Approximation of kernel density estimator

Given a data set $\{x_i\}_{i=1}^n$, the purpose of the approximation is to compute

$$f(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left(-\frac{(x-x_i)^2}{2h^2}\right)$$

fast. To do this, we want to bin the data into bins

$$\mathcal{A}_j = \{x_i\}_{i=1}^n \cap [\min_{i=1,\dots,n} x_i,\ j \cdot \alpha h + \min_{i=1,\dots,n} x_i]$$

where $j = 1, \dots, J$ so that $\bigcup_{j=1}^J \mathcal{A}_j$ cover the range of the data. The bin width is αh, where $\alpha < (\sqrt{2}-1)/2$. For each bin $\mathcal{A}$ we replace

$$f_{\mathcal{A}}(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{x_i \in \mathcal{A}} \exp\left(-\frac{(x-x_i)^2}{2h^2}\right)$$

with

$$\tilde{f}_{\mathcal{A}}(x) = \frac{|\mathcal{A}|}{nh\sqrt{2\pi}} \exp\left(-\frac{(x-x_{\mathcal{A}})^2}{2h^2}\right)$$

where $\mathcal{A} = \{x_i\}_{i=1}^n \cap [x_0, x_0 + \alpha h]$ with $\alpha < 1$ and $x_{\mathcal{A}}$ is the mean of $\mathcal{A}$. Below we will show that the maximal error times h is bounded by $\alpha^2/(2\sqrt{2\pi})$.

Without loss of generality we assume that $x_{\mathcal{A}} = 0$ (otherwise replace each x_i with $x_i - x_{\mathcal{A}}$). Let

$$g(x,y) = \exp\left(-\frac{(x-y)^2}{2h^2}\right).$$

Then

$$\frac{dg}{dy}(x,y) = \frac{1}{h^2}(x-y)\exp\left(-\frac{(x-y)^2}{2h^2}\right)$$

and

$$\frac{d^2g}{dy^2} = \frac{1}{h^4}\left((x-y)^2 - h^2\right)\exp\left(-\frac{(x-y)^2}{2h^2}\right).$$

Using a MacLaurin expansion this gives that

$$\begin{aligned} g(x,x_i) = \exp\left(-\frac{x^2}{2h^2}\right) + \frac{x_i x}{h^2}\exp\left(-\frac{x^2}{2h^2}\right) \\ + \frac{x_i^2}{2h^4}\left((x-\xi_i)^2 - h^2\right)\exp\left(-\frac{(x-\xi_i)^2}{2h^2}\right) \end{aligned}$$

for some ξ_i such that $|\xi_i| < |x_i|$ and $\text{sign}(\xi_i) = \text{sign}(x_i)$. Thus

$$\begin{aligned}
nh\sqrt{2\pi}|(f_{\mathcal{A}}(x) - \tilde{f}_{\mathcal{A}}(x))| &= |\sum_{x_i \in \mathcal{A}} g(x, x_i) - |\mathcal{A}| \exp\left(-\frac{x^2}{2h^2}\right)| \\
&= |\sum_{i\colon x_i \in \mathcal{A}} \frac{x_i^2}{2h^4}\left((x-\xi_i)^2 - h^2\right)\exp\left(-\frac{(x-\xi_i)^2}{2h^2}\right)| \\
&\leq \sum_{i\colon x_i \in \mathcal{A}} \frac{x_i^2}{2h^4}\max((x-\xi_i)^2 - h^2, h^2)\exp\left(-\frac{(x-\xi_i)^2}{2h^2}\right)
\end{aligned}$$

Note that $x_{\mathcal{A}} = 0$ implies that $0 \in [x_0, x_0 + \alpha h]$, which implies that $|x_i| < \alpha h$ for $x_i \in \mathcal{A}$, and thus $|\xi_i| < \alpha h$.

We now have two cases to consider: 1) $\max((x-\xi_i)^2 - h^2, h^2) = h^2$, and 2) $\max((x-\xi_i)^2 - h^2, h^2) = (x-\xi_i)^2 - h^2$. We will show that in both cases, $nh\sqrt{2\pi}|f_{\mathcal{A}}(x) - \tilde{f}_{\mathcal{A}}(x)|$ is bounded by $|\mathcal{A}|\alpha^2/2$, as long as $\alpha < (\sqrt{2}-1)/2 \approx 0.2$.

Case 1): If $\max((x-\xi_i)^2 - h^2, h^2) = h^2$ we get

$$nh\sqrt{2\pi}|f_{\mathcal{A}}(x) - \tilde{f}_{\mathcal{A}}(x)| \leq \sum_{x_i \in \mathcal{A}} \frac{x_i^2}{2h^2} \leq |\mathcal{A}|\frac{\alpha^2}{2}.$$

Case 2): Now we assume that $\max((x-\xi_i)^2 - h^2, h^2) = (x-\xi_i)^2 - h^2$. This means that

$$\begin{aligned}
&\sum_{i\colon x_i \in \mathcal{A}} \frac{x_i^2}{2h^4}\max(((x-\xi_i)^2 - h^2), h^2)\exp\left(-\frac{(x-\xi_i)^2}{2h^2}\right) \\
&\qquad \leq \sum_{i\colon x_i \in \mathcal{A}} \frac{x_i^2}{2h^4}((x-\xi_i)^2 - h^2)\exp\left(-\frac{(x-\xi_i)^2}{2h^2}\right) \\
&\qquad \leq \sum_{i\colon x_i \in \mathcal{A}} \frac{x_i^2}{2h^4}((|x| + \alpha h)^2 - h^2)\exp\left(-\frac{(|x| - \alpha h)^2}{2h^2}\right) \\
&\qquad = |\mathcal{A}|\frac{\alpha^2}{2}((|x|/h + \alpha)^2 - 1)\exp\left(-\frac{(|x|/h - \alpha)^2}{2}\right).
\end{aligned}$$

That $\max((x-\xi_i)^2 - h^2, h^2) = (x-\xi_i)^2 - h^2$ can only occur if $|x| \geq (\sqrt{2} - \alpha)h$. If we further assume that $\alpha < (\sqrt{2}-1)/2$ we also have that $|x|/h - \alpha \geq 1$.

