Graphical Nadaraya Watson estimator

Martin Gjorgjevski

May 2022

Contents

1 Introduction, motivation and notations

1.1 Nonparametric regression and the Nadaraya-Watson estimator

In the classical nonparametric regression setting we are given data $X_1, ..., X_n \in \mathbb{R}^d$ which is either fixed or 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 $Y_1, ..., Y_n$. Various such estimators are proposed in the literature, such as projection estimators which project the observation vector $Y = (Y_1, ..., Y_n)$ onto a subspace spanned by the data $X_1, ..., X_n$ (or potentially some embedding of the data $\phi(X_1), ..., \phi(X_n)$ where $\phi : \mathbb{R}^d \to \mathbb{R}^m$). Another popular type of estimators are the local polynomial estimators which estimate not only the function f in question but also several of its derivatives. For more details on nonparametric regression we refer to [Tsy08].

Kernels A kernel k on $\mathbb{R}^d$ is a symmetric function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. The kernel k is said to be

- positive semi definite if for any $x_1, ..., x_n \in \mathbb{R}^d$, and any $c_1, ..., c_n \in \mathbb{R}$ we have

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

- stationary if for all $x, y \in \mathbb{R}^d$,

$$k(x, y) = k(x - y)$$

- radial basis kernel if for all $x, y \in \mathbb{R}^d$

$$k(x,y) = k(||x-y||)$$

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. Finally, given $S \subseteq \mathbb{R}^d$, and a positive semi definite kernel k on $\mathbb{R}^d$, the restriction of k to $S \times S$ denoted with k_S is a positive semidefinite kernel on S. Conversely, if k_S is a positive semi definite kernel on S, then by letting $k(x,y) = k_S(x,y)$ if $(x,y) \in S \times S$ and $k(x,y) = 0$ otherwise, we get a positive semi definite kernel on $\mathbb{R}^d$. This fact allows us to work with $\mathbb{R}^d$ as ambient space for the data even though in certain situations we will be interested in compact subsets of $\mathbb{R}^d$.

The Nadaraya Watson estimator The Nadaraya Watson estimator is a special case of the local polynomial estimators. We assume that we are given a stationary kernel $k : \mathbb{R} \to \mathbb{R}$, a parameter $h > 0$ known as a bandwith. We also assume that $f : \mathbb{R} \to \mathbb{R}$ is in the Holder class $\Sigma(\beta, L)$, with $0 \leq \beta < 1$, that is for all $x, z \in \mathbb{R}$ we have

$$|f(x) - f(z)| \leq L|x-z|^\beta$$

The Nadaraya Watson estimator $\hat{f}_{NW}(x)$ of $f(x)$ is given by

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

It is easy to see that $\hat{f}_{NW}(x)$ is a solution to the following optimization problem

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

In the fixed design setting, assuming that there exists $\lambda_0 > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ and $x \in [0,1]$ we have $\frac{1}{nh} \sum_{i=1}^n k(\frac{X_i - x}{h}) \geq \lambda_0$ and with an additional assumption about the empirical density of the points $X_1, ..., X_n$, the mean squared error and mean integrated square error of $\hat{f}_n$ go to zero uniformly over $\Sigma(\beta, L)$ at a rate proportional to $n^{-\frac{2\beta}{2\beta+1}}$ ([Tsy08] p.40).

1.2 Latent Position Models

For a positive integer n, a positive definite kernel k on $\mathbb{R}^d$ taking values between 0 and 1 and a density p on $\mathbb{R}^d$ the Latent Position Model $LPM(n,k,p)$ is a model of random graph on n vertices $\{1, 2, ..., n\}$ generated as follows:

1. For each vertex i, $1 \leq i \leq n$, a sample X_i is drawn with distribution p. This variable is known as the latent position of node i

2. For each pair (i,j) with $i < j$,

$$a(i,j) = I(U_{i,j} \leq k(X_i, X_j))$$

   is a Bernoulli variable with parameter $k(X_i, X_j)$ determines whether there is an edge between i and j

3. The samples $X_1, ..., X_n$ are not observed and are assumed to be independent

4. The variables $U_{i,j}$, $1 \leq i < j \leq n$ are uniformly distributed on $[0,1]$, independent among themselves and from X_i, $1 \leq i \leq n$.

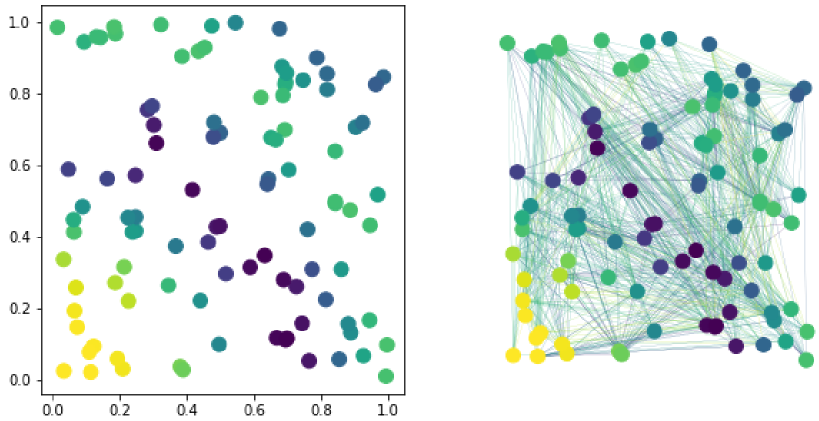

Figure 1: Latent positions and a corresponding graph sampled according to a Latent Position Model

Intuitively this means that we are more likely to observe an edge between two nodes which have positions that are similar with respect to k. Note that under these assumptions, edges which do not have common endpoints appear independently, but if two edges share the endpoint i then the appearence of both of those edges depends on X_i. Also, conditionally on the positions $X_1, ..., X_n$ all edges appear independently. We emphasize that contrary to the classical nonparametric approach we do not get to choose the kernel k. On the other hand, both in the nonparametric estimation ([Tsy08]) and in the random graph literature ([BCL11]) it is common to assume that the kernel depends on the sample size n.

