem, it is clear that one should try to find a generalized ν-approximate distribution function with the smallest possible derivatives. Specifically, of all the sample distribution functions, pick the one that "minimizes" B_K, the bound on the K^{th} derivative. Thus, when using optimization techniques to find the generalized distribution function, one would be justified in introducing penalty terms, penalizing the magnitudes of the derivatives (for example Tikhonov type regularizers [24]).

L_1 convergence follows by using very similar arguments to the previous theorem.

Theorem 3.12 (L_1 convergence to the true density) *Let N data points, x_i be drawn i.i.d. from the distribution $G \in \mathcal{G}$. Let $\sup_x \left|G^{(i)}\right| = A_i$ for $i = 0 \ldots K$, where $K \geq 2$. Fix $\nu > 1$ and $\epsilon > 0$. Let $B_K^\nu(D) = \inf_{H \in \mathcal{H}_D^\nu} \sup_x \left|H^K\right|$. Let $H \in \mathcal{H}_D^\nu$ be a ν-approximate distribution function with $B_K = \sup_x \left|H^K\right| \leq B_K^\nu + \epsilon$ (by the definition of B_K^ν, such a ν-approximate sample distribution function must exist). Then, for any $F \in \mathcal{G}$, as $N \to \infty$, the inequality*

$$\left\langle \| H' - G' \|_F^2 \right\rangle_D \leq 2^{(K-1)} (2A_K + \epsilon)^{\frac{1}{K}} \mathcal{F}(N) \tag{51}$$

where

$$\mathcal{F}(N) = \left[(1+\nu) \left(\frac{2 \log\log(N)}{N} \right)^{\frac{1}{2}} + \frac{2}{N+1} \right]^{1-\frac{1}{K}} \tag{52}$$

holds with probability 1.

Proof: The proof proceeds in an analogous fashion to the proof of theorem 3.11 with $\sup_y S^2(y)$ replaced by $\sup_y |S(y)|$. ■

Once again, the theorem applies to any $\epsilon > 0$ and any $\nu > 1$, therefore, for smooth density functions, we have L_1 convergence at a rate $O((\log\log N/N)^{\frac{1}{2}})$, which is faster than $O(1/N^{\frac{1}{2}-\epsilon})$ for any $\epsilon > 0$.

3.3 Convergence to G^{-1}

We will make the simplifying assumption that $G^{-1}(u)$ is continuous on $[0, 1]$. Therefore the input support is compact. Since $G(x)$ was assumed differentiable, then so is $G^{-1}(u)$, and further because $[0,1]$ is compact, we have that $|dG^{-1}(u)/du|$ is bounded. Let the bound on the derivative be S. Then, the input distribution is bounded in a range of size $2S$.

Let $u_i = i/(N+1)$ and let $H^{-1}(u)$ be the inverse of a generalized distribution function that interpolates the points (for simplicity we let $\nu(N) = 0$ though the same convergence rate is obtained for $\nu(N) < 1/\sqrt{\log\log N}$) $\{u_i, x_i\}$. We would like to analyze the rate of convergence of the integrated squared error $\mathcal{E} = E\left[\int_0^1 d\mu(u)\ (H^{-1}(u) - G^{-1}(u))^2\right]$, where μ is any continuous measure on $[0,1]$. Write

$$x_i = G^{-1}(u_i) + \epsilon_i \tag{53}$$

Because both G^{-1} and H^{-1} are monotonic, the same arguments that led up to (15) can be used to get that for $u \in [u_i, u_{i+1}]$,

$$x_i - G^{-1}(u_{i+1}) \le H^{-1}(u) - G^{-1}(u) \le x_{i+1} - G^{-1}(u_i) \tag{54}$$

from which follows the inequality

$$(H^{-1}(u) - G^{-1}(u))^2 \le (x_i - G^{-1}(u_{i+1}))^2 + (x_{i+1} - G^{-1}(u_i))^2 \tag{55}$$

The density for the order statistic x_i is [4, pg 200]

$$\tilde{g}(x_i) = \frac{N!}{(i-1)!(N-i)!} g(x_i) G(x_i)^{i-1}(1 - G(x_i))^{N-i} \tag{56}$$

We will require the expected value $E[(x_i - G^{-1}(u_{i+1}))^2]$ and $E[(x_{i+1} - G^{-1}(u_i))^2]$. They can be bounded as follows

$$\begin{aligned}
E[(x_i - G^{-1}(u_{i+1}))^2] &= \int_{-\infty}^{\infty} dx_i\ (x_i - G^{-1}(u_{i+1}))^2 \tilde{g}(x_i) && (57)\\
&\overset{(a)}{=} \int_0^1 du\ \beta_{i,N+1-i}(u)(G^{-1}(u) - G^{-1}(u_{i+1}))^2 && (58)\\
&\overset{(b)}{=} \int_0^1 du\ \beta_{i,N+1-i}(u) \frac{dG^{-1}(\zeta)}{du}^2 (u - u_{i+1})^2 && (59)\\
&\le S^2 E[(u - u_{i+1})^2]_{\beta_{i,N+1-i}} && (60)
\end{aligned}$$

where

$$\beta_{\alpha,\beta}(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1} \tag{61}$$

