Statistica Sinica Preprint No: SS-2015-0365.R2	
Title	On the Efficiency of Online Approach to Nonparametric Smoothing of Big Data
Manuscript ID	SS-2015-0365.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0365
Complete List of Authors	Efang Kong and Yingcun Xia
Corresponding Author	Yingcun Xia
E-mail	staxyc@nus.edu.sg
Notice: Accepted version subject to English editing.	

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

Lemma 2.2 *Under (A1) and (A2), we have*

$$AMSE(\tilde{f}_n(\mathbf{x}|\tilde{h}_n)) = \frac{c^4}{4(1-2\alpha)^2}[tr\{\mathcal{H}_f(\mathbf{x})\}]^2 n^{-4\alpha} + \frac{1}{(1+p\alpha)c^p} f(\mathbf{x}) R_2(K) n^{p\alpha-1},$$

which is minimized at

$$\alpha = 1/(4+p), \quad c = \Big(\frac{p(p+2)}{2(p+4)}\Big)^{1/(4+p)} \Big(\frac{R_2(K)f(\mathbf{x})}{[tr\{\mathcal{H}_f(\mathbf{x})\}]^2}\Big)^{1/(4+p)},$$

with minimum

$$\frac{(p+4)^2}{8(p+2)}\Big(\frac{p(p+2)}{2(p+4)}\Big)^{-p/(4+p)} n^{-4/(4+p)} [f(\mathbf{x})R_2(K)]^{4/(4+p)} [tr\{\mathcal{H}_f(\mathbf{x})\}]^{2p/(4+p)}.$$

From Lemma 2.1, for OFFLINE Parzen-Rosenblatt estimator with bandwidth $ch^{-\alpha}$, the optimal values for α and c are

$$\alpha = 1/(p+4), \quad c = \Big\{\frac{f(\mathbf{x})R_2(K)}{tr\{\mathcal{H}_f(\mathbf{x})\}^2}\Big\}^{1/(4+p)} p^{1/(4+p)}. \tag{2.6}$$

So, while the optimal choices of α are identical for OFFLINE and ONLINE, their respective optimal choices for the coefficient c do differ, with the ratio given by

$$\frac{\text{optimal } \tilde{h}_n}{\text{optimal } h_n} = \Big(\frac{p+2}{2(p+4)}\Big)^{1/(4+p)}, \tag{2.7}$$

which is always less than 1. A ratio larger than one is expected, for otherwise, the use of index-specific bandwidths $\tilde{h}_i \propto i^{-\alpha}$ would result in too large a bias for ONLINE, which cannot be compensated for by the accompanying reduction in the variance.

The relative efficiency of the ONLINE Parzen-Rosenblatt estimator against its OFFLINE counterpart

$$\frac{AMSE(\tilde{f}_n(\mathbf{x})|\text{optimal } h_n)}{AMSE(\tilde{f}_n(\mathbf{x})|\text{optimal } \tilde{h}_n)} = 2^{4/(4+p)}\Big(\frac{p+2}{p+4}\Big)^{(2p+4)/(p+4)}. \tag{2.8}$$

This starts at one for $p = 1$, drops to its lowest level of 0.9186 at $p = 4$, and then slowly increases to 1 as $p \to \infty$; see Figure 1.

The nonlinear pattern in the relative efficiency (2.8) versus dimensionality is a result of several confounding factors. These include, for example, the use of a common bandwidth by OFFLINE in contrast to the use of index-specific bandwidths by ONLINE. Also, that the OFFLINE has smaller bias but larger variance, yet the exact opposite holds true for ONLINE. For both OFFLINE and ONLINE, as the dimension increases, the so-called 'curse of