1.3 Notation, framework and main results

Notation 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 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

$$T_k^m(f)(x) = \int_{\mathbb{R}^d} T_k^{m-1}(f(z))k(x, z)p(z)dz$$

Framework Given a Latent Position Model on $n+1$ vertices, we assume that we have additional information $Y_i = f(X_i) + \epsilon_i$, $1 \leq i \leq n$ with $f : \mathbb{R}^d \to \mathbb{R}$ and we are interested in estimating $f(X_{n+1})$. Let us denote $\mathcal{E}_n$ the sigma algebra generated by the edges $\{a(n + 1, i)|1 \leq i \leq n\}$. Analogously to the nonparametric case we can pose this question in three cases:

1. Assuming the other latent positions are known, that is estimate

$$E(f(X_{n+1})|X_1, X_2, ...X_n, \mathcal{E}_n)$$

2. Assuming only that the edges between $n+1$ and the other vertices are known i.e. estimating

$$E(f(X_{n+1})|\mathcal{E}_n)$$

3. Assuming that we know the latent position of the point we are trying to estimate i.e.

$$E(f(X_{n+1})|\mathcal{E}_n, X_{n+1} = x)$$

The first case is in general easier to analyze. Thus, in this report, we focus on the second case. The third special case is of interest in proving results about the second case. Inspired by the classical Nadaraya Watson estimator, we introduce the **Graphical Nadaraya Watson** estimator:

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

More generally, given a graph on n vertices G, a subset of $S \subseteq V(G)$ and variables Y_s for $s \in S$, we define $\hat{f}_{GNW} : V(G) - S \to \mathbb{R}$ with

$$\hat{f}_{GNW}(v) = \begin{cases} \frac{\sum_{s\in S} Y_s a(v,s)}{\sum_{s\in S} a(v,s)} & \text{if } \sum_{s\in S} a(v,s) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

The definition of the estimator is purely graphical, but we will still by a slight abuse of notation write $a(x, X_i)$ in place of $a(n+1, i)$ in the context of setting 3. Similarly, we will use $\hat{f}_{GNW}(x)$ to denote a prediction for a node which is identified with the position x. Also, we will often write X in place of X_{n+1} in setting 2. Thus,

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

We introduce the 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 $\int I(c(x) = 0)d\nu(x) = 0$[1].

Main results In this report we prove...

2 Concentration properties

In this section we show, using concentration inequalities, that for a fixed point $x \in \mathbb{R}^d$ with $c(x) > 0$ the Graphical Nadaraya Watson estimator $\hat{f}_{GNW}$ concentrates towards the quantity

$$\frac{T_k(f)(x)}{T_k(1)(x)} = \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}$$

for all bounded functions $f : \mathbb{R}^d \to \mathbb{R}$. This is done in context with the framework case 3. We also prove a concentration result for $\hat{f}_{GNW}(X)$ where X is random in the context of case 2. The concentration is exponential in the number of samples n and depends on the parameter $c(x) = T_k(1)(x) = \int k(x,z)p(z)dz$. The main idea is as follows: we use concentration inequalities to establish concentration of the numerator and denominator of $\frac{1}{n}\sum_{i=1}^n Y_i a(x, X_i)$ and $\frac{1}{n}\sum_{i=1}^n a(x, X_i)$. This is done in Lemma 1 and Lemma 2. In Theorem 1 we finish the proof using a union bound. In Corollary 2 we integrate the inequalities obtained in Theorem 1 over the support of pdx to get the result for the random case as well.

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

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

[1]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 $\nu = \mu = pdx$

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 ([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. □

In the following corollary we prove a concentration result analogous to Lemma 1 for the case when the

Corollary 1 Suppose that $f(X_1)$ is (essentially) bounded, measurable function with $||f(X_1)||_\infty \leq B$ and that $X, X_1, ..., X_n$ are i.i.d. with density p. Then

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

Proof. Let $U_1, ..., U_n$ be i.i.d. uniform on $[0, 1]$ such that $a(X, X_i) = I(U_i \leq k(X, X_i))$. Consider the indicator function $\phi : \mathbb{R}^{2n+1} \to \mathbb{R}$ given by

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

According to Lemma 1, we have

$$\begin{aligned} E\phi(x, X_1, ..., X_n, U_1, ..., U_n) &= \int \phi(x, x_1, ..., x_n, u_1, ..., u_n) \prod_{i=1}^{n} p(x_i) \prod_{i=1}^{n} dx_i \prod_{i=1}^{n} du_i \\ &= P(|\frac{1}{n}\sum_{i=1}^{n} f(X_i)a(x, X_i) - \int f(z)k(x, z)p(z)dz| \geq t) \leq 2\exp(-\frac{2t^2n}{5B^2}) \end{aligned}$$

