A Short Note on the L_∞ Concentration of the KDE

Yen-Chi Chen
Department of Statistics
University of Washington

The concentration inequality of the kernel density estimator (KDE) from Giné and Guillou (2002) suggests

$$P(\|\widehat{p}_n - \mathbb{E}(\widehat{p}_n)\|_\infty > \epsilon) \leq c_1 e^{-c_2 \cdot nh^d \cdot \epsilon^2}$$

for some constants $c_1, c_2 > 0$. This seems to be inconsistent with other results (see, e.g, Einmahl and Mason 2005; Genovese et al. 2014): $\|\widehat{p}_n - \mathbb{E}(\widehat{p}_n)\|_\infty = O_P\left(\sqrt{\frac{|\log h|}{nh^d}}\right)$. We point out that this concentration inequality *is consistent* with others and the key reason is that the concentration works only if $\epsilon \geq \sqrt{\frac{|\log h|}{nh^d}}$. The lower bound on ϵ, though converges to 0, enforces the convergence rate from the concentration inequality to $O_P\left(\sqrt{\frac{|\log h|}{nh^d}}\right)$, which is consistent with other findings.

1. Main Result

Let $X_1, \cdots, X_n$ be an IID random sample from an unknown density function p with a compact support $\mathbb{K} \subset \mathbb{R}^d$. The kernel density estimator of p is

$$\widehat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right),$$

where K is a smooth function (known as the kernel function) such as the Gaussian and $h > 0$ is the smoothing parameter that controls the amount of smoothing.

Here we focus on the uniform loss (L_∞ error) of $\widehat{p}_n$ from its expectation:

$$\Delta_n = \sup_x |\widehat{p}_n(x) - \mathbb{E}(\widehat{p}_n(x))| = \|\widehat{p}_n - \mathbb{E}(\widehat{p}_n)\|_\infty$$

This quantity is the uniform deviation of $\widehat{p}_n$ from its expected value and it plays a key role in constructing confidence bands of the density function p.

There are three important results about Δ_n.

(LD) **Limiting distribution.** Bickel and Rosenblatt (1973); Rosenblatt et al. (1976) proved that Δ_n converges to an extreme value distribution after properly rescaling. One can also use the KMT approximation (Komlós et al., 1975, 1976) to obtain a similar result. Roughly speaking, they proved that (after rearranging) there exists a constant $A_1 > 0$ such that

$$\sqrt{nh^d}(\Delta_n - \sqrt{|\log h|} A_1) = O_P\left(\frac{1}{\sqrt{|\log h|}}\right),$$

which implies

$$\Delta_n = O\left(\sqrt{\frac{|\log h|}{nh^d}}\right) + O_P\left(\sqrt{\frac{1}{nh^d \cdot |\log h|}}\right). \tag{1}$$

(AS) **Almost sure convergence rate.** Another important result of Δ_n is Giné and Guillou (2002); Einmahl and Mason (2005), where the authors applied the Talagrand's inequality (Talagrand, 1994, 1996; Giné and Guillou, 2001) to the KDE and proved that under weak conditions, there exists a constant $C > 0$ such that

$$\sqrt{\frac{nh^d}{|\log h|}}\Delta_n = C \quad a.s.$$

This implies that

$$\Delta_n = O_{a.s.}\left(\sqrt{\frac{|\log h|}{nh^d}}\right). \tag{2}$$

Note that the same O_P rate has been derived in Yukich (1985).

(CI) **Concentration inequality.** When deriving the almost sure rate in Giné and Guillou (2002), the authors have implicitly proved a concentration inequality of Δ_n: when $h \to 0$, there exists $c_1, c_2 > 0$ such that

$$P(\Delta_n > \epsilon) \leq c_1 e^{-c_2 \cdot nh^d \cdot \epsilon^2} \tag{3}$$

for every

$$\epsilon \geq \sqrt{\frac{|\log h|}{nh^d}}. \tag{4}$$

Note that we use the version from the lecture note of CMU 36-702 (Statistical Machine Learning)[1], 2016 version.

Also note that because we often choose h to be a polynomial of n, $O(|\log h|) = O(\log n)$. So some literature (Genovese et al., 2014; Chen et al., 2015; Chen, 2016) replace $|\log h|$ by $\log n$. We now compare these three results.

