Graphical Nadaraya Watson estimator

Martin Gjorgjevski

May 2022

Contents

1 Introduction, motivation and notations

1.1 Brief overview of nonparametric regression

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 given 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 $\mathcal{F}$ and $\epsilon_1, ..., \epsilon_n$ are assumed to be i.i.d. centered Gaussian with variance σ^2. The goal is to estimate f. The term *nonparametric* stems from the fact that the function class $\mathcal{F}$ can not be parametrized by a subset of $\mathbb{R}^m$ for any $m \in \mathbb{N}$. Typically one makes an assumption about the smoothness of f such as Holder continuity (Holder class $\Sigma(\beta, L)$) or boundedness of its derivatives (Sobolev class $W(\beta, L)$). A linear nonparametric regression estimator for f is an estimator $\hat{f}$ which can be expressed as $\hat{f}(x) = \sum_{i=1}^n Y_i W_{n,i}(x)$ where $W_{n,i}(x)$ depends on $x, X_1, ..., X_n$ but not on the observations $Y_1, ..., Y_n$. We give a brief overview of two popular types of estimators used in nonparametric regression.

Projeciton estimators We assume that the data $X_1, ..., X_n$ is uniformly distributed on $[0, 1]$ and that $f \in L^2([0, 1], dx)$ has a pointwise convergent Fourier expansion with respect to some orthonormal basis $\{\phi_j, j \geq 1\}$ of $L^2([0, 1], dx)$, that is

$$f(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x)$$

holds for all $x \in [0, 1]$, where $\theta_j = \int f(z)\phi_j(z)dz$. The idea is to approximate f by a projection on the linear space spanned by the first N elements in the basis, i.e. to approximate

$$f_N(x) = \sum_{j=1}^{N} \theta_j \phi_j(x)$$

By the law of large numbers it is easy to see that $\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^n Y_i \phi_j(X_i) \to \theta_j$ as $n \to \infty$. In this context, the estimator $\hat{f}_N(x) = \sum_{j=1}^N \hat{\theta}_j \phi_j(x)$ is called a projection estimator [1]. Under suitable

[1] It is easy to see that the projection estimator is a linear estimator

smoothness assumptions (see [Tsy08] p. 55) it is possible to show that if N is of the order $n^{\frac{1}{2\beta+1}}$ then the mean integrated square error $E(\int_0^1(\hat{f}_N(x)-f(x))^2dx)$ goes to 0 at a rate $n^{-\frac{2\beta}{2\beta+1}}$. Here β is an integer related to the level of smoothness of f.

A natural generalization of the projection estimator is the least squares estimator. Given orthonormal basis given $\{\phi_j, j\geq 1\}$, we consider $\phi^{(N)}(x)=(\phi_1(x),...,\phi_N(x))^T$. Then the least squares estimators is given by

$$\hat{f}_{LS}(x)=\phi^{(N)}(x)^T\hat{\theta}_{LS}$$

where

$$\hat{\theta}^{LS}=\arg\min_{\theta\in\mathbb{R}^N}\sum_{i=1}^n(Y_i-\theta^T\phi^{(N)}(X_i))^2$$

There are other popular estimators such as reweighted projection estimators where we consider estimators of the form $\hat{f}_\lambda(x)=\sum_{j\geq 1}\lambda_j\hat{\theta}_j\phi_j(x)$ with $\lambda=(\lambda_j)_{j\geq 1}\in l^2(\mathbb{N})$ or penalized least squares estimators which are given by

$$\hat{\theta}_p=\arg\min_{\theta\in\mathbb{R}^N}(\sum_{i=1}^n(Y_i-\theta^T\phi^{(N)}(X_i))^2+\sum_{i=1}^n b_j|\theta_j|^p)$$

with $p=1$ resulting in the LASSO estimator and $p=2$ resulting in the Tikhonov regularization estimator (also known as Ridge regression estimator).

Kernels and local polynomial estimators A kernel k on $\mathbb{R}^d$ is a symmetric function $k:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$. k is said to be positive semidefinite kernel if for any $x_1,...,x_n\in\mathbb{R}^d$, and any $c_1,...,c_n\in\mathbb{R}$ we have

$$\sum_{i,j=1}^n k(x_i,x_j)c_ic_j\geq 0$$

It is said to be stationary if $k(x,y)=k(x-y)$ and it is said to be radial basis kernel if $k(x,y)=k(||x-y||)$. A common way to construct kernels on $\mathbb{R}^d$ is to take tensor product of one dimensional kernels, that is if $k_1,...,k_d$ are kernels on $\mathbb{R}$ then

$$k(x,y)=\prod_{j=1}^d k_j(x_j,y_j)$$

is a kernel on $\mathbb{R}^d$, where x_j and y_j are the j-th component of x and y respectively. If $k_1,...,k_d$ are positive semidefinite, then so is k.