Finally, we have

$$\begin{aligned} P(|\frac{1}{n}\sum_{i=1}^{n} f(X_i)a(X_i, X) - \int f(z)k(X, z)p(z)dz| \geq t) &= E\phi(X, X_1, ..., X_n, U_1, ..., U_n) \\ &= \int [\phi(x, x_1, .., x_n.u_1, ..., u_n) \prod_{i=1}^{n} p(x_i) \prod_{i=1}^{n} dx_i \prod_{i=1}^{n} du_i]p(x)dx \\ &= \int E\phi(x, X_1, ..., X_n, U_1, ..., U_n)p(x)dx \\ &\leq 2\exp(-\frac{2t^2n}{5B^2}) \int p(x)dx = 2\exp(-\frac{2t^2n}{5B^2}) \end{aligned}$$

□

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. □

Corollary 2 Suppose that and $\epsilon_1, ..., \epsilon_n$ are i.i.d. centered Gaussian variables with variance σ^2, $X, X_1, ..., X_n$ are i.i.d. with density p. Then

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

Proof. The result follows by Lemma 2 using the same method that was used to derive Corollary 1 from Lemma 1. We will omit the details. □

Theorem 1 (Concetnration of $\hat{f}_{GNW}(x)$ with x fixed) 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 $B_{\delta_2}^c$ 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 $B_{\delta_2}^c$, 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. □

Theorem 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 $r > 0$ and $0 < \delta < 3B$ we have $\delta^2 = \frac{r\delta}{6B}$. Applying Corollary 1, Corollary 2, and a union bound, we get

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

where $H(B,\sigma^2) = \min(\frac{c_1}{\sigma^2}, \frac{1}{90B^2})$

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 $x \in \mathbb{R}^d$.

For $\delta, r > 0$ and $f : \mathbb{R}^d \to \mathbb{R}$ bounded, let

$$C_r = \{\int k(X,z)p(z)dz \geq r\} = \{c(X) \geq r\}$$
$$A_\delta(f) = \{|\frac{1}{n}\sum_{i=1}^n f(X_i)a(X,X_i) - \int f(z)k(X,z)p(z)dz| \geq \delta\}$$
$$N_\delta = \{|\frac{1}{n}\sum_{i=1}^n \epsilon_i a(X,X_i)| \geq \delta\}$$

Let $\delta_1, \delta_2, \delta_3 > 0$ to be specified later. On $C_r \cap A_{\delta_1}(f)^c \cap A_{\delta_2}(1)^c \cap N_\delta^c$ we have

$$\frac{1}{n}\sum_{i=1}^n a(X,X_i) > c(X) - \delta_2 \geq r - \delta_2 \geq \frac{r}{2}$$

as soon as $\delta_2 < \frac{r}{2}$. Furthermore the same calculation as in Theorem 1 gives

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

Now we choose $\delta_1 = \delta_3 = \frac{r\delta}{6}$ and $\delta_2 = \min(\frac{r}{2}, \frac{r\delta}{6B})$, we get that on $C_r \cap A_{\delta_1}(f)^c \cap A_{\delta_2}(1)^c \cap N_\delta^c$ we have

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

To conclude, we note that when $\delta < 3B$ we have

$$\begin{aligned} P(C_r^c \cup A_{\delta_1}(f) \cup A_{\delta_2}(1) \cup N_\delta) &\leq P(c(X) < r) + 2\exp(-\frac{r^2\delta^2 n}{90B^2}) + 2\exp(-\frac{r^2\delta^2 n}{90B^2}) + 2\exp(-\frac{c_1 r^2\delta^2 n}{\sigma^2}) \\ &\leq P(C(X) < r) + 6\exp(-H(B,\sigma^2)r^2\delta^2 n) \end{aligned}$$

where $H(B,\sigma^2) = \min(\frac{c_1}{\sigma^2}, \frac{1}{90B^2})$.

$\square$

We conclude this section with several remarks on the presented results.

Remarks

Remark 1 (Influence of the noise and the bounedness constant) Larger constants $H(B,\sigma^2) = \min(\frac{C}{\sigma^2}, \frac{1}{90B^2})$ give better concentration rate in Theorem 1 and Theorem 2. On the other hand, $H(B,\sigma) \to \infty$ if and only if $B \to 0$ and $\sigma^2 \to 0$. In particular, letting $B \to \infty$ ruins the concentration property.

Remark 2 (Generalization of the noise) The proof of Lemma 2 relies on sub-gaussian inequalities. Those inequalities hold true for a wider class of probability distributions, namely for subgaussian variables. Thus similar results hold if one assumes that the variables ϵ_i are i.i.d subgaussian.

Remark 3 (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. Under weaker assumption such as $f(X_1) \in L^2$ one can use Chebyshev or Markov inequalities to find a concentration rate. One way to slightly generalize the function class (while preserving the strong concentration rate) 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 replace McDiarmid's bounded difference inequality with Hoeffding's inequality to obtain similar concentration result, where the constant B is replaced by an upper bound of the ψ_2 subgaussian norm $\psi_2(X_1)$.

Remark 4 (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 $\hat{f}_{NW}$, and the dimensionality of the data will play an important role here.

Remark 5 (Comparisson to classical Nadaraya Watson estimator) It is also easy to show with slight modification of the presented proofs, that the classical Nadaraya Watson estimator $\hat{f}_{NW}$ given by

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

satisfies

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

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 $c(x)$. Indeed, if we take

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

then one can easily show that $EF(X_1, ..., X_n, U_1, ..., U_n) = 0$ and similar ideas as in Lemma 1 apply. We omit the details.

