Graphical Nadaraya Watson estimator

Martin Gjorgjevski

May 2022

Contents

1 Motivation and introduction

In the classical nonparametric regression setting we are given data $X_1, ..., X_n \in \mathbb{R}^d$ i.i.d. with density p. We are also provided with noisy observations $Y_i = f(X_i) + \epsilon_i$ with $f : \mathbb{R}^d \to \mathbb{R}$ unknown and in some suitable class of functions and $\epsilon_1, ..., \epsilon_n$ are assumed to be i.i.d. centered Gaussian with variance σ^2. The goal is to estimate f. A popular approach for this task is the Nadaraya Watson estimator [Tsy08]

$$\hat{f}_{NW}(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i k(\frac{x-X_i}{h})}{\sum_{i=1}^n k(\frac{x-X_i}{h})} & \text{if } \sum_{i=1}^n k(\frac{x-X_i}{h}) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $k : \mathbb{R}^d \to \mathbb{R}$ is a kernel and $h > 0$ is a parameter known as bandwith.

In our setting we assume that the data $X, X_1, ..., X_n$ is latent, independent and X has possibly different distribution from $X_1, ..., X_n$ which are i.i.d., and in addition to the noisy observations $Y_1, ..., Y_n$ we observe a random graph associated with the data $X, X_1, ..., X_n$ generated as follows: for any two points x, y a Bernoulli variable $a(x, y)$ with parameter $k(x, y)$ determines whether there is an edge between x and y. Here, $k : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ is a kernel which measures similarity between two points. Intuitively this means that we are more likely to observe an edge between two variables that are similar with respect to k. Typically we are interested in the case when $X = x$ is deterministic or in the case where X has the same distribution as $X_1, ..., X_n$.

We are interested in estimating f in this setting. Inspired by the classical Nadaraya Watson estimator, we introduce the **Graphical Nadaraya Watson** estimator:

$$\hat{f}_{GNW}(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i a(x,X_i)}{\sum_{i=1}^n a(x,X_i)} & \text{if } \sum_{i=1}^n a(x, X_i) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

We introduce the expected connection parameter

$$c(x) = Ek(x, X_1) = \int k(x, z)p(z)dz$$

If $c(x) = 0$ then $k(x, X_i) = 0$ almost surely and consequently $\sum_{i=1}^n a(x, X_i) = 0$ almost surely, so $\hat{f}_{GNW}(x) = 0$. Thus in order to have nontrivial estimator μ almost surely, we need to assume $\int I(c(x) = 0)d\mu(x) = 0$[1]

In this report we are investigating the concentration and L^2 convergence properties of this estimator.

[1] This condition reads as $c(x) > 0$ when $\mu = \delta_x$ is a Dirac measure at x and $\int I(c(x) = 0)p(x)dx = 0$ when X has density p

2 Concentration properties

Lemma 1 Suppose that $f(X_1)$ is (essentially) bounded, measurable function, $||f(X_1)||_\infty \le B$. Then

$$P(|\frac{1}{n}\sum_{i=1}^{n} f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz| \ge t) \le 2\exp(-\frac{2t^2 n}{5B^2})$$

Proof. For $i = 1,...,n$ we can write $a(x,X_i) = I(U_i \le k(x,X_i))$ where U_i are i.i.d. uniform variables on $[0,1]$ independent from the $X_i's$ and $\epsilon_i's$. Define

$$F(x_1,...,x_n,u_1,...,u_n) = \frac{1}{n}\sum_{i=1}^{n}[f(x_i)I(u_i \le k(x,x_i)) - \int f(z)k(x,z)p(z)dz]$$

Note that $EF(X_1,...,X_n,U_1,...,U_n) = 0$. We will verify that F satisfies the hypothesis of McDiarmid's bounded difference inequality ([Ver18] Thm 2.9.1). Changing one of the $x_i's$ gives:

$$|F(x_1,...,x_i,...,x_n,u_1,...,u_n) - F(x_1,...,x_i^{'},...,x_n,u_1,...,u_n)| =$$
$$\frac{1}{n}|I(u_i \le k(x,x_i))f(x_i) - I(u_i \le k(x,x_i^{'}))f(x_i^{'})| \le \frac{2B}{n}$$

Changing one of the $u_i's$ gives:

$$|F(x_1,...,x_n,u_1,...u_i,...,u_n) - F(x_1,...,x_n,u_1,...u_i^{'},...,u_n)| =$$
$$\frac{1}{n}|[I(u_i \le k(x,x_i)) - I(u_i^{'} \le k(x,x_i))]f(x_i)| \le \frac{B}{n}$$

Hence F has the $(c_1,,,c_n,c_{n+1},...,c_{2n})$ bounded difference property with $c_1 = c_2 = ... = c_n = \frac{2B}{n}$ and $c_{n+1} = ... = c_{2n} = \frac{B}{n}$, giving $\sum_{i=1}^{2n} c_i^2 = \frac{5B^2}{n}$. The result now follows immediately from McDiarmid's inequality.

□

Lemma 2 Suppose that $w_1,...,w_n$ and $\epsilon_1,...,\epsilon_n$ are independent, $|w_i| \le 1$ and ϵ_i are centered Gaussian variables with variance σ^2. Then

$$P(|\frac{1}{n}\sum_{i=1}^{n} w_i\epsilon_i| \ge t) \le 2\exp(-\frac{3ct^2 n}{8\sigma^2})$$

where $c > 0$ is an absolute constant.

