Graphical Nadaraya Watson estimator

Martin Gjorgjevski

May 2022

Contents

1 Motivation and introduction

In the classical nonparametric regression setting we are given data $X_1, ..., X_n$ i.i.d. with density p. We are also provided with noisy observations $Y_i = f(X_i) + \epsilon_i$ with f 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]

$$\hat{f}_{NW}(x) = \frac{\sum_{i=1}^n Y_i k(\frac{x-X_i}{h})}{\sum_{i=1}^n k(\frac{x-X_i}{h})}$$

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

In our setting we assume that the data $X_1, ..., X_n$ is latent, and that in addition to the noisy observations $Y_1, ..., Y_n$ we observe a random graph associated with the data $X_1, ..., X_n$ 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, $k : \mathbb{R}^2 \to [0, 1]$ 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. We are interested in estimating f in this setting. Inspired by the classical Nadaraya Watson estimator, we introduce the **Graphical Nadaraya Watson** estimator:

$$\hat{f}_{GNW}(x) = \frac{\sum_{i=1}^n Y_i a(x, X_i)}{\sum_{i=1}^n a(x, X_i)}$$

In this report we are investigating the convergence of this estimator. Our main result can be stated as follows:

Theorem If $||f||_\infty \leq B$, $Ek(x, X_1) = \int k(x, z)p(z)dz > 0$ and $\delta \leq 4B$ then

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

with probability at least $1 - 8\exp(-H\delta^2 n)$ where $H > 0$ is a constant that depends on B, σ^2, k and p but not on n and δ.

The assumption $Ek(x, X_1) > 0$ is natural, as $Ek(x, X_1) = 0$ implies that almost surely $k(x, X_i) = 0$ and hence we don't observe any edges between x and the latent data. The boundedness assumption can be somewhat loosened, see the remark section. The precise result is Theroem 1.

2 Main results

Lemma 1 Suppose that f is bounded, measurable function with $||f||_\infty \leq B$. Then

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

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

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

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:

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

Changing one of the $u_i's$ gives:

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

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 centered and independent, $|w_i| \leq 1$ and ϵ_i are Gaussian variables with variance σ^2. Then

$$P(|\frac{1}{n}\sum_{i=1}^{n} w_i\epsilon_i| \geq t) \leq 2\exp(-Ct^2n)$$

where C depends on σ^2 but not on n (In particular one can take $C = \frac{9\sqrt{e}}{4\sigma^2}$).

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

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

We have

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

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

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

In particular this shows that

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

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

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

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

□

Theorem 1 Suppose that $||f||_\infty \leq B$ and $Ek(x,X_1) = \int k(x,z)p(z)dz > 0$. Then for $0 < \delta < 4B$ and $H(B,\sigma^2,k,p) = \min\{\frac{(\int k(x,z)p(z)dz)^2}{160B^2}, \frac{C(\int k(x,z)p(z)dz)^2}{64B^2\sigma^2}, \frac{1}{128\sigma^2}\}$ we have

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

except on a set of probability no larger than $8\exp(-H(B,\sigma^2,k,p)\delta^2 n)$

Proof. We have

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

We focus on the third term in the right hand side of the last display:

$$\begin{aligned}
\frac{\int f(z)k(x,z)p(z)dz}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz} &= \int f(z)k(x,z)p(z)dz[\frac{1}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} - \frac{1}{\int k(x,z)p(z)dz}] \\
&= \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz} \frac{\frac{1}{n}\sum_{i=1}^n [a(x,X_i) - \int k(x,z)p(z)dz]}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}
\end{aligned}$$

Let $\delta > 0$ and denote

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

Choosing $\delta_2 \leq \frac{1}{2}\int k(x,z)p(z)dz$, on $(A_{\delta_1} \cup B_{\delta_2} \cup C_{\delta_3} \cup D_{\delta_4})^c$ we have:

$$\begin{aligned}
|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| &\leq |\frac{\frac{1}{n}\sum_{i=1}^n [f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz]}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}| \\
&+ |\frac{\frac{1}{n}\sum_{i=1}\epsilon_i[a(x,X_i) - \int k(x,z)p(z)dz]}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}| \\
&+ |\frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz} \frac{\frac{1}{n}\sum_{i=1}^n [a(x,X_i) - \int k(x,z)p(z)dz]}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}| \\
&+ |\int k(x,z)p(z)dz \frac{\frac{1}{n}\sum_{i=1}^n \epsilon_i}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}| \\
&\leq \frac{\delta_1 + \delta_3 + \delta_2 B + \delta_4 \int k(x,z)p(z)dz}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} \\
&\leq \frac{2(\delta_1 + \delta_2 B + \delta_3)}{\int k(x,z)p(z)dz} + 2\delta_4
\end{aligned}$$