Remark 6 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 Bias

The quantity $\frac{T_k(f)(x)}{T_k(1)(x)}$ discussed in the previous section is morally speaking, the expectation of $\hat{f}_{GNW}(x)$. Thus it is of interest to study $|\frac{T_k(f)}{T_k(1)} - f|$, a quantity which might be considered as the bias of the estimator $\hat{f}_{GNW}$. At this point we emphasize again that the kernel k is built in the definition of a Latent Position Model, and contrary to classical kernel based estimators, we can not choose k. With that in mind, we will provide some conditions on k under which $|\frac{T_k(f)}{T_k(1)} - f|$ is small. Suppose that $k : \mathbb{R}^d \to [0, K]$ is bounded and symmetric function. We suppose that the kernel in the Latent position model depends on the sample size in the following way:

$$k_n(x, z) = \lambda_n k(\frac{x - z}{h_n})$$

Here n is the sample size, $\lambda_n > 0$, $h_n > 0$ are positive numbers. It is common to assume that $\lambda_n \to 0$ [BCL11] and $h_n \to 0$ [Tsy08] as $n \to \infty$. We assume for simplicity that $\lambda_n \leq \frac{1}{K}$ [2]

[2] This assumption ensures that for all $x, z \in \mathbb{R}^d$, $0 \leq k_n(x, z) \leq 1$. Otherwise we would have to deal with the convex projection of k_n onto the interval $[0, 1]$.

Lemma 3 Suppose that $0 \in \text{supp}(k)$ and f is α-Holder continuous with $0 < \alpha \leq 1$.

- If k has compact support and X is a random variable with distribution p, then
$$||\frac{T_{k_n}(f)(X)}{T_{k_n}(1)(X)} - f(X)||_\infty = \sup_{x \in \text{supp}(p)} |\frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)} - f(x)| \leq C h_n^\alpha$$
where C is an absolute constant depending only on k and L.
- If $\int ||y||^2 k(y)dy < \infty$, p is β Holder continuous, with $0 < \beta \leq 1$ and $p(x) > 0$ then there exists $n(x) \in \mathbb{N}$ such that for all $n \geq n(x)$,
$$|\frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)} - f(x)| \leq c h_n^\alpha$$
with $c > 0$ an absolute constant depending on k and the Holder constants of f and p.

Proof. By a change of variables, we have
$$T_{k_n}(f)(x) = \lambda_n \int f(z) k(\frac{x-z}{h_n}) p(z) dz = \lambda_n h_n^d \int f(x + h_n y) k(y) p(x + h_n y) dy$$
Note that if $x \in \text{supp}\, p$, then by our assumption that k doesnt vanish in a neighbourhood of 0, We have
$$T_{k_n}(1)(x) = \lambda_n \int k(\frac{x-z}{h_n}) p(z) dz = \lambda_n h_n^d \int k(y) p(x + h_n y) dy > 0$$
$$\begin{aligned} |\frac{T_k(f)(x)}{T_k(1)(x)} - f(x)| &= |\frac{\int [f(x + h_n y) - f(x)] k(y) p(x + h_n y) dy}{\int k(y) p(x + h_n y) dy}| \\ &\leq L h_n^\alpha \frac{\int ||y||^\alpha k(y) p(x + h_n y) dy}{\int k(y) p(x + h_n y) dy} \end{aligned} \tag{2}$$
We note that the factor $\lambda_n > 0$ plays no role in this proof since it cancels out in the expression $\frac{T_{k_n}(f)}{T_{k_n}(1)}$.

- Suppose now that k is compactly supported and let $M > 0$ be a constant such that $k(x) = 0$ for $|x| > M$. Then, from (2) we get
$$|\frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)} - f(x)| \leq L M^\alpha h_n^\alpha$$
This bound is independent of $x \in \text{supp}\, p$ and of f in the Holder class, and this proves the first claim.
- Now suppose $p(x) > 0$ and p is β-Holder continuous. Then for any $G(y)$ such that both $\int |G(y)| k(y) dy < \infty$ and $\int ||y||^\beta |G(y)| k(y) dy < \infty$ we have
$$\begin{aligned} |\int G(y) k(y) p(x + h_n y) dy - p(x) \int G(y) k(y) dy| &= |\int G(y) k(y) [p(x + h_n y) - p(x)] dy| \\ &\leq h_n^\beta \int ||y||^\beta |G(y)| k(y) dy \end{aligned} \tag{3}$$
In particular, for $G(y) = ||y||^\alpha$ and $G(y) = 1$ we get that for n sufficiently large (potentially depending on x)
$$\int ||y||^\alpha k(y) p(x + h_n y) dy \leq c_1 p(x)$$
and
$$\int k(y) p(x + h_n y) dy \geq c_2 p(x)$$
where $c_1, c_2 > 0$ depend on k and the Holder constants of p. Hence, using these estimates in (2), we get
$$|\frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)} - f(x)| \leq c h_n^\alpha$$
where $c > 0$ is an absolute constant depending on k and the Holder constants of f and p.

□

Lemma 4 Suppose that p is β-Holder continuous, i.e. there is an $L > 0$ such that for all $x, z \in \mathbb{R}^d$

$$|p(x) - p(z)| \leq L||x - z||^\beta$$

Suppose also that $p(x) > 0$.

