EECS 598: Statistical Learning Theory, Winter 2014 **Topic 19**

Kernel Density Estimation

Lecturer: Clayton Scott *Scribe: Yun Wei, Yanzhen Deng*

1 Introduction

Let f be a density on $\mathbb{R}^d$, i.e. $f \geq 0$ and $\int f(x)dx = 1$. Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f$. Let ϕ be a function s.t. $\int \phi(x)dx = 1$, called a *kernel*, and denote

$$\phi_\sigma(x) := \frac{1}{\sigma^d}\phi(\frac{x}{\sigma})$$

for $\sigma > 0$. σ is called the *bandwidth*. The *kernel density estimator* (KDE) is

$$\widehat{f}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \phi_\sigma(x - X_i).$$

Example. 1) Gaussian kernel: $\phi(x) = (2\pi)^{-\frac{d}{2}} e^{-\frac{\|x\|^2}{2}}$
2) There are some common kernels like triangle kernel and box kernel. See Figure 1 for their graph in one dimension.

2 L^p Space

For $f : \mathbb{R} \to \mathbb{R}$ and $0 < p < \infty$, define

$$\|f\|_p = (\int |f(x)|^p \, dx)^{\frac{1}{p}}$$

and

$$L^p = \{f \mid \|f\|_p < \infty\}.$$

If $p \geq 1$ and we identify f and g when $\|f - g\|_p = 0$ (thus defining equivalence classes) then L^p is a normed vector space, where the triangle inequality is given by Minkowski's Inequality. For a full development, see [1].

Definition 1 (Convolution)**.** *Given f, g, the convolution $f * g$ is the function*

$$f * g(x) = \int f(y)g(x - y)dy = \int g(y)f(x - y)dy.$$

Young's Inequality shows that the convolution of L^1 functions is still an L^1 function.

Lemma 1 (Young's Inequality)**.** *If $f, g \in L^1$, then $f * g \in L^1$ and $\|f * g\|_1 \leq \|f\|_1 \|g\|_1$.*

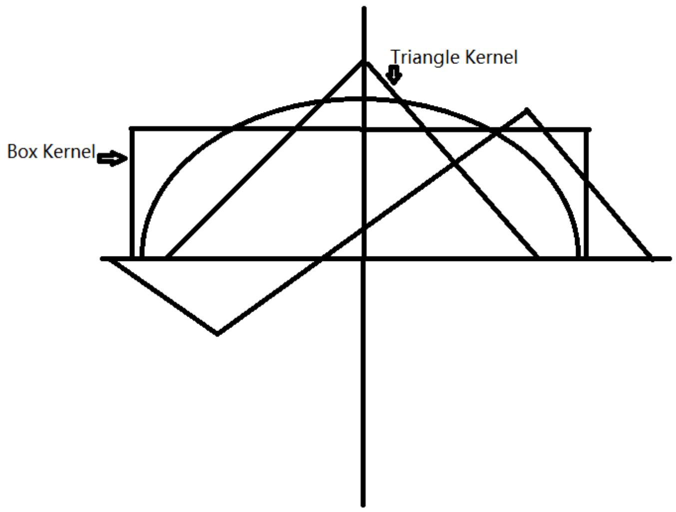

Figure 1: This is a picture showing some kernels. In addition to the uniform and triangular kernels, the third example is an arbitrary kernel illustrating that a kernel need not be 1) positive, or 2) symmetric.

Proof.

$$\begin{aligned}
\|f * g\|_1 &= \int |f * g(x)| dx \\
&= \int | \int f(y) g(x-y) dy | dx \\
&\leq \int (\int |f(y) g(x-y)| dy) dx \\
&= \int |f(y)| (\int |g(x-y)| dx) dy \quad \text{(By Tonelli Theorem)} \\
&= \int |f(y)| \|g\|_1 dy \quad \text{(By substitution: } u = x - y\text{)} \\
&= \|f\|_1 \|g\|_1 .
\end{aligned}$$

□

We state the next result without proof.