Another popular approach to nonparametric regression is the local polynomial estimator which we now define. We assume that we are given a kernel $K:\mathbb{R}\to\mathbb{R}$, a parameter $h>0$ known as a bandwith and an integer $l\geq 0$.

$$\hat{\theta}(x)=\arg\min_{\theta\in\mathbb{R}^{l+1}}\sum_{i=1}^n(Y_i-\theta^TU(\frac{X_i-x}{h}))^2k(\frac{X_i-x}{h})$$

A popular approach for this task is the Nadaraya Watson estimator [Tsy08]

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

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

1.2 Latent Position Models

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:

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

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

1.3 Notations

Throughout this report all random variables are considered on a joint probability space $(\Omega, \mathcal{F}, P)$. The latent variables $X_1, ..., X_n$ are assumed to be independent with distribution which is absolutely continouos with respect to Lebesgue measure on $\mathbb{R}^d$ with density p. Given a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$, the associated integral operator $T_k : L^1(\mathbb{R}^d, \mathcal{B}_d, pdx) \to L^\infty(\mathbb{R}^d, \mathcal{B}_d, pdx)$ is given by

$$T_k(f)(x) = \int f(z)k(x,z)p(z)dz$$

Here $\mathcal{B}_d$ is the Borel σ-algebra on $\mathbb{R}^d$ and pdx stands for the probability measure μ on $\mathbb{R}^d$ which is given by $\mu(B) = \int_B p(x)dx$ (that is, the probability measure associated with the latent data X_1). Note that T_k depends on the distribution p. Moreover, it is easy to see $||T_k(f)||_\infty \leq ||f||_{L^1}$. As pdx is a probability measure, compositions of T_k of any order $m \geq 1$ are well defined, and

$$T_k^m(f)(x) = \int_{\mathbb{R}^d} T_k^{m-1}(f(z))k(x,z)p(z)dz = \int_{(\mathbb{R}^d)^{\otimes n}} f(z_1)(\prod_{i=1}^{m-1} k(z_i, z_{i+1}))k(z_m, x) \prod_{i=1}^{m} p(z_i)dz_i$$

We introduce the connection parameter of order m

$$c_m(\cdot) = T_k^m(1)(\cdot)$$

In the case $m = 1$, we use the notation $c(x)$ in place of $c_1(x)$. In particular,

$$c(x) = \int_{\mathbb{R}^d} k(x,z)p(z)dz = Ek(x, X_1)$$

This parameter plays a crucial role in our analysis. 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 ν almost surely, we need to assume $\int I(c(x) = 0)d\nu(x) = 0$[2].

2 Concentration properties

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

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

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. uniform variables on $[0,1]$ independent from the $X_i's$ and $\epsilon_i's$. Define

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

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 (**vershynin˙2018** Thm 2.9.1). Changing one of the $x_i's$ gives:

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

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

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.

□

Corollary 1 Suppose that $f(X_1)$ is (essentially) bounded, measurable function with $||f(X_1)||_\infty \leq B$ and that X is independent from and with the same distribution as $X_1,...,X_n$. Then

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

Proof. Nearly the same argument as Corollary 2. Thus ommited at the moment. □

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

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

where $c > 0$ is an absolute constant.

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

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

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 (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

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

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. □

Corollary 2 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

$$\phi(x, X_1, ..., X_n, U_1, ..., U_n) = I(|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| \geq \delta)$$

We note that by Theorem 1,

$$E\phi(x, X_1, ..., X_n, U_1, ..., U_n) = P(|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| \geq \delta) \leq 6\exp(-H(B,\sigma^2)c(x)^2\delta^2 n)$$

Then

$$\begin{aligned}
P(|\hat{f}_{GNW}(X) - \frac{\int f(z)k(X,z)p(z)dz}{\int k(X,z)p(z)dz}| \geq \delta) &= E\phi(X, X_1, ...X_n, U_1, U_2, ..., U_n) \\
&= E(E\phi(X, X_1, ..., X_n, U_1, .., U_n|X)) \\
&= \int_{\mathbb{R}^d} P(|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| \geq \delta)p(x)dx \\
&\leq \int_{\mathbb{R}^d} 6\exp(-H(B,\sigma^2)c(x)^2\delta^2 n)p(x)dx \\
&\leq 6\exp(-H(B,\sigma^2)r^2\delta^2 n) + 6\int_{\mathbb{R}^d} I(c(x) < r)p(x)dx) \\
&= 6\exp(-H(B,\sigma^2)r^2\delta^2 n) + 6P(\int k(X,z)p(z)dz < r)
\end{aligned}$$