- if $\lambda_n h_n^d = \omega(\frac{1}{\sqrt{n}})$ then $\hat{f}_{GNW}(x) \to f(x)$ in probability.
- if $\lambda_n h_n^d = \omega(\sqrt{\frac{\log n}{n}})$ then $\hat{f}_{GNW}(x) \to f(x)$ almost surely.

Proof. We begin by observing that $c_n(x) = T_{k_n}(1)(x)$ is important in the concentration of $\hat{f}_{GNW}$. In order to keep the concentration property we need to have

$$\lim_{n\to\infty} c_n(x)^2 n = \infty \tag{4}$$

We recall the expression

$$c_n(x) = \lambda_n \int k(\frac{x-z}{h_n})p(z)dz = \lambda_n h_n^d \int k(y)p(x + h_n y)dy$$

Using the β-Holder assumption on p, we have

$$\begin{aligned}|\lambda_n^{-1} h_n^{-d} c_n(x) - p(x) \int k(y)dy| &= |\int k(y)[p(x + h_n y) - p(x)]dy| \\ &\leq \int k(y)|p(x + h_n y) - p(x)|dy \\ &\leq L h_n^\beta \int ||y||^\beta k(y)dy\end{aligned}$$

From here it is easy to see that there are $c_1, c_2 > 0$ such that for n sufficiently large (potentially depending on x),

$$c_1 p(x)\lambda_n h_n^d \leq c_n(x) \leq c_2 p(x)\lambda_n h_n^d \tag{5}$$

- Suppose that $\lambda_n h_n^d = \omega(\frac{1}{\sqrt{n}})$. Then using Theorem 1 and 5 we get that $\hat{f}_{GNW}(x) \to f(x)$ as $n \to \infty$.
- Suppose that $\lambda_n h_n^d = \omega(\sqrt{\frac{\log n}{n}})$. Then for n sufficiently large, $\exp(-c_1\lambda_n^2 h_n^{2d} n) \leq n^{-(1+r)}$ and hence by Borel-Cantelli lemma,

  $$\hat{f}_{GNW}(x) - \frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)} \to 0 \textit{ almost surely as } n \to \infty$$

  Now the result follows from Lemma 3 (which is a deterministic statement).

□

Remarks

Remark 7 (The effect of λ_n) As mentioned in the proof of Lemma 3, λ_n has no effect on the bias term. However, if we want to keep the consistency properties of $\hat{f}_{GNW}$ via Lemma 4, we see that shrinking λ_n forces h_n to increase, and as can be seen from Lemma 3 this loosens the bound on the bias term. In this sense the assumption $\lambda_n \geq \lambda > 0$ is optimal for convergence properties of $\hat{f}_{GNW}$.

Remark 8 (Bias-variance tradeoff) For the sake of simplicity, we suppose that $\lambda_n = 1$. Then Theorem 1 states that $P(|\hat{f}_{GNW}(x) - \frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)}| \geq \delta) \leq c_1 \exp(-c\delta^2 h_n^{2d} n)$. Since we want this probability to be small we need to have $h_n^d n \to \infty$. In fact the for the purpose of low variance, large values of h_n are good. However, for the purpose of low bias, as per Lemma 3, small values of h_n are preferred. In particular, we see that if $h_n = \omega(\frac{1}{n^{1/2d}})$ and $h_n = o(1)$ then a good bias-variance tradeoff has been achieved and consistency properties of $\hat{f}_{GNW}$ follow.

Remark 9 (Curse of dimensionality) We observe the well known phenomenon known as the curse of dimensionality, which states that sample complexities grow exponentially in the dimension of the data.

4 L^2 convergence

In this section we study the L^2 convergence of $\hat{f}_{GNW}$ at a fixed point x. We assume that $c(x) > 0$ and $f(X_1) \in L^{2+\rho}$, with $\rho > 0$. We prove that $E(\hat{f}_{GNW}(x) - \frac{T_k(f)(x)}{c(x)})^2 \to 0$ at a rate $\frac{1}{n}$. If $c(x) = 1$ then for every $1 \leq i \leq n$, $k(x, X_i) = 1$ almost surely, and hence $a(x, X_i) = 1$ almost surely. In this case $\hat{f}_{GNW}(x) = \frac{1}{n}\sum_{i=1}^n Y_i$, that is $\hat{f}_{GNW}$ coincides with empirical average estimator and the variance of $\hat{f}_{GNW}(x)$ is given by

$$Var(\hat{f}_{GNW}(x)) = \frac{1}{n^2}\sum_{i=1}^n Var(Y_i) = \frac{1}{n}Var(Y_1) = \frac{1}{n}(v^2 + \sigma^2)$$

where $v^2 = Var(f(X_1))$. Thus we restrict our attention to the case where $0 < c(x) < 1$.