Proof. Consider the sub-gaussian norm of $w_1\epsilon_1$ defined as

$$||w_1\epsilon_1||_{\psi_2} = \inf\{t > 0 : E\exp(w_1\epsilon_1)^2/t^2) \le 2\}$$

We have

$$E\exp((w_1\epsilon_1)^2/t^2) \le E\exp(\epsilon_1^2/t^2) = \frac{1}{\sqrt{1-\frac{2\sigma^2}{t^2}}}$$

as soon as t is chosen such that $1 - \frac{2\sigma^2}{t^2} > 0$. Choosing $t = \sqrt{\frac{8\sigma^2}{3}}$ we get

$$E\exp((w_1\epsilon_1)^2/t^2) \le 2$$

In particular this shows that

$$||w_1\epsilon_1||_{\psi_2}^2 \le \frac{8\sigma^2}{3}$$

Using the General Hoeffding's inequality ([Ver18] Thm 2.6.3), we have

$$P(|\frac{1}{n}\sum_{i=1}^{n} w_i\epsilon_i| \ge t) \le 2\exp(-\frac{3ct^2 n}{8\sigma^2})$$

with $c > 0$ an absolute constant. This concludes the proof. □

Theorem 1 (Concetnration in the deterministic case) Suppose that $||f(X_1)||_\infty \leq B$ and $c(x) = Ek(x, X_1) = \int k(x,z)p(z)dz > 0$. Then for $0 < \delta < 3B$ and $H(B,\sigma^2) = \min\{\frac{1}{90B^2}, \frac{C}{\sigma^2}\}$ we have

$$|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| < \delta$$

with probability at least $1 - 6\exp(-H(B,\sigma^2)c(x)^2\delta^2 n)$.

Proof. Let $\delta > 0$ and denote

$$\begin{aligned}
A_\delta &= \{|\frac{1}{n}\sum_{i=1}^n f(x_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz| \geq \delta\} \\
B_\delta &= \{|\frac{1}{n}\sum_{i=1}^n a(x,X_i) - c(x)| \geq \delta\} \\
C_\delta &= \{|\frac{1}{n}\sum_{i=1}^n \epsilon_i a(x,X_i)| \geq \delta\}
\end{aligned}$$

Let $\delta_1, \delta_2, \delta_3 > 0$, to be specified later. Choosing $\delta_2 \leq \frac{1}{2}c(x)$, on $B^c_{\delta_2}$ we have $\frac{1}{n}\sum_{i=1}^n a(x,X_i) \geq \frac{1}{2}c(x)$ and in particular $\sum_{i=1}^n a(x,X_i) > 0$. Hence on $B^c_{\delta_2}$, we have

$$\begin{aligned}
\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{c(x)} &= \frac{\frac{1}{n}\sum_{i=1}^n Y_i a(x,X_i)}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} - \frac{\int f(z)k(x,z)p(z)dz}{c(x)} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n [f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz]}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} + \frac{\frac{1}{n}\sum_{i=1}\epsilon_i a(x,X_i)}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} \\
&+ \int f(z)k(x,z)p(z)dz[\frac{1}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} - \frac{1}{c(x)}]
\end{aligned} \tag{1}$$

In addition, on $(A_{\delta_1} \cup B_{\delta_2} \cup C_{\delta_3})^c$, we have

$$\begin{aligned}
|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{c(x)}| &\leq |\frac{\frac{1}{n}\sum_{i=1}^n [f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz]}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}| \\
&+ |\frac{\frac{1}{n}\sum_{i=1}\epsilon_i a(x,X_i)}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}| \\
&+ |\frac{\int f(z)k(x,z)p(z)dz}{c(x)} \frac{\frac{1}{n}\sum_{i=1}^n [a(x,X_i) - c(x)]}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}| \\
&\leq \frac{\delta_1 + \delta_3 + \delta_2 B}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} \\
&\leq \frac{2(\delta_1 + \delta_2 B + \delta_3)}{c(x)}
\end{aligned}$$

Finally, setting

$$\delta_1 = \delta_3 = \frac{\delta c(x)}{6}, \delta_2 = \frac{\delta c(x)}{6B}$$

we get

$$|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| \leq \delta$$

on $(A_{\delta_1} \cup B_{\delta_2} \cup C_{\delta_3})^c$.
By Lemma 1, we have $P(A_{\delta_1}) \leq 2\exp(-\frac{2\delta_1^2 n}{5B^2})$ and $P(B_{\delta_2}) \leq 2\exp(-\frac{2\delta_2^2 n}{5})$
By Lemma 2 we have $P(C_{\delta_3}) \leq 2\exp(-\frac{C\delta_3^2 n}{\sigma^2})$ where $C > 0$ is a constant.
Now

$$\begin{aligned}
P(A_{\delta 1} \cup B_{\delta_2} \cup C_{\delta_3}) &\leq P(A_{\delta_1}) + P(B_{\delta_2}) + P(C_{\delta_3}) \\
&\leq 6\exp(-H(B,\sigma^2)c(x)^2\delta^2 n)
\end{aligned}$$

which completes the proof. $\square$

Corollary 1 Suppose that $X, X_1, ..., X_n$ are i.i.d. with density p such that

$$\int_{\mathbb{R}^d} I(c(x) = 0)p(x)dx = 0$$

Then for any $r > 0$,

$$P(|\hat{f}_{GNW}(X) - \frac{\int f(z)k(X,z)p(z)dz}{\int k(X,z)p(z)dz}| \geq \delta) \leq 6\exp(-H(B,\sigma^2)r^2\delta^2 n) + 6P(\int K(X,z)p(z)dz < r)$$