□

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 3 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

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

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

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

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

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

and hence

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

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

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

Proof. Recalling (1), we have:

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

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 Generalizations

4.1 Second order GNW estimator $\hat{f}_{GNW,2}$

The proposed estimator $\hat{f}_{GNW}$ does not take advantage of the graph structure of the data. The estimator at a vertex v is based only on neighbours of v. In order to account for the potential influence of vertices which are not direct neighbours of v, we introduce the weights[3]

$$w_2(X_i, X) = \sum_{j=1, j \ne i}^n a(X_i, X_j) a(X_j, X)$$

We introduce the **Second order GNW estimator**:

$$\hat{f}_{GNW,2}(x) = \frac{\sum_{i=1}^n Y_i w_2(X_i, x)}{\sum_{i=1}^n w_2(X_i, x)}$$

[3]At this point we have not stated anything about self edges in the observed graph. As long as the variables $a(X_i, X_i)$ are bounded and independent, their contribution will vanish for large n so to simplify the exposition we assume that $a(X_i, X_i) = 0$.

Lemma 6 With probability at least $1-(2n+2)\exp(\frac{-2\delta^2(n-1)}{5B})$,

$$|\frac{1}{n(n-1)}\sum_{i=1}^n f(X_i)w_2(X_i,X) - \int\int f(z)k(w,z)k(w,X)p(z)p(w)dzdw| \le 2\delta$$

Proof.

$$\begin{aligned}\frac{1}{n(n-1)}\sum_{i=1}^n f(X_i)w_2(X_i,X) &= \frac{1}{n(n-1)}\sum_{j=1}^n[\sum_{i\neq j} f(X_i)a(X_i,X_j)]a(X_j,X)\\ &= \frac{1}{n}\sum_{j=1}^n[\frac{1}{n-1}\sum_{i\neq j} f(X_i)a(X_i,X_j) - \int f(z)k(X_j,z)p(z)dz]a(X_j,X)\\ &+ \frac{1}{n}\sum_{j=1}^n[\int f(z)k(X_j,z)p(z)dz]a(X_j,X)\end{aligned}$$

Given $1\le j\le n$, according to Corolary 1 applied to the $n-1$ variables $X_1,...X_{j-1},X_{j+1},...,X_n$, we have

$$|\frac{1}{n-1}\sum_{i\neq j} f(X_i)a(X_i,X_j) - \int f(z)k(X_j,z)p(z)dz| \ge \delta$$

with probability $\le 2\exp(-\frac{2\delta^2(n-1)}{5B})$ Hence, with probability $\ge 1-2n\exp(-\frac{2\delta^2(n-1)}{5B})$

$$|\frac{1}{n}\sum_{j=1}^n[\frac{1}{n-1}\sum_{i\neq j} f(X_i)a(X_i,X_j) - \int f(z)k(X_j,z)p(z)dz]a(X_j,X)| \le \frac{\delta}{n}\sum_{j=1}^n a(X_j,X) \le \delta$$

Applying Corolary 1 with $f_1(x)=\int f(z)k(x,z)p(z)dz$ (which is also bounded by B) , we have

$$|\frac{1}{n}\sum_{j=1}^n[\int f(z)k(X_j,z)p(z)dz]a(X_j,X) - \int\int f(z)k(w,z)k(w,X)p(z)p(w)dzdw| \ge \delta$$

with probability $\le 2\exp(-\frac{2\delta^2 n}{5B})$.

Hence with probability at least $1-(2n+2)\exp(\frac{-2\delta^2(n-1)}{5B})$, we have

$$|\frac{1}{n(n-1)}\sum_{i=1}^n f(X_i)w_2(X_i,X) - \int\int f(z)k(w,z)k(w,X)p(z)p(w)dzdw| \le 2\delta$$

□

Theorem 3 Assme that $P(\int\int k(X,w)k(w,z)p(w)p(z)dwdz=0)=0$. For any $r>0$

$$|\hat{f}_{GNW,2}(X) - \frac{\int\int f(z)k(z,w)k(w,X)p(z)p(w)dwdz}{\int\int k(z,w)k(w,X)p(z)p(w)dwdz}| \le \frac{(4r+2)\delta}{r^2}$$

with probability $\ge 1-P(\int\int k(X,z)k(z,w)p(z)p(w)dzdw<r) - c_1 n\exp(-H(B,\sigma^2)\delta^2(n-1))$.