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. As $\hat{f}_{GNW}(x) = 0$ on E_n, we have

$$E(\hat{f}_{GNW}(x) - \frac{T_k(f)(x)}{c(x)})^2 = (\frac{T_k(f)(x)}{c(x)})^2(1 - c(x))^n + E_*(\hat{f}_{GNW}(x) - \frac{T_k(f)(x)}{c(x)})^2$$

We emphasize the trivial inequality $E_*(Z) \leq E(Z)$ whenever Z is a nonnegative random variable. We also denote the event $A_n(\delta) = \{|\frac{1}{n}\sum_{i=1}^n a(x, X_i) - c(x)| \geq \delta\}$. We need to control the L^2 norm of various quantities. Recalling (1), and using Cauchy-Schwarz's inequality with $n = 3$, we have:

$$\begin{aligned} E_*|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}|^2 \leq 3E_*|&\frac{\frac{1}{n}\sum_{i=1}^n f(X_i)a(x, X_i) - \int f(z)k(x,z)p(z)dz}{\frac{1}{n}\sum_{i=1}^n a(x, X_i)}|^2 \\ &+ 3E_*|\frac{\sum_{i=1}^n \epsilon_i a(x, X_i)}{\sum_{i=1}^n a(x, X_i)}|^2 \\ &+ 3|\int f(z)k(x,z)p(z)dz|^2 E_*|\frac{1}{\frac{1}{n}\sum_{i=1}^n a(x, X_i)} - \frac{1}{c(x)}|^2 \end{aligned} \tag{6}$$

This is again done via concentration inequalities. To be precise, the only quantity which concentrates in this setting is the denominator $\frac{1}{n}\sum_{i=1} a(x, X_i)$. Unlike in the previous section, we can not simply ignore bad sets where concentration does not hold because the behaviour of $\hat{f}_{GNW}$ on such sets can affect the L^2 norm. To go around this issue, we assume that f is $L^{2+\rho}$ for some $\rho > 0$. In the remarks we comment on a regularized version of $\hat{f}_{GNW}$ which converges towards $\frac{T_k(f)(x)}{c(x)}$ for the class of L^2 functions. Lemma 3 is a useful result which is used for bounding the second and third summand in (6). Corollary 3, Lemma 4 and Lemma 5 bound the third, second and first summand in (6) respectively.

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

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

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

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

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. □

Corollary 3 For n sufficiently large[3], we have

$$\begin{aligned}
E_*|\frac{1}{\frac{1}{n}\sum_{i=1}^n a(x, X_i)} - \frac{1}{c(x)}|^2 &\leq \frac{[2\log(2) + 3\log(n) + 2\log(c(x))]^2}{n^2} + \frac{1}{2n(n^2c(x)^2+1)} \\
&\leq \frac{25\log^2(n)}{n^2} + \frac{4\log^2(c(x))}{n^2} + \frac{1}{2n(n^2c(x)^2+1)}
\end{aligned}$$

Proof. Set

$$f(\delta) = K_1\delta^p + K_2\exp(-K_3\delta^2)$$

with $K_1 = (\frac{2}{c(x)^2})^p$, $K_2 = 2^p(n^p + \frac{1}{c(x)^p})$ and $K_3 = 2n$. The goal is to minimize f in δ so that we get the tightest possible bound from Lemma 3. We have

$$f^{'}(\delta) = pK_1\delta^{p-1} - 2K_2K_3\delta\exp(-K_3\delta^2)$$

For general p it is not possible to find explicit solution to $f^{'}(\delta) = 0$, but for $p = 2$, $f^{'}(\delta) = 0$ is equivalent to

$$2K_1 = 2K_2K_3\exp(-K_3\delta)$$

From here we compute that $\delta = \frac{1}{2n}\log(2n(n^2c(x)^4 + c(x)^2))$ is optimal rate and consequently,

$$E_*|\frac{1}{\frac{1}{n}\sum_{i=1}^n a(x, X_i)} - \frac{1}{c(x)}|^2 \leq \frac{[\log(2n(n^2c(x)^4 + c(x)^2)]^2}{n^2} + \frac{1}{2n(n^2c(x)^2+1)}$$

as soon as n is large enoguh so that $\delta < \frac{c(x)}{2}$. This bound is tighter than what is claimed in the corollary. We trade off a bit of the tightness for a simpler upper bound. This is achieved by replacing the expression $n^2c(x)^4 + c(x)^2$ by $2n^2c(x)^2$. To obtain the second inequality, we expand

[3] explicit value of n_0 such that for $n > n_0$ this works is available and can be found within the proof

$$
\begin{aligned}
(2\log(2)+3\log(n)+2\log(c(x)))^2 &= (2\log(2)+3\log(n))^2+4\log^2(c(x))+2(2\log(2)+3\log(n))(\log(c(x)))\\
&\le (2\log(2)+3\log(n))^2+4\log^2(c(x))\\
&\le 25\log(n)+4\log^2(c(x))
\end{aligned}
$$

where we used the fact that $0 < c(x) < 1$ in the first inequality, and replaced $\log(2)$ with $\log(n)$ in the second inequality. □

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

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

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 $p=1$.

□

Lemma 5 Suppose that $f(X_1) \in L^{2+\rho}$ for some $\rho > 0$. Then

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

Proof. Consider $A_n(\delta) = \{|\frac{1}{n}\sum_{i=1}^n a(x,X_i) - c(x)| \ge \delta\}$. On $A_n(\delta)^c$, we have

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

as soon as $\delta \le \frac{1}{2}c(x)$. Set $\delta = \frac{1}{2}c(x)$.