Proof. Under the assumption of the theorem,

$$P(\int K(X,z)p(z)dz = 0) = \int I(c(x) = 0)p(x)dx = 0$$

so that $\int K(X,z)p(z)dz > 0$ almost surely and $c(x) > 0$ for dp-almost every $x \in \mathbb{R}^d$. Define

$$\phi(X_1, ..., X_n, x) = I(|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| \geq \delta)$$

We note that by Theorem 1,

$$E\phi(X_1, ..., X_n, x) = P(|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| \geq \delta) \leq 6\exp(-H(B,\sigma^2)c(x)^2\delta^2 n)$$

Then

$$\begin{aligned}
P(|\hat{f}_{GNW}(X) - \frac{\int f(z)k(X,z)p(z)dz}{\int k(X,z)p(z)dz}| \geq \delta) &= E\phi(X_1, ...X_n, X) \\
&= \int_{\mathbb{R}^d} [\int_{(\mathbb{R}^d)^n} \phi(z_1, ..., z_n, x)p(z_1)p(z_2)...p(z_n)dz_1dz_2...dz_n]p(x)dx \\
&= \int_{\mathbb{R}^d} E\phi(X_1, ..., X_n, x)p(x)dx \\
&\leq \int_{\mathbb{R}^d} 6\exp(-H(B,\sigma^2)c(x)^2\delta^2 n)p(x)dx \\
&\leq 6\exp(-H(B,\sigma^2)r^2\delta^2 n) + 6\int_{\mathbb{R}^d} I(c(x) < r)p(x)dx) \\
&= 6\exp(-H(B,\sigma^2)r^2\delta^2 n) + 6P(\int k(X,z)p(z)dz < r)
\end{aligned}$$

□

Remarks

Remark 1 (Generalization of the noise) Lemma 1 and Lemma 2 show that the noise term always concentrates around 0 with exponential rate in n. Moreover one can generalize the results with sub-gaussian noise.

Remark 2 (Generalization of the function class) It is easy to see that as long as $E|f(X_1)k(x,X_1)| = \int |f(z)|k(x,z)p(z)dz < \infty$, the strong law of large numbers states that

$$\hat{f}_{GNW}(x) \to \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}$$

In particular, if $E|f(X_1)| = \int |f(z)|p(z)dz < \infty$ then the last display holds for all values of x for which $c(x) > 0$. However, it is not clear how to obtain concentration results for such a weak assumption. One way to slightly generalize the function class is to consider functions f for which $f(X_1)$ is sub-gaussian i.e. there exists $t > 0$ s.t.

$$E\exp(\frac{f^2(X_1)}{t^2}) = \int \exp(\frac{f^2(z)}{t^2})p(z)dz < \infty$$

With such an assumption on f it is possible to reason as in Lemma 2 to obtain similar concentration result.

Remark 3 (Generalization of the domain of the latent data) Throughout this report we have assumed that the latent data $X_1, ..., X_n$ belongs to $\mathbb{R}^d$. Using the notion of sub-gaussian variables it is possible to allow for the data $X_1, ..., X_n$ to be in essentially any abstract space as long as it is still independent and $||f(X_1)||_{\psi_2} < \infty$. In particular the dimensionality of the data plays no role in the approximation of $\hat{f}_{NW}$ by $\hat{f}_{GNW}$. However, we still have to take into account that our ultimate goal is to estimate f, and not $\hat{f}_{NW}$.

Remark 4 (Comparisson to classical Nadaraya Watson estimator) It is also easy to show with slight alteration of the presented proofs, that with $\hat{f}_{NW}(x) = \frac{\sum_{i=1}^n Y_i k(x,X_i)}{\sum_{i=1}^n k(x,X_i)}$,

$$|\hat{f}_{GNW}(x) - \hat{f}_{NW}(x)| \leq \delta$$

with probability at least $1 - c_1 \exp(-c_2\delta^2 n)$ for some constants $c_1, c_2 > 0$ depending on B, σ^2, k and p and $c(x)$.

Remark 5 Assuming that $\inf_{x\in\mathbb{R}^d} c(x) \geq r > 0$ gives $P(\int k(X,z)p(z)dz < r) = 0$ so that $\hat{f}_{GNW}(X)$ concentrates around $\frac{\int f(z)k(X,z)p(z)dz}{\int k(X,z)p(z)dz}$ with overwhelming probability. In that case, an application of Borel-Cantelli's lemma gives almost sure convergence. This is the case if for example $p(z)$ is compactly supported density (i.e. the data $X_1, ..., X_n$ are drawn i.i.d. from some compact set) and $c(x) > 0$ for all x in the support of p. In general, there is a penalty term $P(\int k(X,z)p(z)dz < r)$ which is highly dependent on the kernel k. However it is still true that $\hat{f}_{GNW}(X)$ converges in probability towards $\frac{\int f(z)k(X,z)p(z)dz}{c(X)}$.