Proof. Denote

$$C_r = \{\int\int k(X,w)k(w,z)p(w)p(z)dwdz \ge r\}$$
$$A_\delta(f) = \{|\frac{1}{n(n-1)}\sum_{i=1}^n f(x_i)w_2(x,X_i) - \int\int f(z)k(z,w)k(w,X)p(z)p(w)dzdw| \ge \delta\}$$

Applying Lemma 6 with $f=1$, we have

$$|\frac{1}{n(n-1)}\sum_{i=1}^n w_2(X_i,X) - \int\int k(w,z)k(w,X)p(z)p(w)dzdw| \le 2\delta$$

with probability at least $1-(2n+2)\exp(-\frac{2\delta^2 n}{5})$. In particular $\hat{f}_{GNW,2}(X)$ is well defined on $C_r \cap A_\delta(1)$ for any $\delta < \frac{r}{2}$. On this event we have

$$\hat{f}_{GNW,2}(X) = \frac{\frac{1}{n(n-1)}\sum_{i=1}^n f(X_i)w_2(X_i,X) - \int\int f(z)k(w,z)k(w,X)p(z)p(w)dzdw}{\frac{1}{n(n-1)}\sum_{i=1}^n w_2(X,X_i)}$$
$$+\frac{\int\int f(z)k(w,z)k(w,X)p(z)p(w)dzdw}{\frac{1}{n(n-1)}\sum_{i=1}^n w_2(X_i,X)} + \frac{\sum_{i=1}^n \epsilon_i w_2(X_i,X)}{\sum_{i=1}^n w_2(X_i,X)}$$

and

$$\frac{1}{\frac{1}{n(n-1)}w_2(X_i,X)} \leq \frac{2}{r}$$

Using the same technique as in Lemma 6, together with subgaussian concentration inequalities we can show that[4]

$$|\frac{1}{n(n-1)}\sum_{i=1}^n \epsilon_i w_2(X_i,X)| \geq \delta$$

holds with probability less than $c_1 n\exp(-C(\sigma^2)\delta^2(n-1))$ where $c_1, C(\sigma^2) > 0$.
On $C_r \cap A_\delta(1) \cap A_\delta(f)$ we have

$$|\frac{\frac{1}{n(n-1)}\sum_{i=1}^n f(X_i)w_2(X_i,X) - \int\int f(z)k(w,z)k(w,X)p(z)p(w)dzdw}{\frac{1}{n(n-1)}\sum_{i=1}^n w_2(X,X_i)}| \leq \frac{2\delta}{r}$$

Lastly, on $C_r \cap A_\delta(1)$ we have

$$|\frac{1}{\frac{1}{n(n-1)}\sum_{i=1}^n w_2(X_i,X)} - \frac{1}{\int\int k(X,z)k(z,w)p(z)p(w)dzdw}| \leq \frac{2}{r^2}\delta$$

On $C_r \cap A_\delta(1)^c \cap A_\delta(f)^c \cap N_\delta^c$ we have

$$|\hat{f}_{GNW,2}(X) - \frac{\int\int f(z)k(z,w)k(w,X)p(z)p(w)dwdz}{\int\int k(z,w)k(w,X)p(z)p(w)dwdz}| \leq \frac{4\delta}{r} + \frac{2\delta}{r^2}$$

Finally, a union bound gives

$$P(C_r^c \cup A_\delta(1) \cup A_\delta(f) \cup N_\delta) \leq P(\int\int k(X,z)k(z,w)p(z)p(w)dzdw < r) + c_1 n\exp(-H(B,\sigma^2)\delta^2(n-1))$$

□

Corollary 4 If $r = \inf_{x\in supp(p)} \int\int k(x,z)k(w,z)p(z)p(w)dzdw > 0$ then

$$|\hat{f}_{GNW,2}(X) - \frac{\int\int f(z)k(z,w)k(w,X)p(z)p(w)dwdz}{\int\int k(z,w)k(w,X)p(z)p(w)dwdz}| \leq \frac{(4r+2)\delta}{r^2}$$

with probability $\geq 1 - c_1 n\exp(-H(B,\sigma^2)\delta^2(n-1))$.