Theorem 1 (See Folland, Thm 8.14)**.** *Let $f \in L^p$, and $\phi \in L^1$ with $\int \phi(x) dx = a$. Then for any $r > 0$, $f * \phi_r \in L^p$ and*

$$\lim_{r \downarrow 0} \|f * \phi_r - af\|_p = 0.$$

3 L^2 consistency

Theorem 2. *Let $f \in L^2$ be a density, $\phi \in L^1 \bigcap L^2$ with $\int \phi(x)dx = 1$. Assume $X_1, X_2, \dots, X_n \overset{iid}{\sim} f$. If $\sigma \to 0$ and $n\sigma^d \to \infty$ as $n \to \infty$, then*

$$\|\widehat{f}_n - f\|_2 \overset{i.p.}{\to} 0.$$

Proof. By the triangle inequality,

$$\|\widehat{f}_n - f\|_2 \le \|\widehat{f}_n - f * \phi_\sigma\|_2 + \|f * \phi_\sigma - f\|_2.$$

The second term $\to 0$ as $\sigma \to 0$, by Theorem 1, since $f \in L^2, \phi \in L^1$. The first term converges i.p. to zero according to Lemma 2. □

Lemma 2. *If f is a density, $\phi \in L^2$, and $X_1, X_2, \dots, X_n \overset{iid}{\sim} f$, then*

$$\|\widehat{f}_n - f * \phi_\sigma\|_2 \overset{i.p.}{\to} 0$$

provided $n\sigma^d \to \infty$.

Proof. Observe

$$\begin{aligned}
\Pr\{\|\widehat{f}_n - f * \phi_\sigma\|_2 > \epsilon\} &= \Pr\{\|\widehat{f}_n - f * \phi_\sigma\|_2^2 > \epsilon^2\} \\
&\le \mathbb{E}\{\|\widehat{f}_n - f * \phi_\sigma\|_2^2\}/\epsilon^2
\end{aligned}$$

by Markov's Inequality. So it suffices to show $\mathbb{E}\{\|\widehat{f}_n - f * \phi_\sigma\|_2^2\} \to 0$. Note $\mathbb{E}$ is an integral operator, and therefore by Tonelli's Theorem we can interchange the order of integration:

$$\mathbb{E}\{\|\widehat{f}_n - f * \phi_\sigma\|_2^2\} = \int \mathbb{E}\{(\widehat{f}_n(x) - f * \phi_\sigma(x))^2\}dx.$$

Write

$$\widehat{f}_n(x) - f * \phi_\sigma(x) = \frac{1}{n}\sum_{i=1}^n Z_i,$$

where $Z_i = \phi_\sigma(x - X_i) - f * \phi_\sigma(x)$. Note that Z_i are iid and $\mathbb{E}(Z_i) = 0$ because

$$\mathbb{E}\phi_\sigma(x - X_i) = \int \phi_\sigma(x - x_i) f(x_i) dx_i = f * \phi_\sigma(x).$$

The variance of Z_i is

$$\begin{aligned}
\mathbb{E}(Z_i^2) &= \mathrm{Var}(\phi_\sigma(x - X_i)) \\
&= \mathbb{E}\{(\phi_\sigma(x - X_i))^2\} - [\mathbb{E}\{\phi_\sigma(x - X_i)\}]^2 \\
&\le \mathbb{E}\{(\phi_\sigma(x - X_i))^2\} \\
&= \int f(y)\phi_\sigma(x - y)^2 dy \\
&= f * \phi_\sigma^2(x) \\
&= \frac{1}{\sigma^d} f * (\phi^2)_\sigma(x),
\end{aligned}$$

where the last step follows from the fact $\phi_\sigma^2(x) = [\frac{1}{\sigma^d}\phi(\frac{x}{\sigma})]^2 = \frac{1}{\sigma^d}[\frac{1}{\sigma^d}\phi^2(\frac{x}{\sigma})] = \frac{1}{\sigma^d}(\phi^2)_\sigma$. Thus

$$\mathbb{E}\{(\frac{1}{n}\sum_{i=1}^n Z_i)^2\} = \frac{1}{n}\mathbb{E}\{Z_1^2\} \leq \frac{1}{n\sigma^d} f * (\phi^2)_\sigma(x)$$

and therefore

$$\begin{aligned}
\int \mathbb{E}\{(\widehat{f}_n(x) - f * \phi_\sigma(x))^2\}dx &\leq \frac{1}{n\sigma^d}\int f * (\phi^2)_\sigma(x)dx \\
&= \frac{1}{n\sigma^d}\|f * (\phi^2)_\sigma(x)\|_1 \\
&\leq \frac{1}{n\sigma^d}\|f\|_1\|(\phi^2)_\sigma\|_1 \quad \text{(by Young's Inequality)} \\
&= \frac{1}{n\sigma^d}\|\phi\|_2^2 \qquad (\|f\|_1 = 1) \\
&\to 0,
\end{aligned}$$

since $n\sigma^d \to \infty$ and $\phi \in L^2$. □

Remark.
(1) The condition $f \in L^2$ excludes certain densities such as

$$f(x) = \frac{1}{1-r}x^{-r},\ 0 < x < 1,$$

where $\frac{1}{2} < r < 1$.
(2) $\phi \in L^2$ is satisfied by all common kernels.