3 L^2 convergence

In this section we study the L^2 convergence of $\hat{f}_{GNW}$ at a fixed point x. We assume that $c(x) > 0$.

Lemma 3 Suppose that X_i are i.i.d Bernoulli variables with parameter $c > 0$. Set

$$Y_n = \begin{cases} \frac{n}{\sum_{i=1}^n X_i} & \text{if } \sum_{i=1}^n X_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Then for all $\frac{c}{2} > \delta > 0$, $p \geq 1$

$$E|Y_n - \frac{1}{c}|^p \leq c^{n-p} + (\frac{2\delta}{c^2})^p + 2^p(n^p + \frac{1}{c^p})\exp(-2\delta^2 n)$$

Proof. Let us denote the event $E_n = \{\sum_{i=1}^n X_i = 0\}$. Then $P(E_n) = c^n$ and

$$E|Y_n - \frac{1}{c}|^p I(E_n) = \frac{1}{c^p}P(E_n) = c^{n-p}$$

Next, denote $A_n(\delta) = \{|\frac{1}{n}\sum_{i=1}^n X_i - c| \geq \delta\}$. On $A_n(\delta) \cap E_n^c$ we have

$$\frac{1}{n}\sum_{i=1}^n X_i \geq \frac{1}{n}$$

Using the fact that $x \to x^p$ is convex for $p \geq 1$, we have

$$\begin{aligned} E|Y_n - \frac{1}{c}|^p I(A_n(\delta) \cap E_n^c) &\leq 2^{p-1}(E([|\frac{n}{\sum_{i=1}^n X_i}|^p + \frac{1}{c^p}]I(A_n(\delta) \cap E_n^c)) \\ &\leq 2^{p-1}(n^p + \frac{1}{c^p})P(A_n(\delta) \cap E_n^c) \\ &\leq 2^{p-1}(n^p + \frac{1}{c^p})P(A_n(\delta)) \\ &\leq 2^p(n^p + \frac{1}{c^p})\exp(-2\delta^2 n) \end{aligned}$$

where once again we used McDiarmid's inequality in the last line.

Finally, on $A_n(\delta)^c$ we have $|\frac{1}{n}\sum_{i=1}^n X_i - c| < \delta$ and in particular $\frac{1}{n}\sum_{i=1}^n X_i \geq c - \delta > \frac{c}{2}$.

Hence,

$$
\begin{aligned}
E(|Y_n - \frac{1}{c}|^p I(A_n(\delta)^c) &= E(|\frac{c - \frac{1}{n}\sum_{i=1}^n X_i}{\frac{1}{n}(\sum_{i=1}^n X_i)c}|^p I(A_n(\delta)^c)) \\
&\leq (\frac{2\delta}{c^2})^p P(A_n(\delta)^c) \\
&\leq (\frac{2\delta}{c^2})^p
\end{aligned}
$$

We note that as soon as $\delta < c$, $E_n \subseteq A_n(\delta)$ and hence the result follows by spliting the expectation in three parts as above. $\square$

The event $E_n = \{\sum_{i=1}^n a(x, X_i) = 0\}$ has probability $(1 - c(x))^n$. In this section, for ease of notation we denote by $E_*(\cdot)$ the expection over the event E_n^c and with $E(\cdot)$ the standard expectation. We emphasize the trivial inequality $E_*(Z) \leq E(Z)$ whenever Z is a nonnegative random variable. We also denote the event $A_n(\delta) = \{|\frac{1}{n}\sum_{i=1}^n a(x, X_i) - c(x)| \geq \delta\}$.

Corollary 2 For any $0 < r < 1$,

$$
E_*|\frac{1}{\frac{1}{n}\sum_{i=1}^n a(x, X_i)} - \frac{1}{c(x)}|^2 \leq \frac{1}{n^r}(1 + o(1))
$$

Proof. Setting $\delta = \frac{1}{n^{\frac{r}{2}}}c(x)$ in Lemma 3 yields the claimed result. $\square$

Lemma 4 For all $\frac{c(x)}{2} > \delta > 0$, we have

$$
E_*(\frac{\sum_{i=1}^n \epsilon_i a(x, X_i)}{\sum_{i=1}^n a(x, X_i)})^2 \leq \frac{\sigma^2}{n}(\frac{1}{c(x)} + \frac{2\delta}{c(x)^2} + 2(n + \frac{1}{c(x)})\exp(-2\delta^2 n))
$$

Proof. Set $w_i = \frac{a(x,X_i)}{\sum_{i=1}^n a(x,X_i)}$. Then $w_1, ..., w_n$ are indpendent from $\epsilon_1, ...\epsilon_n$ and as the ϵ_i's are centered,

$$
E_*((\sum_{i=1}^n \epsilon_i w_i)^2) = \sum_{i=1}^n E_*(\epsilon_i^2 w_i^2) = \sigma^2 E_*(\sum_{i=1}^n w_i^2)
$$

But $w_i^2 = \frac{a(x,X_i)^2}{(\sum_{i=1}^n a(x,X_i))^2} = \frac{a(x,X_i)}{(\sum_{i=1}^n a(x,X_i))^2}$ and hence

$$
\sum_{i=1}^n w_i^2 = \frac{1}{\sum_{i=1}^n a(x, X_i)}
$$

We get

$$
E_*(\sum_{i=1}^n \epsilon_i w_i)^2 = \frac{\sigma^2}{n} E_*(\frac{n}{\sum_{i=1}^n a(x, X_i)})
$$

The conclusion follows from Lemma 3 with $p = 1$.