Proof. Follows immediately from Theorem 3, as

$$P(\int\int k(X,z)k(z,w)p(z)p(w)dzdw < r) = \int_{\mathbb{R}^d} I(\int\int k(x,w)k(w,z)p(w)p(z)dwdz < r)p(x)dx = 0$$

□

[4] The technical details can be provided later if necessary

4.2 m-th order GNW estimator $\hat{f}_{GNW,m}$

Given $1 \leq m \leq n$, we introduce the weights

$$w_m(X_i, X) = \sum_{J_i} \prod_{j=0}^{m-1} a(X_{i_j}, X_{i_{j+1}})$$

Here, $J_i = (i, i_1, ..., i_{m-1})$ is a m-tuple of distinct indicies with the convention that $i_0 = i$ and X_{i_m} is identified with X and the sum is taken over all such m-tuples J_i. We introduce the **GNW estimator of order m**:

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

Lemma 7 Assume $||f(X_1)||_\infty \leq B$. Then

$$|\frac{(n-m)!}{n!}\sum_{i=1}^n f(X_i)w_m(X_i,X) - \frac{(n-(m-1))!}{n!}\sum_{i=1}^n T_k(f)(X_i)w_{m-1}(X_i,X)| \geq \delta$$

with probability $\leq 2n^{m-1}\exp(-\frac{2\delta^2(n-(m-1))}{5B})$.

Proof.

$$\begin{aligned}
\frac{(n-m)!}{n!}\sum_{i=1}^n f(X_i)w_m(X_i,X) =& \frac{(n-m)!}{n!}\sum_{I=(i_0,i_1,\ldots,i_{m-1})} f(X_{i_0})\prod_{j=0}^{m-1} a(X_{i_j},X_{i_{j+1}}) \\
=& \frac{(n-m)!}{n!}\sum_{J=(i_1,\ldots,i_{m-1})}[\sum_{i_0\notin J} f(X_{i_0})a(X_{i_0},X_{i_1})]\prod_{j=1}^{m-1} a(X_{i_j},X_{i_{j+1}}) \\
=& \frac{(n-(m-1))!}{n!}\sum_J[\frac{\sum_{i_0\notin J} f(X_{i_0})a(X_{i_0},X_{i_1})}{n-(m-1)}]\prod_{j=1}^{m-1} a(X_{i_j},X_{i_{j+1}})
\end{aligned}$$

For fixed $(m-1)$-tuple J of distinct indices, applying Corollary 1 on the $n-(m-1)$ variables $X_{i_0}, i_0 \notin J$, we have

$$|\frac{\sum_{i_0\notin J} f(X_{i_0})a(X_{i_0},X_{i_1})}{n-(m-1)} - T_k(f)(X_{i_1})| \geq \delta$$

has probability $\leq 2\exp(-\frac{2\delta^2(n-(m-1)}{5B})$. There are exactly $\frac{n!}{(n-(m-1))!}$ distinct $(n-(m-1)$-tuples J. Applying Corollary 1 to every such tupple we get

$$|\frac{(n-m)!}{n!}\sum_{i=1}^n f(X_i)w_m(X_i,X) - \frac{(n-(m-1))!}{n!}\sum_{i=1}^n T_k(f)(X_i)w_{m-1}(X_i,X)| \geq \delta$$

with probability $\leq 2\frac{n!}{(n-(m-1))!}\exp(-\frac{2\delta^2(n-(m-1))}{5B})$

□

Theorem 4 There is a polynomial p_m of degree m such that the event

$$|\hat{f}_{GNW,m}(X) - \frac{T_k^m(f)(X)}{T_k^m(1)(X)}| \geq \frac{(r\alpha+\beta)\delta}{r^2}$$

has probability $\leq P(T_k^m(X) < r) + p_m(n)\exp(-H(B,\sigma)\delta^2(n-(m-1)))$

Proof. Given $1 \leq j \leq m$, applying Lemma 7, we get

$$\Delta_j = |\frac{(n-j)!}{n!}\sum_{i=1}^n T_k^{m-j}(1)(X)w_j(X_i,X) - \frac{(n-(j-1))!}{n!}\sum_{i=1}^n T_k^{m-(j-1)}(1)(X)w_{j-1}(X_i,X))| \geq \delta$$

with probability $\leq 2n^{j-1}\exp(-2\delta^2(n-(j-1))/5B)$

$$|\frac{(n-m)!}{n!}\sum_{i=1}^n w_m(X_i,X) - T_k^m(1)(X)| \leq \sum_{j-1}^m \Delta_j \leq m\delta$$

with probability $\geq 1 - p_m(n)\exp(-c_1\delta^2(n-(m-1)))$ where p_m is a polynomial with degree m. Denote

$$C_r^m = \{T_k^m(1)(X) \geq r\}$$
$$A_\delta = \{|\frac{(n-m)!}{n!}\sum_{i=1}^n w_m(X_i, X) - T_k^m(1)(X)| \geq m\delta\}$$