Finally, setting

$$\delta_1 = \delta_3 = \frac{\delta \int k(x,z)p(z)dz}{8}, \delta_2 = \frac{\delta \int k(x,z)p(z)dz}{8B}, \delta_4 = \frac{\delta}{8}$$

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 $(A_{\delta_1} \cup B_{\delta_2} \cup C_{\delta_3} \cup D_{\delta_4})^c$.
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.
Finally, it is easy to show (for example by using Chernoff's bound) that $P(D_{\delta_4}) \leq 2\exp(-\frac{\delta_4^2 n}{2\sigma^2})$
Now

$$\begin{aligned}P(A_{\delta 1} \cup B_{\delta_2} \cup C_{\delta_3} \cup D_{\delta_4}) &\leq P(A_{\delta_1}) + P(B_{\delta_2}) + P(C_{\delta_3}) + P(D_{\delta_4}) \\ &\leq 8\exp(-H(B,\sigma^2,k,p)\delta^2 n)\end{aligned}$$

which completes the proof. □

Corollary 1 Under the assumptions and with the notation of Theorem 1, suppose that X is independent of the latent data $X_1, ..., X_n$ with density q such that

$$Ek(X, X_1) = \int\int k(x,z)p(z)q(x)dzdx > 0$$

Then

$$P(|\hat{f}_{GNW}(X) - \frac{\int f(z)k(X,z)p(z)dz}{\int k(X,z)p(z)dz}| \geq \delta) \leq 8\exp(-H(B,\sigma^2,k,p)\delta^2 n)$$

In particular, when X is random and independent from the latent data, $\hat{f}_{GNW}(X) \to \frac{\int f(z)k(X,z)p(z)dz}{\int k(X,z)p(z)dz}$ almost surely[1]

Proof. Write $\phi(X_1, ..., X_n, x) = I(|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| \geq \delta)$. We note that by Theroem 1, $E\phi(X_1, ..., X_n, x) = P(|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| \geq \delta) \leq 8\exp(-H(B,\sigma^2,k,p)\delta^2 n)$ Then

$$\begin{aligned}P(|\hat{f}_{GNW}(X) - \frac{\int f(z)k(X,z)p(z)dz}{\int k(X,z)p(z)dz}| \geq \delta) &= E\phi(X_1, ...X_n, X) \\ &= \int_{\mathbb{R}} [\int_{\mathbb{R}^n} \phi(z_1, ..., z_n, x)p(z_1)p(z_2)...p(z_n)dz_1dz_2...dz_n]q(x)dx \\ &= \int_{\mathbb{R}} E\phi(X_1, ..., X_n, x)q(x)dx \\ &\leq 8\exp(-H(B,\sigma^2,k,p)\delta^2 n)\int_{\mathbb{R}} q(z)dz \\ &= 8\exp(-H(B,\sigma^2,k,p)\delta^2 n)\end{aligned}$$

□

In particular, if X is independent from $X_1, ...X_n$ and with the same distribution, then under the mild assumption that $Ek(X_1, X_2) = \int\int k(x,z)p^2(z)dz > 0$, we get the result from corollary 1.

[1]In contrast to the deterministic case, this is still a random variable dependent on X

3 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, the arguments used require only sub-gaussian noise, so one can generalize the result 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

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

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 $Ek(x,X_1) > 0$. 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 $\mathbb{R}$. Using the notion of sub-gaussian variables it is possible to avoid the framework of McDiarmid's bounded differences inequality and thus we can 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 $\hat{f}_{NW}$. Hence we will see the impact of the dimensionality of data when we approximate f by $\hat{f}_{NW}$

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.

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

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

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

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

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.