$\square$

Lemma 5 Suppose that $f(X_1) \in L^{2+\rho}$ for some $\rho > 0$. Then for $\delta < \frac{c(x)}{2}$ we have

$$
E_*(\frac{\frac{1}{n}\sum_{i=1}^n f(X_i)a(x, X_i) - \int f(z)k(x, z)p(z)dz}{\frac{1}{n}\sum_{i=1}^n a(x, X_i)})^2 \leq \frac{4}{nc(x)^2}||f(X_1)||_{L^2}^2 + 2^{\frac{1}{1+\frac{2}{\rho}}+\frac{1}{2}} n^2(||f(X_1)||_{L^{2+\rho}})^{\frac{1}{2}} \exp(-\frac{2\delta^2 n}{1+\frac{2}{\rho}})
$$

Proof. Consider $A_n(\delta) = \{|\frac{1}{n}\sum_{i=1}^n a(x, X_i) - c(x)| \geq \delta\}$. On $A_n(\delta)^c$, we have $\frac{1}{n}\sum_{i=1} a(x, X_i) \geq \frac{1}{2}c(x)$ as soon as $\delta < \frac{1}{2}c(x)$. For ease of notation, set

$$
W_i = f(X_i)a(x, X_i) - \int f(z)k(x, z)p(z)dz
$$

Then W_i are i.i.d, centered and

$$\begin{aligned}
E_*(\frac{\frac{1}{n}\sum_{i=1}^n W_i}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} I(A_n(\delta)^c))^2 &\le \frac{4}{c(x)^2} E(\frac{1}{n}\sum_{i=1}^n W_i)^2 \\
&= \frac{4}{nc(x)^2} Var(W_1) \\
&= \frac{4}{nc(x)^2} EW_1^2 \\
&= \frac{4}{nc(x)^2}[\int f(z)^2 k(x,z)p(z)dz - (\int f(z)k(x,z)p(z)dz)^2]
\end{aligned}$$

Next on $A_n(\delta)$ under $E_*(\cdot)$ we have $\frac{1}{n}\sum_{i=1}^n a(x,X_i) \ge \frac{1}{n}$ and

$$\begin{aligned}
E_*([\frac{\frac{1}{n}\sum_{i=1}^n W_i}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}]^2 I(A_n(\delta))) &\le E((\sum_{i=1}^n W_i)^2 I(A_n(\delta))) \\
&\le n\sum_{i=1}^n EW_i^2 I(A_n(\delta)) \\
&\le n\sum_{i=1}^n [EW_i^{2+\rho}]^{\frac{1}{1+\frac{\rho}{2}}} [P(A_n(\delta))]^{\frac{1}{1+\frac{2}{\rho}}} \\
&\le 2^{\frac{1}{1+\frac{2}{\rho}}} n^2 (E|W_1|^{2+\rho})^{\frac{1}{1+\frac{\rho}{2}}} \exp(-\frac{2\delta^2 n}{1+\frac{2}{\rho}})
\end{aligned}$$

Here, we used the basic Cauchy-Schwarz inequality in line 2 and Holder's inequality with $p = 1+\frac{\rho}{2}$ and $q = 1+\frac{2}{\rho}$ in line 3. Finally, by conditional Jensen's inequality, we have

$$\begin{aligned}
|W_1|^{2+\rho} &= |f(X_1)a(x,X_1) - Ef(X_2)a(x,X_2)|^{2+\rho} \\
&= |E(f(X_1)a(x,X_1) - f(X_2)a(x,X_2)|X_1)|^{2+\rho} \\
&\le E(|f(X_1)a(x,X_1) - f(X_2)a(x,X_2)|^{2+\rho}|X_1)
\end{aligned}$$

and hence

$$||W_1||_{L^{2+\rho}} \le ||f(X_1)a(x,X_1) - f(X_2)a(x,X_2)||_{L^{2+\rho}} \le 2||f(X_1)||_{L^{2+\rho}}$$

We conclude by breaking the expectation on $A_n(\delta)$ and $A_n(\delta)^c$. □

Theorem 2 (L^2 convergence of $\hat{f}_{GNW}$) Suppose that $f(X_1) \in L^{2+\rho}$ for some $\rho > 0$. Then for any $0 < r < 1$ we have

$$E_*(\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz})^2 \le \frac{1}{n^r}(1+o(1))$$

Proof. Recalling (1), we have:

$$\begin{aligned}
E_*|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}|^2 \le & 3E_*|\frac{\frac{1}{n}\sum_{i=1}^n f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}|^2 \\
&+ 3E_*|\frac{\sum_{i=1}^n \epsilon_i a(x,X_i)}{\sum_{i=1}^n a(x,X_i)}|^2 \\
&+ 3|\int f(z)k(x,z)p(z)dz|^2 E_*|\frac{1}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} - \frac{1}{c(x)}|^2
\end{aligned}$$

The three sumands on the right hand side of the last display go to zero by Corollary 2, Lemma 4 and Lemma 5 at the stated rate. □