(3) Recall ϕ need not be symmetric w.r.t. the origin. Thus, the consistency result holds for

$$\phi(x) = \mathbf{1}_{\{x \in B(x_0, r)\}},$$

where

$$B(x_0, r) = \{x : \|x - x_0\|_2 \leq r\},$$

r is s.t. $\int \phi(x)dx = 1$, and $x_0 = (0, 0, \cdots, 0, 10^{100})$. This may seem bizarre, but as an exercise you are asked to make sense of this example.

4 L^1-Consistency

In this section we will show that the L^1 error converges to 0 in probability. To keep things simpler, we will assume f has compact support, although this is not necessary for L^1 consistency.

Theorem 3. *If f is a density with compact support, $\phi \in L^1$ s.t. $\int \phi(x)dx = 1$, and $X_1, X_2, ..., X_n \overset{i.i.d.}{\sim} f$, then*

$$\|\widehat{f}_n - f\|_1 \xrightarrow{i.p.} 0,$$

provided that $\sigma \to 0$ and $n\sigma^d \to 0$ and $n \to \infty$.

Proof. Note that

$$\|\widehat{f}_n - f\|_1 \leq \|\widehat{f}_n - f * \phi_\sigma\|_1 + \|f * \phi_\sigma - f\|_1.$$

By Theorem 1, we know that $\|f * \phi_\sigma - f\|_1 \to 0$, so it remains to show convergence to zero of $\|\widehat{f}_n - f * \phi_\sigma\|_1$.

Let $\mathcal{C}_c = \{g : \mathbb{R}^d \to \mathbb{R} \,|\, g \text{ is bounded and has compact support}\}$. It is a well-known fact in analysis [1] that $\mathcal{C}_c$ is dense in L^1. Thus for any fixed $\epsilon > 0$, we can take $\psi \in \mathcal{C}_c$ s.t. $\|\phi - \psi\|_1 < \epsilon$. Denote $\widehat{f}_n^c(x) = \frac{1}{n}\sum_{i=1}^n \psi_\sigma(x - X_i)$. Note that

$$\|\widehat{f}_n - f * \phi_\sigma\|_1 \le \|\widehat{f}_n - \widehat{f}_n^c\|_1 + \|\widehat{f}_n^c - f * \psi_\sigma\|_1 + \|f * \psi_\sigma - f * \phi_\sigma\|_1.$$

By Young's Inequality,

$$\|f * \psi_\sigma - f * \phi_\sigma\|_1 = \|f * (\psi_\sigma - \phi_\sigma)\|_1 \le \|f\|_1 \|\psi_\sigma - \phi_\sigma\| = \|\psi - \phi\|_1 < \epsilon.$$

The first term is bounded by

$$\|\widehat{f}_n - \widehat{f}_n^c\|_1 \le \frac{1}{n}\sum_{i=1}^n \|\psi_\sigma(x - X_i) - \phi_\sigma(x - X_i)\|_1 < \epsilon.$$

Since ϵ is arbitrary, we only need to prove that $\|\widehat{f}_n^c - f * \psi_\sigma\|_1 \to 0$ *i.p.* Denote by S_f and S_ψ the supports of f and ψ, respectively. We know that S_f and S_ψ are both compact sets and S_{ψ_σ} is also compact and shrinks as $\sigma \to 0$. Thus,

$$\begin{aligned}
\|\widehat{f}_n^c - f * \psi_\sigma\|_1 &= \int_{S_f \cup S_{\psi_\sigma}} |\widehat{f}_n^c - f * \psi_\sigma| \, dx \\
&= \int_{S_f \cup S_\psi} |\widehat{f}_n^c - f * \psi_\sigma| \, dx && \text{(for } \sigma \le 1) \\
&= \int |\widehat{f}_n^c - f * \psi_\sigma| \mathbf{1}_{S_f \cup S_\psi} dx \\
&\le \|\widehat{f}_n^c - f * \psi_\sigma\|_2 \|\mathbf{1}_{S_f \cup S_\psi}\|_2. && \text{(Hölder's Inequality)}
\end{aligned}$$