(a) follows by using (56) and making a transformation of variables to $u = G(x_i)$, and (b) by Taylor's theorem, for some $\zeta(u)$. A similar calculation yields $E[(x_{i+1} - G^{-1}(u_i))^2] \leq S^2 E[(u-u_i)^2]_{\beta_{i+1,N-i}}$ The expectations are with respect to the relevent β-distributions and can be calculated using standard methods. Then, straightforward bounds for these expectations gives that for $u \in [1/(N+1), N/(N+1)]$,

$$E\left[(H^{-1}(u) - G^{-1}(u))^2\right] \leq S^2 \frac{N-1}{4(N+1)^2} \tag{62}$$

In the two end cells $[0, 1/(N+1)]$ and $[N/(N+1), 1]$, the difference between H^{-1} and G^{-1} is bounded by $2S$. Putting all this together, we find

$$\begin{aligned} E\left[\int_0^1 d\mu(u)\ (H^{-1}(u) - G^{-1}(u))^2\right] &= \sum_{i=0}^{N} \int_{\frac{i}{N+1}}^{\frac{i+1}{N+1}} d\mu(u)\ E\left[(H^{-1}(u) - G^{-1}(u))^2\right] \\ &\leq 4S^2 \mu\left(\left[0, \frac{1}{N+1}\right] \bigcup \left[\frac{N}{N+1}, 1\right]\right) + \\ &\quad S^2 \frac{N-1}{4(N+1)^2} \int_{\frac{1}{N+1}}^{\frac{N}{N+1}} d\mu(u) \\ &\leq S^2 \left(\frac{8B}{N+1} + \frac{N-1}{4(N+1)^2}\right) \end{aligned} \tag{63}$$

where the last inequality follows because $d\mu(u)$ was assumed to be a continuous measure and therefore can be bounded on this compact set (by B). Thus we have proved the following theorem

Theorem 3.13 (L_2 convergence to G^{-1}) *Let the data set D consist of N data points, x_i drawn i.i.d. from the (true) distribution $G \in \mathcal{G}$. Assume that G^{-1} is continuous on the closed interval $[0,1]$ and let its derivative be bounded by S. For every $H \in \mathcal{H}_D^0$ that is zero outside the range of G^{-1} and for any continuous measure $d\mu(u)$ on $[0,1]$, the inequality*

$$\left\langle \| H^{-1} - G^{-1} \|_\mu^2 \right\rangle_D \leq S^2 \left(\frac{8B}{N+1} + \frac{N-1}{4(N+1)^2}\right) \tag{64}$$

holds. ■

Thus we have convergence to G^{-1} provided the conditions of theorem 3.13 hold, and further that the rate of convergence is $O(1/N)$. In a similar manner, one obtains L_1 convergence.

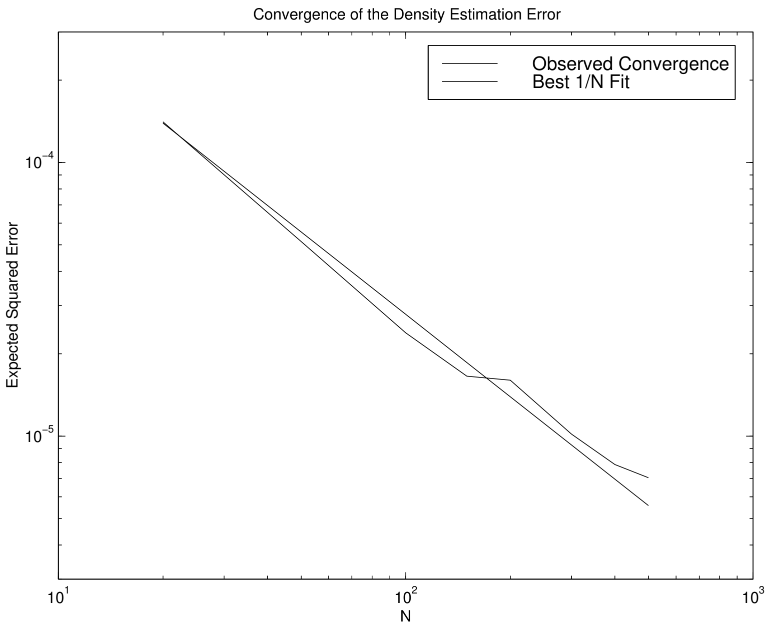
Theorem 3.14 (L_1 convergence to G^{-1}) *Let the data set D consist of N data points, x_i drawn i.i.d. from the (true) distribution $G \in \mathcal{G}$. Assume that G^{-1} is continuous on the closed interval $[0,1]$ and let its derivative be bounded by S. For every $H \in \mathcal{H}_D^0$ that is zero outside the range of G^{-1} and for any continuous measure $d\mu(u)$ on $[0,1]$, the inequality*

$$\left\langle \int_0^1 d\mu(u) \left| H^{-1} - G^{-1} \right| \right\rangle_D \leq S \left(\frac{4B}{N+1} + \frac{\sqrt{N-1}}{2(N+1)} \right) \tag{65}$$

holds.