If $m\delta < r/2$, then on $C_r^m \cap A_\delta^c$ we have

$$\frac{1}{\frac{(n-m)!}{n!}\sum_{i=1}^n w_m(X_i, X)} \leq \frac{1}{r - m\delta} \leq \frac{2}{r}$$

Following a similar technique as in Theorem 3, we can arrive at a similar result[5]. □

4.3 Deterioration of concentration for $\hat{f}_{GNW,m}$

5 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:

$$\textit{Rectangular: } k(x,y) = \frac{1}{2}I(|x-y| < h)$$
$$\textit{Triangular: } k(x,y) = (1 - \frac{|x-y|}{h})I(|x-y| \leq h)$$
$$\textit{Parabolic (Epanechnikov): } k(x,y) = \frac{3}{4}(1 - (\frac{x-y}{h})^2)I(|x-y| \leq h)$$
$$\textit{Gaussian: } k(x,y) = \exp(-\frac{(x-y)^2}{h})$$
$$\textit{Laplacian: } k(x,y) = \exp(-\frac{|x-y|}{h})$$

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.

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 $f(x)$.

References

[AB02] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Reviews of Modern Physics* 74.1 (Jan. 2002), pp. 47–97. DOI: 10.1103/revmodphys.74.47. URL: https://doi.org/10.1103%2Frevmodphys.74.47.

[BCL11] Peter J. Bickel, Aiyou Chen, and Elizaveta Levina. "The method of moments and degree distributions for network models". In: *The Annals of Statistics* 39.5 (Oct. 2011). DOI: 10.1214/11-aos904. URL: https://doi.org/10.1214%2F11-aos904.

[BJR07] Béla Bollobás, Svante Janson, and Oliver Riordan. "The phase transition in inhomogeneous random graphs". In: *Random Structures and Algorithms* 31.1 (2007), pp. 3–122. DOI: 10.1002/rsa.20168. URL: https://doi.org/10.1002%2Frsa.20168.

[Cha15] Sourav Chatterjee. "Matrix estimation by Universal Singular Value Thresholding". In: *The Annals of Statistics* 43.1 (Feb. 2015). DOI: 10.1214/14-aos1272. URL: https://doi.org/10.1214%2F14-aos1272.

[KG00] Vladimir Koltchinskii and Evarist Giné. "Random Matrix Approximation of Spectra of Integral Operators". In: *Bernoulli* 6.1 (2000), pp. 113–167. ISSN: 13507265. URL: http://www.jstor.org/stable/3318636 (visited on 06/01/2022).

[5]More details should be added, but the argument is essentially the same once we take care of the denominator and use Lemma 7 when appropriate