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

$$\begin{aligned}
E_*\left(\frac{\frac{1}{n}\sum_{i=1}^n W_i}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} I(A_n(\delta)^c)\right)^2 &\leq \frac{4}{c(x)^2} E\left(\frac{1}{n}\sum_{i=1}^n W_i\right)^2 \\
&= \frac{4}{nc(x)^2} Var(W_1) \\
&= \frac{4}{nc(x)^2} EW_1^2 \\
&= \frac{4}{nc(x)^2}\left[\int f(z)^2 k(x,z)p(z)dz - \left(\int f(z)k(x,z)p(z)dz\right)^2\right] \\
&\leq \frac{4}{nc(x)^2}\int f(z)^2 k(x,z)p(z)dz \\
&\leq \frac{4}{nc(x)^2}\int f(z)^2 p(z)dz \\
&= \frac{4}{nc(x)^2} ||f(X_1)||_{L_2}^2 \\
&\leq \frac{4}{nc(x)^2} ||f(X_1)||_{L_{2+\rho}}^2
\end{aligned}$$

where we used the well known Lyapunov's inequality in the last line. 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

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

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

$$\begin{aligned}
|W_1|^{2+\rho} &= |f(X_1)a(x,X_1) - Ef(X_2)a(x,X_2)|^{2+\rho} \\
&= |E(f(X_1)a(x,X_1) - f(X_2)a(x,X_2)|X_1,U_1)|^{2+\rho} \\
&\leq E(|f(X_1)a(x,X_1) - f(X_2)a(x,X_2)|^{2+\rho}|X_1,U_1)
\end{aligned}$$

and hence

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

Again we obtain tighter inequalities than the presented ones. To obtain the form stated in the lemma, note that $2^{\frac{1}{1+\frac{2}{\rho}}} \leq 2$. We conclude by splitting the expectation on $A_n(\delta)$ and $A_n(\delta)^c$. □

Theorem 3 (L^2 convergence of $\hat{f}_{GNW}$) Suppose that $f(X_1) \in L^{2+\rho}$ for some $\rho > 0$. Then

$$nE_*\left(\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}\right)^2 \leq G(c(x), ||f(X_1)||_{2+\rho}, \sigma^2)$$

Proof. The hard work has already been done. Recalling (6), Corollary 3, Lemma 4 and Lemma 5, we have

$$
\begin{aligned}
E_*(\hat{f}_{GNW}(x) - \frac{T_k(f)(x)}{c(x)})^2 \leq 12(\frac{1}{nc(x)^2} &+ n^2 \exp(-\frac{\frac{1}{2}c(x)^2 n}{1+\frac{2}{\rho}})||f(X_1)||^2_{L^{2+\rho}} \\
&+ 3\frac{\sigma^2}{n}(\frac{1}{c(x)} + \frac{2\delta}{c(x)^2} + 2(n + \frac{1}{c(x)})\exp(-2\delta^2 n) \\
&+ 3(\frac{25\log^2(n)}{n^2} + \frac{4\log^2(c(x))}{n^2} + \frac{1}{2n(n^2c(x)^2+1)})||f(X_1)||^2_{2+\rho}
\end{aligned} \tag{7}
$$

From here we see that

$$
E_*(\hat{f}_{GNW}(x) - \frac{T_k(f)(x)}{c(x)})^2 \leq \frac{1}{n}G(c(x), ||f(X_1)||_{2+\rho}, \sigma^2)
$$

□

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 recalling (1) one only needs to show that $E_*|\frac{\frac{1}{n}\sum_{i=1}^n f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}|^p$ and $E_*|\frac{1}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} - \frac{1}{c(x)}|^p$ 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 argue by analogy with 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 $|\frac{\sum_{i=1}^n f(X_i)a(x,X_i)}{n}|^p$ is bounded in $L^{1+\frac{\rho}{p}}$ and hence uniformly integrable. Keeping in mind that $\frac{\sum_{i=1}^n f(X_i)a(x,X_i)}{n} \to \int f(z)k(x,z)p(z)dz$ in probability, it follows that

$$
E|\frac{\sum_{i=1}^n [f(X_i)a(x,X_i) - \int f(z)k(x,z)p(z)dz]}{n}|^p \to 0
$$

as $n \to \infty$.

Remark 7 (Regularization) The main problem with the estimator $\hat{f}_{GNW}$ is that in the case when there are too few edges. We can fix the L^2 convergence issue by considering the **Regularized Graphical Nadaraya Watson** estimator:

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

where $\lambda(\alpha,\beta) = \alpha I(\frac{1}{n}\sum_{i=1}^n a(x,X_i) \leq \beta c(x))$ with $\alpha \geq 0$ 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 $\hat{f}_{RGNW,\alpha,\beta}$. For notational convenience, let

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

We note that on $A_n(\delta)$ we have

$$
\sum_{i=1}^n a(x,X_i) + n\lambda(\alpha,\beta) \geq \min(\alpha,\beta)nc(x)
$$

so that

$$E([\frac{\sum_{i=1}^{n} W_i}{\sum_{i=1}^{n} a(x, X_i) + n\lambda(\alpha, \beta)}]I(A_n(\delta))^2 \leq G(x)E(\frac{1}{n}[\sum_{i=1}^{n} W_i]I(A_n(\delta)))^2$$

where $G(x) = \frac{1}{\min(\alpha,\beta)^2 c(x)^2}$. 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.