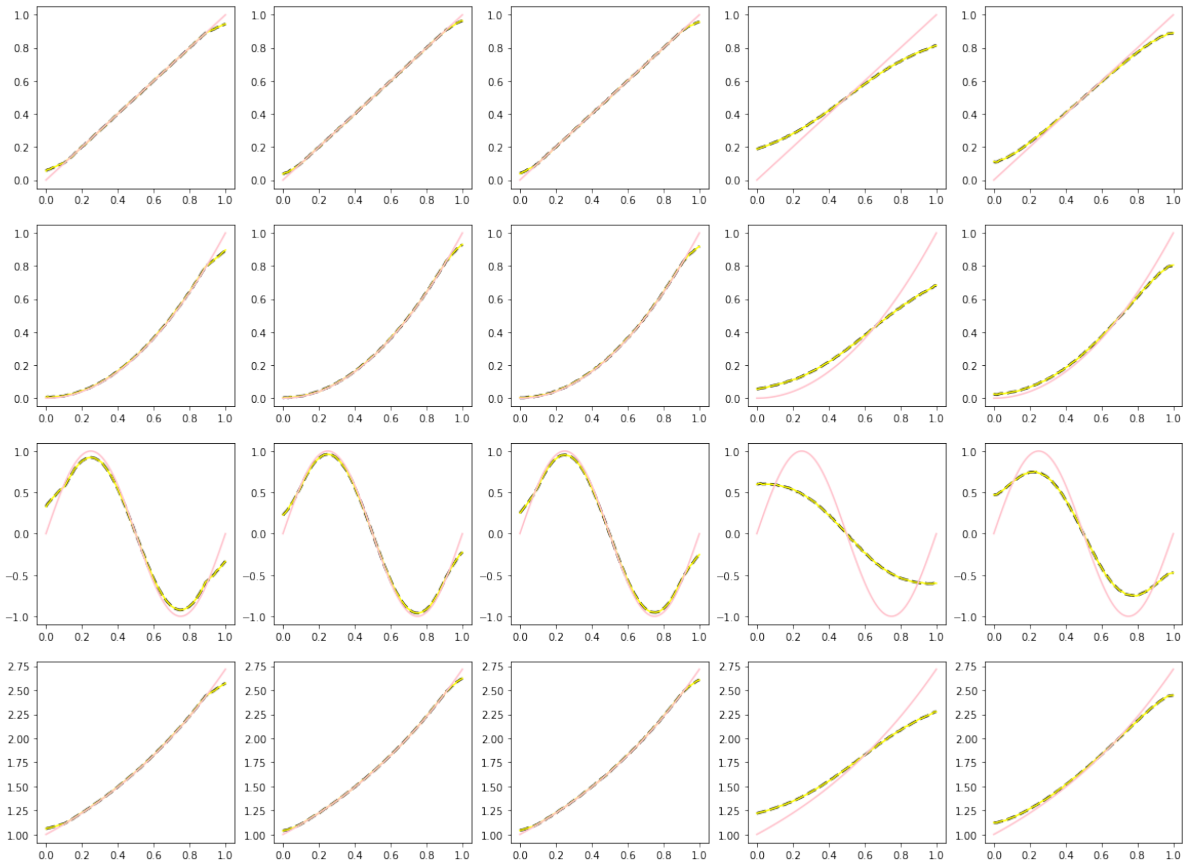

Figure 1: Each column represents a kernel, in the order listed above (rectangular, triangular, Epanechnikov, Gaussian, Laplacian). Each row represents a function in the following order $x, x^2, \sin(2\pi x), \exp(x)$. The pink line represents the true function, the yellow solid line is the plot of $\hat{f}_{GNW}$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

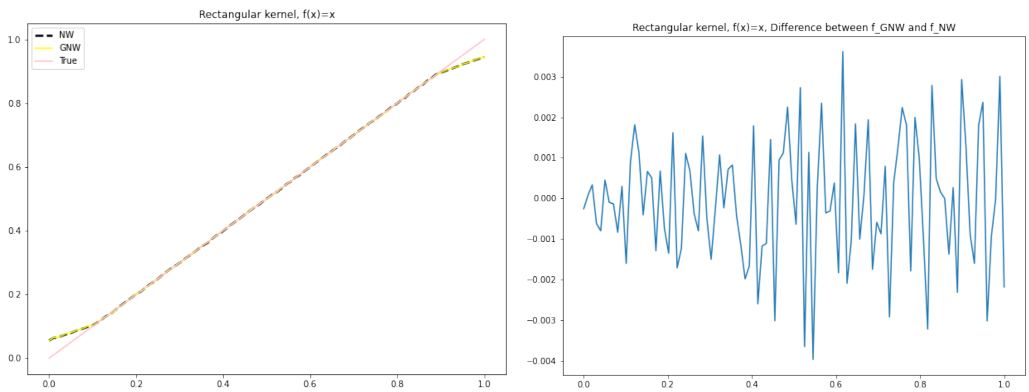

Figure 2: Left: comparison of $\hat{f}_{GNW}$, $\hat{f}_{NW}$ and f (solid yellow line, dashed black line and solid pink line, respectively. Right: Plot of $\hat{f}_{GNW} - \hat{f}_{NW}$.

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 $f(x)$.