(LD) and (AS): consistent. Intuitively, (AS) is consistent with (LD) because in (LD), the dominating quantity is $O\left(\sqrt{\frac{|\log h|}{nh^d}}\right)$, a deterministic sequence and the randomness is at rate $O_P\left(\sqrt{\frac{1}{nh^d \cdot |\log h|}}\right)$, which converges faster than the dominating one (though the rate difference is very slow: $O_P(|\log h|)$). Thus, one would expect that $\sqrt{\frac{nh^d}{|\log h|}}\Delta_n$ converges to a fixed quantity and the remaining stochastic fluctuation eventually die out.

(AS) and (CI): inconsistent (but this is incorrect!). When we compare (AS) to (CI), the result does not seem to be consistent at the first glance because in equation (3), the dependence of ϵ on n and h is through $nh^d\epsilon^2$. This seems to suggest that the rate will be $O_P(\sqrt{\frac{1}{nh^d}})$ by equating them to be a constant. However, this is *incorrect*! The main problem of the above derivation comes from the bound on ϵ. Equation (3) is correct *only if* $\epsilon \geq \sqrt{\frac{|\log h|}{nh^d}}$ (equation (4)).

1. http://www.stat.cmu.edu/~larry/=sml/

(AS) and (CI): consistent. The restriction on ϵ actually constrains the rate to be $O_P\left(\sqrt{\frac{|\log h|}{nh^d}}\right)$. To see this, we first rewrite equation (3) using $t^2 = nh^d\epsilon^2$:

$$\begin{aligned}
&P(\Delta_n > \epsilon) \leq c_1 e^{-c_2 \cdot nh^d \cdot \epsilon^2} \\
\Longrightarrow &P(\sqrt{nh^d}\Delta_n > \sqrt{nh^d}\epsilon) \leq c_1 e^{-c_2 \cdot nh^d \cdot \epsilon^2} \\
\Longrightarrow &P(\sqrt{nh^d}\Delta_n > t) \leq c_1 e^{-c_2 t^2},
\end{aligned}$$

when $t \geq \sqrt{|\log h|}$. Here you see that we cannot pick the right-hand-side arbitrarily small because of the lower bound on t. The above result directly leads to a bound on $\mathbb{E}(\sqrt{nh^d}\Delta_n)$:

$$\begin{aligned}
\mathbb{E}(\sqrt{nh^d}\Delta_n) &= \int_0^\infty P(\sqrt{nh^d}\Delta_n > t)dt \\
&= \int_{\sqrt{|\log h|}}^\infty P(\sqrt{nh^d}\Delta_n > t)dt + \int_0^{\sqrt{|\log h|}} P(\sqrt{nh^d}\Delta_n > t)dt \\
&\leq O(h^{-c_3}) + \int_0^{\sqrt{|\log h|}} 1dt \\
&= O(h^{-c_3}) + O(\sqrt{|\log h|}) = O(\sqrt{|\log h|}),
\end{aligned}$$

where c_3 is a positive constant. Thus, $\mathbb{E}(\Delta_n) = O\left(\sqrt{\frac{|\log h|}{nh^d}}\right)$ and by Markov's inequality

$$\Delta_n = O_P\left(\sqrt{\frac{|\log h|}{nh^d}}\right),$$

which agrees with the bounds from (LD) and (AS).

Take-home message. When using a concentration inequality to derive a convergence rate, we have to be careful about the range where the concentration holds. The convergence rate depends not only on how ϵ and n are associated but also on the valid range of ϵ.

References

PJ Bickel and M Rosenblatt. On some global measures of the deviations of density function estimates. *The Annals of Statistics*, 1(6):1071–1095, 1973.

Yen-Chi Chen. Generalized cluster trees and singular measures. *arXiv preprint arXiv:1611.02762*, 2016.

Yen-Chi Chen, Christopher R Genovese, and Larry Wasserman. Asymptotic theory for density ridges. *The Annals of Statistics*, 43(5):1896–1928, 2015.

Uwe Einmahl and David M Mason. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 2014.

Evarist Giné and Armelle Guillou. On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. In *Annales de l'IHP Probabilités et statistiques*, volume 37, pages 503–522, 2001.

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent rv'-s, and the sample df. i. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1975.

J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent rv's, and the sample df. ii. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1976.

M Rosenblatt et al. On the maximal deviation of k-dimensional density estimates. *The Annals of Probability*, 4(6):1009–1015, 1976.

M Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76, 1994.

Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

JE Yukich. Laws of large numbers for classes of functions. *Journal of multivariate analysis*, 17(3):245–260, 1985.