The proof is analogous to the L_2 convergence, with an application of Jensen's inequality to bound $E[|u - u_i|]$ by $\sqrt{E[(u-u_i)^2]}$.

4 Simulation Results



A five hidden unit two layer neural network was used to perform the mapping $x_i \to i/(N+1)$. For various N, the resulting density estimation error was computed for over 100 runs for each N. Plotted are the results on a Log-Log scale. For comparison, also shown is the best $1/N$ fit.

Figure 1: Convergence of the density estimation error

We have demonstrated convergence of the "smoothest" approximate interpolator. Here we consider its implementation briefly. One possibility which we implement here is using a neural network. It is demonstrated in [13, 14, 1] that a large enough neural network can implement an arbitrary mapping and its derivatives on compacta. Suppose that the distribution function satisfying the conditions of theorem 3.11 is $H(x)$. Then, we can approximate $H(x)$ and its derivatives to an arbitrary accuracy (on compacta) using a large enough Neural Network. We demonstrate the convergence results on simulated data using a neural network in the figure. A five hidden unit neural network was trained to estimate the sample distribution function, for which the density estimate is the derivative of the learned function. The density estimation error was then computed. Shown in figure 1 is the behavior of $\log(E)$ versus $\log(N)$, and for comparison, we show a slope -1 curve. The optimal linear fit had slope -0.97.

5 Acknowledgments

We would like to acknowledge Yaser Abu-Mostafa and the Caltech Learning Systems Group for their useful input.

References

[1] A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.

[2] A. Barron, L. Györfi, and E. van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transactions on Information Theory*, 38(5):1437–1454, September 1992.

[3] A. R. Barron and C.-H. Sheu. Approximation of density functions by sequences of exponential families. *Annals of Statistics*, 19(3):1347–1369, 1991.

[4] P. Billingsley. *Probability and Measure.* Wiley Series in Probability and Mathematical Statistics. Wiley, 1986.

[5] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[6] J. Bretagnolle and C. Huber. Estimation des densités: Risque minimax. *Z. Wahrsch. Verw. Gebiete*, 47:119–137, 1979.

[7] K. L. Chung. An estimate concerning the Kolmogorov limit distribution. *Transactions of the American Mathematical Society*, 67:36–50, 1949.

[8] E. Csáki. An iterated logarithm law for semimartingales and its application to the empirical distribution function. *Studia Sci. Math. Hung.*, 3:287–292, 1968.

[9] L. Devroye and L. Györfi. No empirical probability measure can converge in total variation sense for all distributions. *Annals of Statistics*, 18:1496–1499, 1990.

[10] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* J. Wiley & Sons, 1973.

[11] S. Y. Efroimovich and M. S. Pinsker. Estimation of square integrable probability density of a random variable. *Problems Inform. Transmission*, 18:175–189, 1983.

[12] H. Finkelstein. The law of the iterated logarithm for empirical distributions. *Annals of Mathematical Statistics*, 42(2):607–615, 1971.

[13] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[14] K. Hornik, M. Stinchcombe, and H. White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3:551–560, 1990.

[15] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[16] D. W. Scott. *Multivariate Density Estimation, Theory, Practice and Visualization.* John Wiley & Sons, Inc., New York, 1988.

[17] R. Serfling. *Approximation Theorems of Mathematical Statistics.* Wiley Series in Probability and Mathematical Statistics. Wiley, 1980.

[18] B. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, UK, 1993.

[19] N. V. Smirnov. An approximation to the distribution laws of random quantiles detremined by empirical data. *Uspehi Mat. Nauk*, 10:179–206, 1944.

[20] C. J. Stone. Optimal rates of convergence for non-parametric estimators. *Annals of Statistics*, 8:1348–1360, 1980.

[21] C. J. Stone. Optimal global rates of convergence for non-parametric regression. *Annals of Statistics*, 10:1040–1053, 1982.

[22] C. J. Stone. Optimal uniform rate of convergence for non-parametric estimators of a density function or its derivatives. In M. H. Rezvi, J. S. Rustagi, and D. Siegmund, editors, *Recent Advances in Statistics: Papers in honor of Herman Chernoff on His Sixtieth Birthday*, pages 393–406, New York, 1983. Academic.

[23] C. J. Stone. Large sample inference for log-spline models. *Annals of Statistics*, 18(2):717–741, 1990.

[24] A. N. Tikhonov and V. I. Arsenin. *Solutions of Ill-Posed Problems.* Scripta Series in Mathematics. Distributed solely by Halsted Press, Winston; New York, 1977. Translation Editor: Fritz, John.