Remarks

Remark 6 (L^p convergence for $p > 1$ in the noiseless case) Under the classical assumption that $c(x) > 0$ and in addition $f \in L^{p+\rho}$ and $\sigma^2 = 0$, it is possible to show that

$$E|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}|^p \to 0$$

as $n \to \infty$. Indeed, in the noiseless case one only needs to show that $||\frac{\frac{1}{n}\sum_{i=1}^n f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}||_{L^p}$ and $||\frac{1}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} - \frac{1}{c(x)}||_{L^p}$ go to zero. The second term does indeed go to zero by Lemma 3. The first term can be broken over two events $A_n(\delta)$ of low probability and $A_n(\delta)^c$ of high probability. On the low probability event $A_n(\delta)$ the assumption $f \in L^{p+\rho}$ allows us to replicate the L^2 argument. On the high probability event $A_n(\delta)$, one can use the fact that $f(X_i)$ are $L^{p+\rho}$ bounded to conclude that $|f(X_i)|^p$ are $L^{1+\frac{\rho}{p}}$ bounded and hence uniformly integrable. Further it can be shown that $|\frac{\sum_{i=1}^n [f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz]}{n}|^p$ is uniformly integrable and hence $E|\frac{\sum_{i=1}^n [f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz]}{n}|^p \to 0$ as $n \to \infty$.

Remark 7 (Regularization) We can easily fix the L^2 convergence issue by considering the **Regularized Graphical Nadaraya Watson** estimator:

$$\hat{f}_{RGNW,\alpha,\beta}(x) = \frac{\sum_{i=1}^n Y_i a(x,X_i)}{\sum_{i=1}^n a(x,X_i) + \alpha n I(\frac{1}{n}\sum_{i=1}^n a(x,X_i) \le \beta c(x))}$$

with $\alpha \ge 0$ and $0 < \beta < 1$. The idea behind this regularization is to penalize extreme events when we observe too few edges. We note that for $\alpha = 0$ we recover $\hat{f}_{GNW}(x)$. Moreover, taking $\delta = (1-\beta)c(x)$, and using McDiarmid's inequality we get that

$$\hat{f}_{RGNW,\alpha,\beta}(x) = \hat{f}_{GNW}(x)$$

with probability at least $1 - \exp(-2(1-\beta)^2 c(x)^2 n)$, so that the concentration properties from the previous section as well as the analysis for the L^2 convergence on the set $A_n(\delta)^c$ still hold for $\hat{f}_{RGNW,\alpha,\beta}$. We note that on $A_n(\delta)$ we have

$$\sum_{i=1}^n a(x,X_i) + n\alpha c(x) I(\frac{1}{n}\sum_{i=1}^n a(x,X_i) \le \beta c(x)) \ge \min(\alpha,\beta) n c(x)$$

so that

$$E_{A_n(\delta)}(\frac{\sum_{i=1}^n f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz}{\sum_{i=1}^n a(x,X_i) + \alpha n I(\frac{1}{n}\sum_{i=1}^n a(x,X_i) \le \beta c(x))})^2 \le G(x) E_{A_n(\delta)}(\frac{1}{n}\sum_{i=1}[f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz])^2$$

where $G(x) = \frac{1}{\min(\alpha,\beta)^2 c(x)^2}$ and $E_{A_n(\delta)}$ is the expectation over the event $A_n(\delta)$. In this case the assumption $f \in L^2$ is sufficient to ensure convergence. However, if we asssume that $f \in L^{2+\rho}$ for some $\rho > 0$, then an application of Holder's inequality yields much stronger convergence rate compared to the standard Graphical Nadaraya Watson estimator. The parameters α and β in practice can be chosen with cross validation.

4 Simulations

We test empirically the performance of $\hat{f}_{GNW}$. We assume that the latent data $X_1, ..., X_n$ is i.i.d. uniform on $[0,1]$ and we compare $\hat{f}_{GNW}(x) = \frac{\sum_{i=1}^n Y_i a(x,X_i)}{\sum_{i=1}^n a(x,X_i)}$, $\hat{f}_{NW}(x) = \frac{\sum_{i=1}^n Y_i k(x,X_i)}{\sum_{i=1}^n k(x,X_i)}$ and $f(x)$. We choose a sample size of $n = 50000$. The variance is set to $\sigma^2 = 0.01$, and the bandwith is set to $h = 0.11$. We consider the following five kernels:

Rectangular: $k(x,y) = \frac{1}{2} I(|x-y| < h)$

Triangular: $k(x,y) = (1 - \frac{|x-y|}{h}) I(|x-y| \le h)$

Parabolic (Epanechnikov): $k(x,y) = \frac{3}{4}(1 - (\frac{x-y}{h})^2) I(|x-y| \le h)$