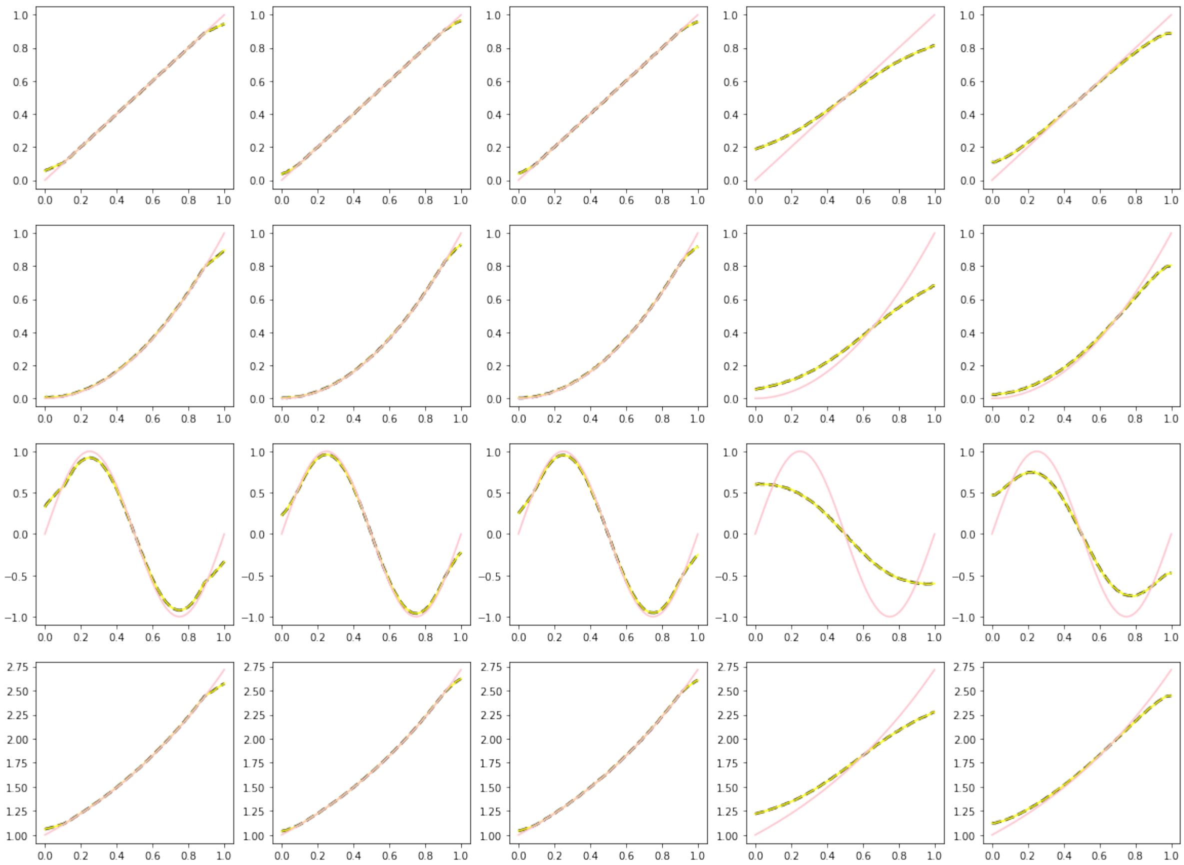

Figure 1: Each column represents a kernel, in the order listed above (rectangular, triangular, Epanechnikov, Gaussian, Laplacian). Each row represents a function in the following order $x, x^2, \sin(2\pi x), \exp(x)$. The pink line represents the true function, the yellow solid line is the plot of $\hat{f}_{GNW}$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

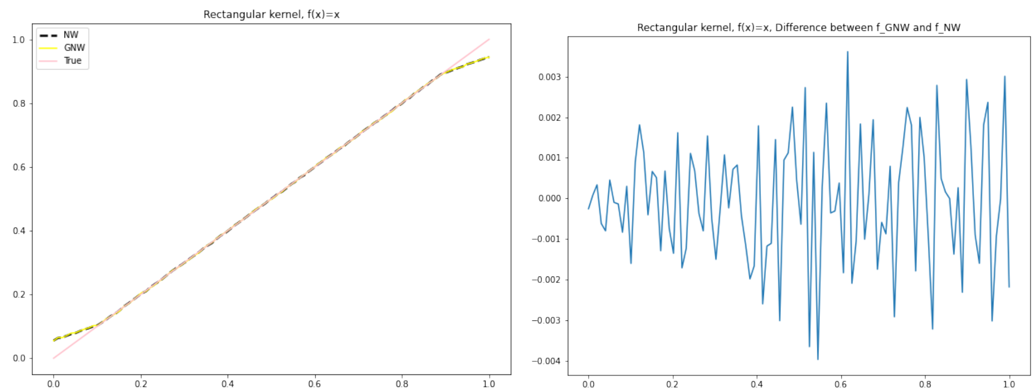

Figure 2: Left: comparison of $\hat{f}_{GNW}$, $\hat{f}_{NW}$ and f (solid yellow line, dashed black line and solid pink line, respectively. Right: Plot of $\hat{f}_{GNW} - \hat{f}_{NW}$.

[Oli09] Roberto Imbuzeiro Oliveira. *Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges*. 2009. DOI: 10.48550/ARXIV.0911.0600. URL: https://arxiv.org/abs/0911.0600.

[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.

[SN97] Tom Snijders and Krzysztof Nowicki. "Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure". In: *Journal of Classification* 14 (Jan. 1997), pp. 75–100. DOI: 10.1007/s003579900004.