The second equality holds when $\sigma < 1$, which implies $S_f \cup S_{\psi_\sigma} \subseteq S_f \cup S_\psi$. Since ψ is in L^1 and bounded with compact support, it is also in L^2. Thus by Lemma 2, $\|\widehat{f}_n^c - f * \psi_\sigma\|_2 \to 0$ *i.p.* Now $\|\mathbf{1}_{S_f \cap S_\psi}\|_2$ is the square root of the volume of a compact set and thus is finite. Therefore $\|\widehat{f}_n^c - f * \psi_\sigma\|_1 \to 0$ *i.p.* □

Remark. The reason that we care about L^1 error is the following equality called Scheffe's Identity: if f, g are densities and $\mathcal{B}$ is the set of Borel sets, then:

$$\begin{aligned}
\|f - g\|_1 &= \int_{f>g} (f - g)(x)dx - \int_{f<g} (g - f)(x)dx \\
&= \int_{f>g} (f - g)(x)dx - [\int (g - f)(x)dx - \int_{f>g} (g - f)(x)dx] \\
&= 2\int_{f>g} (f - g)(x)dx \\
&= 2\sup_{B \in \mathcal{B}} |\int_B f(x)dx - \int_B g(x)dx|
\end{aligned}$$

Scheffe's Identity shows that small L^1 error leads to accurate probability estimation.

5 Strong Consistency

If we add the constraint that the kernel be nonnegative, then weak L^1 consistency implies strong L^1 consistency.

Theorem 4. *Assume* $\phi \geq 0$ *and* $\int \phi(x)dx = 1$*. If* $X_1, X_2, ..., X_n \overset{i.i.d.}{\sim} f$*, then* $\|\widehat{f}_n - f\|_1 \to 0$ *i.p. implies* $\|\widehat{f}_n - f\|_1 \to 0$ *a.s.*

Proof. Let $S = (X_1, ..., X_n)$ and $S'_i = (X_1, ..., X_{i-1}, X'_i, X_{i+1}, ..., X_n)$. Write $\widehat{f}_n = \widehat{f}_{n,S}$, using the new subscript to indicate the sample. Denote $\phi_n(S) = \|\widehat{f}_{n,S} - f\|_1$. Then

$$\begin{aligned}
|\phi_n(S) - \phi_n(S'_i)| &\leq \|\widehat{f}_{n,S} - \widehat{f}_{n,S'_i}\|_1 && \text{(reverse triangle inequality)} \\
&= \frac{1}{n}\int |\phi_\sigma(x - X_i) - \phi_\sigma(x - X'_i)|dx \\
&\leq \frac{1}{n}\int |\phi_\sigma(x - X_i)|dx + \int |\phi_\sigma(x - X'_i)|dx \\
&= \frac{2}{n}. && (\phi \text{ nonnegative})
\end{aligned}$$

By the bounded difference inequality,

$$\Pr(\phi_n(S) - \mathbb{E}[\phi_n(S)] \geq \epsilon) \leq e^{-n\epsilon^2/2}.$$

Fix $\epsilon > 0$. By weak consistency, $\exists N$ s.t. $n \geq N \Rightarrow \mathbb{E}\phi_n(S) < \frac{\epsilon}{2}$. Then for $n \geq N$,

$$\Pr(\phi_n(S) \geq \epsilon) \leq \Pr(\phi_n(S) - \mathbb{E}\phi_n(S) \geq \epsilon/2) \leq e^{-n\epsilon^2/8}.$$

This upper bound decrease geometrically. Therefore

$$\sum_{n=1}^{\infty} \Pr(\phi_n(S) \geq \epsilon) < \infty$$

and Borel-Cantelli implies $\phi_n(S) \to 0$ *a.s.* □

Exercises

1. Make sense of the third remark after the proof of L^2 consistency.
2. What does Bernstein's inequality imply about $\frac{1}{n}\sum Z_i$ in the proof of Lemma 2? Is this observation useful in any way?
3. Remove the assumption in Theorem 3 that f has compact support.

References

[1] Gerald B. Folland, *Real Analysis: Modern Techniques and Their Applications*, Wiley, 1999