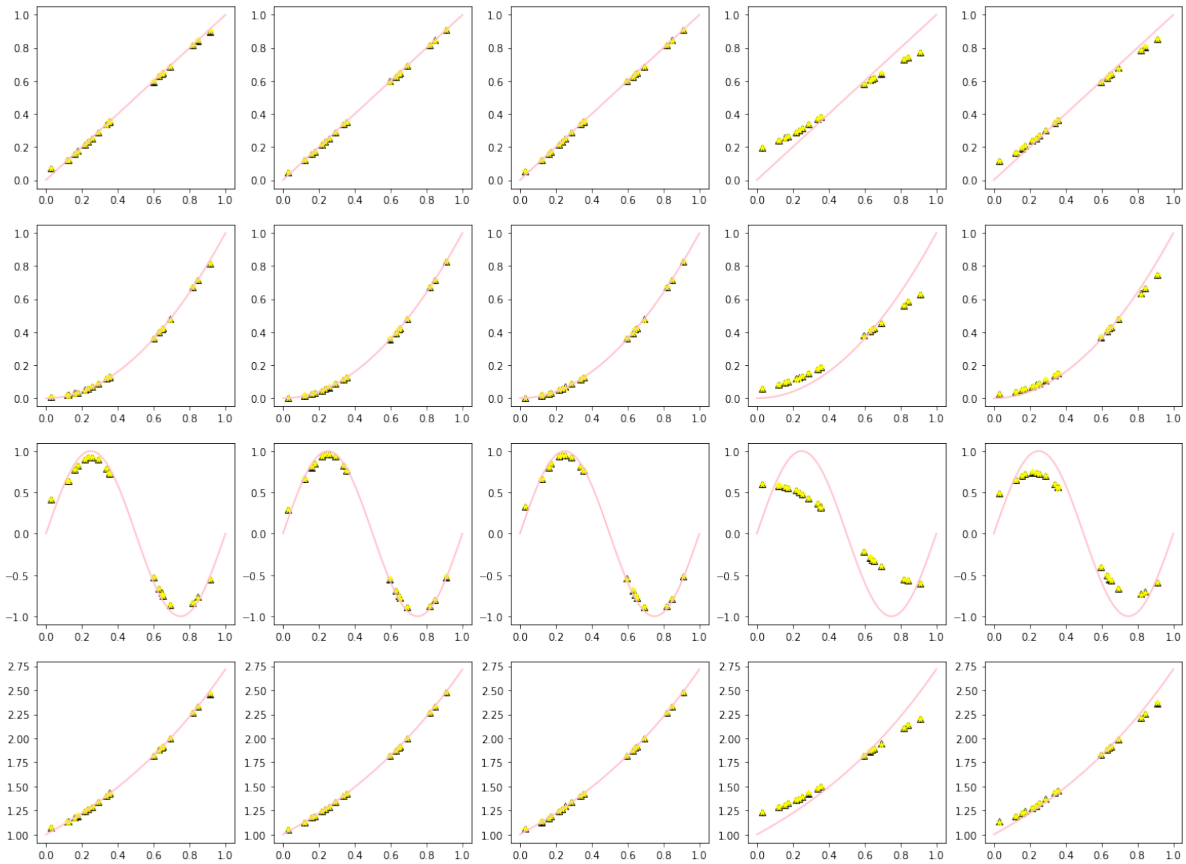

Figure 3: Each column represents a kernel in the order listed above. Each row represents a function as in Figure 1. We represent $\hat{f}_{GNW}$ with yellow triangle, $\hat{f}_{NW}$ with black star symbol and the true function with solid pink line.

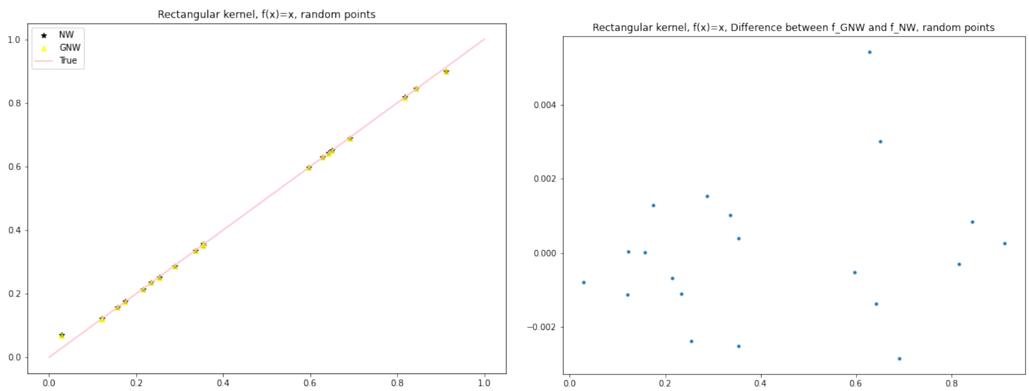

Figure 4: Left: comparison of scatter plots of $\hat{f}_{GNW}$, $\hat{f}_{NW}$ and the plot of f, represented with yellow triangles, black stars and solid pink line. Right: scatter plot of $\hat{f}_{GNW} - \hat{f}_{NW}$.

References

[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.