Gaussian: $k(x,y) = \exp(-\frac{(x-y)^2}{h})$

Laplacian: $k(x,y) = \exp(-\frac{|x-y|}{h})$

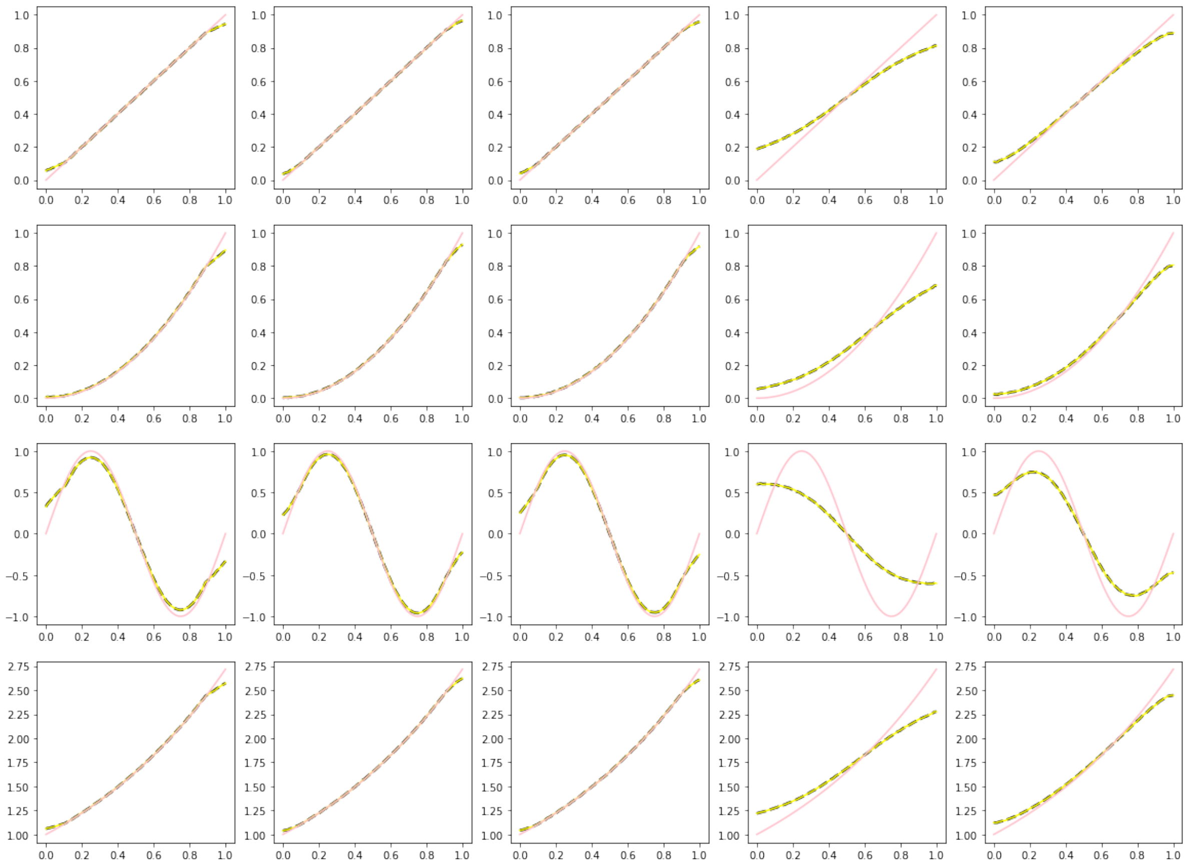

Figure 1: Each column represents a kernel, in the order listed above (rectangular, triangular, Epanechnikov, Gaussian, Laplacian). Each row represents a function in the following order $x, x^2, \sin(2\pi x), \exp(x)$. The pink line represents the true function, the yellow solid line is the plot of $\hat{f}_{GNW}$ and the black dashed line represents $\hat{f}_{NW}$.

Simulation 1 For 100 equally spaced points on $[0, 1]$, we compute $\hat{f}_{GNW}(x), \hat{f}_{NW}$ and $f(x)$ and plot their graphs.

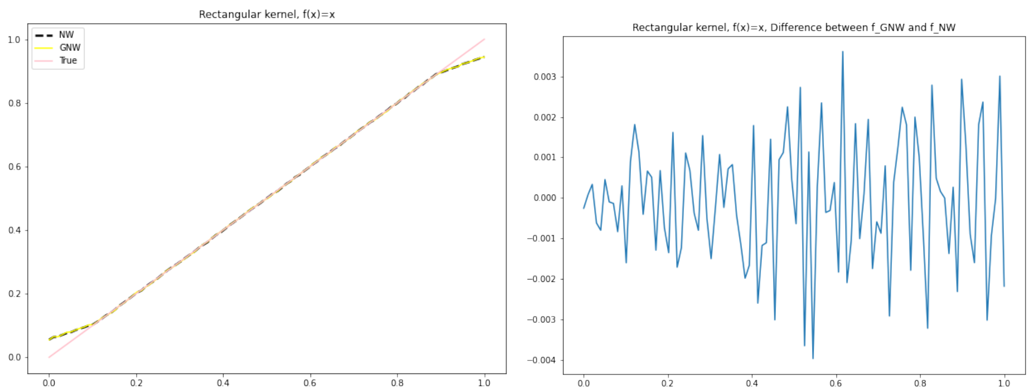

Figure 2: Left: comparison of $\hat{f}_{GNW}$, $\hat{f}_{NW}$ and f (solid yellow line, dashed black line and solid pink line, respectively. Right: Plot of $\hat{f}_{GNW} - \hat{f}_{NW}$.