Let $q(y) = ((y+\alpha)^2 - 1)\exp(-(y-\alpha)^2/2)$. If we can show that $q(y) \leq 1$ when $y \geq \sqrt{2} - \alpha$ and $\alpha < (\sqrt{2}-1)/2$ we are done. First note that under these conditions we also have that $y - \alpha > 1$.

Now

$$q'(y) = \left(2(y+\alpha) - (y-\alpha)((y+\alpha)^2 - 1)\right)\exp(-\frac{(y-\alpha)^2}{2}).$$

Let $r(w) = 2(w + 2\alpha) - w((w + 2\alpha)^2 - 1)$ so that

$$q'(y) = r(y - \alpha)\exp(-(y - \alpha)^2/2).$$

Clearly

$$q'(y) < 0 \Leftrightarrow r(y - \alpha) < 0.$$

Furthermore,

$$r'(w) = -3w^2 - 8\alpha w + (3 - 4\alpha^2) = -3(w + 4\alpha/3)^2 + 3 + 4\alpha^2/3$$

so $r'(w) < 0$ if $w + 4\alpha/3 > \sqrt{1 + 4\alpha^2/9}$. Thus $r'(y - \alpha) < 0$ if $y + \alpha/3 > \sqrt{1 + 4\alpha^2/9}$, which holds as long as $y - \alpha > 1$ since then

$$y + \alpha/3 > 1 + 4\alpha/3 > 1 + 2\alpha/3 > \sqrt{1 + 4\alpha^2/9},$$

where the last inequality comes from the triangle inequality.

This means that in the range of interest, $r(y - \alpha)$ is strictly decreasing, which means that $q'(y) > 0 \Leftrightarrow y < y_0$ for some y_0 and the maximum of $q(y)$ is attained at $q(y_0)$. Furthermore,

$$\begin{aligned} r(y - \alpha) = 2(y + \alpha) - (y - \alpha)((y + \alpha)^2 - 1) &\leq 2(y + \alpha) - ((y + \alpha)^2 - 1) \\ &= -((y + \alpha)^2 - 2(y + \alpha) + 1) + 2 = -(y + \alpha - 1)^2 + 2 \end{aligned}$$

is less than zero when $y > \sqrt{2} + 1 - \alpha$, so $y_0 \leq \sqrt{2} + 1 - \alpha$. Letting $y = t + \sqrt{2} - \alpha$ it is sufficient to show that $q(t + \sqrt{2} - \alpha)$ is bounded by 1 for $t \in [0, t_0]$ with $t_0 = y_0 - \sqrt{2} - \alpha \leq 1$.

We have that for $t \in [0, t_0]$ and $\alpha < (\sqrt{2} - 1)/2$

$$\begin{aligned} q'(t + \sqrt{2} - \alpha) &\leq [2(t + \sqrt{2}) - (t + \sqrt{2} - 2\alpha)((t + \sqrt{2})^2 - 1)]\exp(-1/2) \\ &< [2(t + \sqrt{2}) - (t + 1)((t + \sqrt{2})^2 - 1)]\exp(-1/2) \\ &= [-t^3 - (1 + 2\sqrt{2})t^2 - (2\sqrt{2} - 1)t + (2\sqrt{2} - 1)]\exp(-1/2) \end{aligned}$$

so

$$\begin{aligned} q(t_0 + \sqrt{2} - \alpha) &= q(\sqrt{2} - \alpha) + \int_0^{t_0} q'(s + \sqrt{2} - \alpha)\,ds \\ &\leq \exp(-\frac{1}{2})\left[1 + \int_0^{t_0} (-s^3 - (1 + 2\sqrt{2})s^2 - (2\sqrt{2} - 1)s + (2\sqrt{2} - 1))\,ds\right] \\ &= \exp(-\frac{1}{2})\left[1 - \frac{1}{4}t_0^4 - \frac{1 + 2\sqrt{2}}{3}t_0^3 - \frac{2\sqrt{2} - 1}{2}t_0^2 + (2\sqrt{2} - 1)t_0\right], \end{aligned}$$

which is easily numerically verified to be less than 1 for $0 \leq t_0 \leq 1$.

We have now shown that in both cases above

$$nh\sqrt{2\pi}|f_{\mathcal{A}}(x) - \tilde{f}_{\mathcal{A}}(x)| \leq |\mathcal{A}|\alpha^2/2,$$

thus

$$h|f(x) - \sum_{\mathcal{A}_j} \tilde{f}_{\mathcal{A}_j}(x)| \leq \frac{\alpha^2}{2\sqrt{2\pi}},$$

where

$$\mathcal{A}_j = \{x_i\}_{i=1}^n \cap [\min_{i=1,\ldots,n} x_i,\ j \cdot \alpha h + \min_{i=1,\ldots,n} x_i]$$

and $j = 1, \ldots, J$ so that $\bigcup_{j=1}^J \mathcal{A}_j$ cover the range of $\{x_i\}_{i=1}^n$.