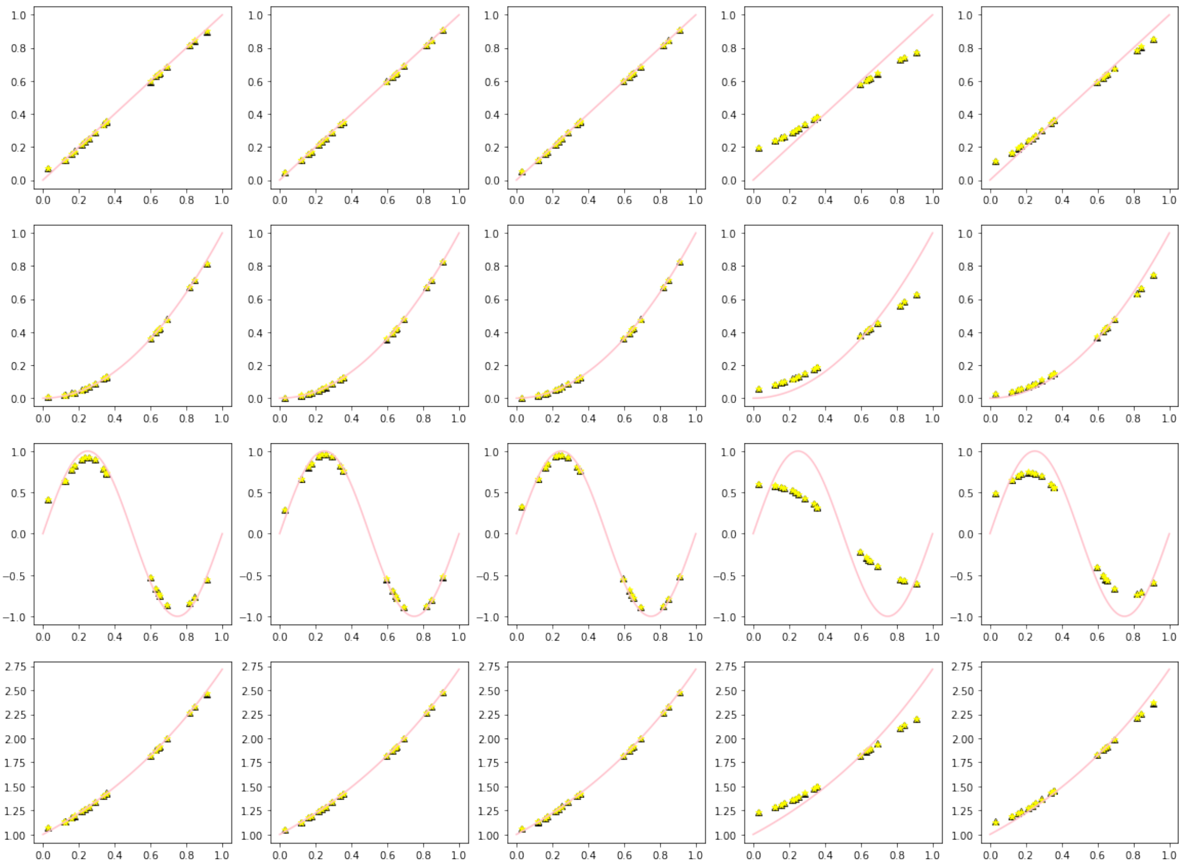

Figure 3: Each column represents a kernel in the order listed above. Each row represents a function as in Figure 1. We represent $\hat{f}_{GNW}$ with yellow triangle, $\hat{f}_{NW}$ with black star symbol and the true function with solid pink line.

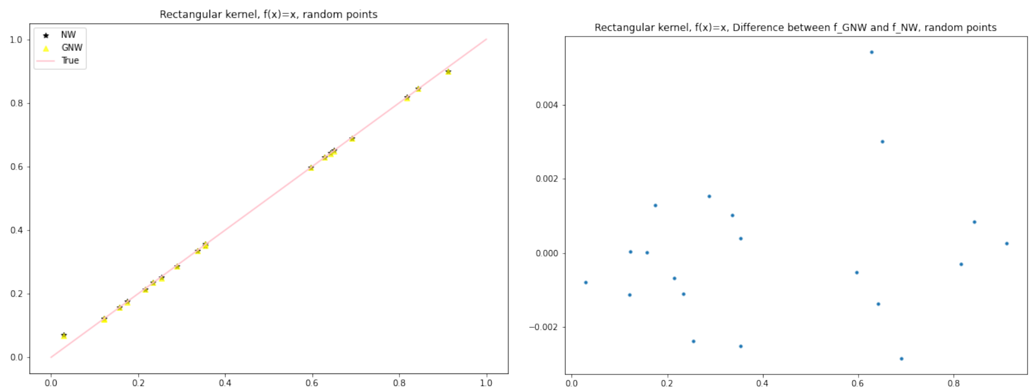

Figure 4: Left: comparison of scatter plots of $\hat{f}_{GNW}$, $\hat{f}_{NW}$ and the plot of f, represented with yellow triangles, black stars and solid pink line. Right: scatter plot of $\hat{f}_{GNW} - \hat{f}_{NW}$.

[TSP13] Minh Tang, Daniel L. Sussman, and Carey E. Priebe. "Universally consistent vertex classification for latent positions graphs". In: *The Annals of Statistics* 41.3 (June 2013). DOI: 10.1214/13-aos1112. URL: https://doi.org/10.1214%2F13-aos1112.

[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.