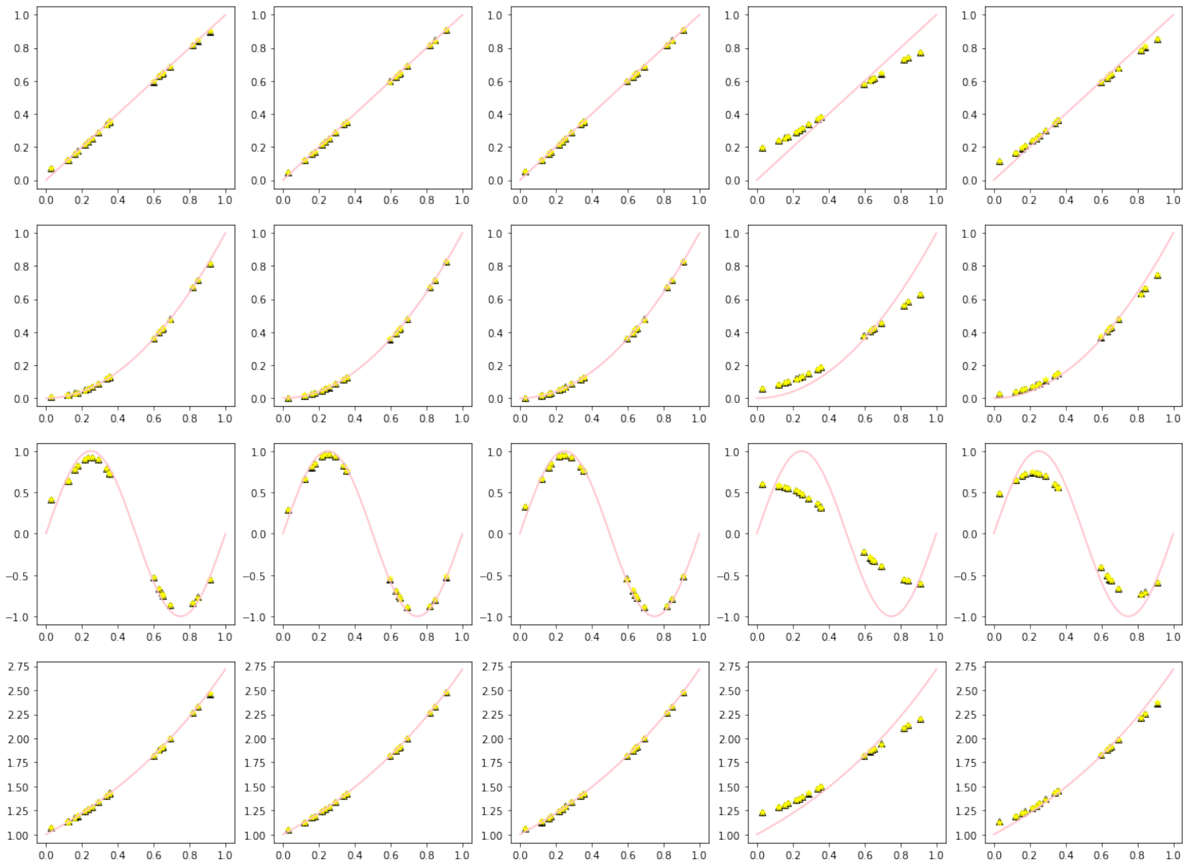

Figure 3: Each column represents a kernel in the order listed above. Each row represents a function as in Figure 1. We represent $\hat{f}_{GNW}$ with yellow triangle, $\hat{f}_{NW}$ with black star symbol and the true function with solid pink line.

Simulation 2 For 20 points chosen independently with uniform distribution on $[0, 1]$, we compute $\hat{f}_{GNW}, \hat{f}_{NW}$ and plot them agains the graph of $f(x)$.

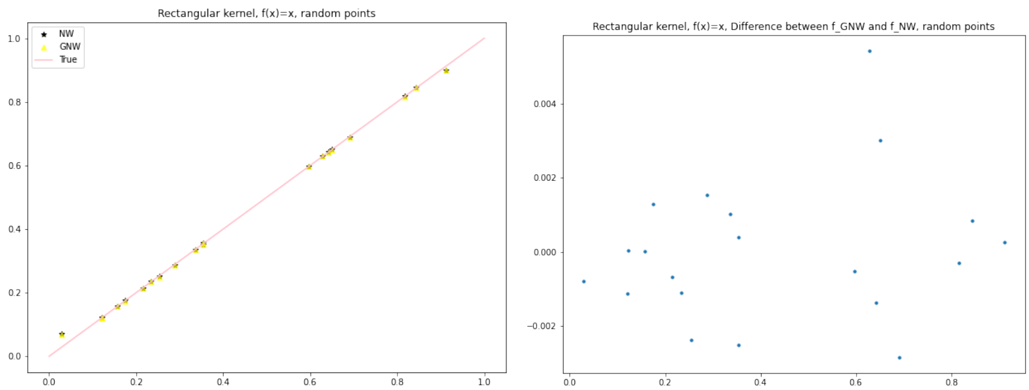

Figure 4: Left: comparison of scatter plots of $\hat{f}_{GNW}$, $\hat{f}_{NW}$ and the plot of f, represented with yellow triangles, black stars and solid pink line. Right: scatter plot of $\hat{f}_{GNW} - \hat{f}_{NW}$.

References

[RBD10] Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. “On Learning with Integral Operators”. In: *Journal of Machine Learning Research* 11 (Feb. 2010), pp. 905–934. DOI: 10.1145/1756006.1756036.

[Tsy08] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated, 2008. ISBN: 0387790519.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. DOI: 10.1017/9781108231596.