5 Higher order GNW estimators

In this section we discuss a generalization of the Graphical Nadaraya Watson which averages the observations Y_i over vertices which have fixed graph distance m from X. Using Corollary 1 we show that this estimator concentrates around the quantity $\frac{T_k^m(f)(X)}{c_m(X)}$ with probability

5.1 Second order GNW estimator $\hat{f}_{GNW,m}$

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[4]

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

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

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

Lemma 6 Suppose that $||f(X_1)||_\infty \leq B$, and $X, X_1, ..., X_n$ are i.i.d. with density p. Then

$$P(|\frac{1}{n(n-1)}\sum_{i=1}^{n} f(X_i)w_2(X_i, X) - T_k^2(f)(X)| \geq 2\delta) \leq (2n+2)\exp(\frac{-2\delta^2(n-1)}{5B})$$

Proof. Set

$$S_j = \frac{1}{n-1}\sum_{i\neq j} f(X_i)a(X_i, X_j)$$

We compute:

$$\begin{aligned}
\frac{1}{n(n-1)}\sum_{i=1}^{n} f(X_i)w_2(X_i, X) &= \frac{1}{n(n-1)}\sum_{j=1}^{n}[\sum_{i\neq j} f(X_i)a(X_i, X_j)]a(X_j, X) \\
&= \frac{1}{n}\sum_{j=1}^{n}[\frac{1}{n-1}\sum_{i\neq j} f(X_i)a(X_i, X_j) - \int f(z)k(X_j, z)p(z)dz]a(X_j, X) \\
&+ \frac{1}{n}\sum_{j=1}^{n}[\int f(z)k(X_j, z)p(z)dz]a(X_j, X) \\
&= \frac{1}{n}\sum_{j=1}^{n}[S_j - T_k(f)(X_j)]a(X_j, X) + \frac{1}{n}\sum_{j=1}^{n} T_k(X_j)a(X_j, X)
\end{aligned} \tag{8}$$

[4]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 $a(X_i, X_i) = 0$.

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

$$P(|S_j - T_k(f)(X_j)| \ge \delta) \le 2\exp(-\frac{2\delta^2(n-1)}{5B})$$

Hence, by a union bound we have

$$\begin{aligned} P(|\frac{1}{n}\sum_{j=1}^{n}[S_j - T_k(f)(X_j)]a(X_j, X)| \ge \delta)) &\le \sum_{i=1}^{n} P(|S_j - T_k(f)(X_j)| \ge \delta) \\ &\le 2n\exp(-\frac{2\delta^2(n-1)}{5B}) \end{aligned}$$

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

$$P(|\frac{1}{n}\sum_{j=1}^{n} T_k(f)(X_j)a(X_j, X) - T_k^2(f)(X)| \ge \delta) \le 2\exp(-\frac{2\delta^2 n}{5B})$$

Finally, combining the last two displays together with (8), we have

$$\begin{aligned} P(|\frac{1}{n(n-1)}\sum_{i=1}^{n} f(X_i)w_2(X_i, X) - T_k^2(f)(X)| \ge 2\delta) &\le P(|\frac{1}{n}\sum_{j=1}^{n}[S_j - T_k(f)(X_j)]a(X_j, X)| \ge \delta) \\ &+ P(|\frac{1}{n}\sum_{j=1}^{n} T_k(f)(X_j)a(X_j, X) - T_k^2(f)(X)| \ge \delta) \\ &\le (2n+2)\exp(-\frac{2\delta^2(n-1)}{5B}) \end{aligned}$$

□

Theorem 4 For any $r > 0$,

$$P(|\hat{f}_{GNW,2}(X) - \frac{T_k^2(f)(X)}{c_2(X)}| \ge \frac{(4r+2)\delta}{r^2}) \le P(c_2(X) < r) + c_1 n\exp(-H(B,\sigma^2)\delta^2(n-1))$$

Proof. Denote

$$\begin{aligned} C_r &= \{c_2(X) \ge r\} = \{\int\int k(X,w)k(w,z)p(w)p(z)dwdz \ge r\} \\ A_\delta(f) &= \{|\frac{1}{n(n-1)}\sum_{i=1}^{n} f(x_i)w_2(x, X_i) - T_k^2(f)(X)| \ge \delta\} \\ N_\delta &= \{|\frac{1}{n(n-1)}\sum_{i=1}^{n} \epsilon_i w_2(X_i, X)| \ge \delta\} \end{aligned}$$

As soon as $\delta < \frac{r}{2}$, on $C_r \cap A_\delta(1)^c$ we have

$$\begin{aligned} \hat{f}_{GNW,2}(X) = &\frac{\frac{1}{n(n-1)}\sum_{i=1}^{n} f(X_i)w_2(X_i, X) - T_k^2(f)(X)}{\frac{1}{n(n-1)}\sum_{i=1}^{n} w_2(X, X_i)} \\ &+ \frac{T_k^2(f)(X)}{\frac{1}{n(n-1)}\sum_{i=1}^{n} w_2(X_i, X)} + \frac{\sum_{i=1}^{n}\epsilon_i w_2(X_i, X)}{\sum_{i=1}^{n} w_2(X_i, X)} \end{aligned}$$

and

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

In particular $\hat{f}_{GNW,2}(X)$ is well defined on $C_r \cap A_\delta(1)^c$

Using the same technique as in Lemma 6, together with subgaussian concentration inequalities we can show that[5]

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

$$P(|\frac{1}{n(n-1)}\sum_{i=1}^{n}\epsilon_i w_2(X_i,X)| \geq \delta) \leq c_1 n \exp(-C(\sigma^2)\delta^2(n-1))$$

where $c_1, C(\sigma^2) > 0$.
On $C_r \cap A_\delta(1)^c \cap A_\delta(f)^c$ we have

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

Next, on $C_r \cap A_\delta(1)^c$ we have

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

Finally, on $C_r \cap A_\delta(1)^c \cap A_\delta(f)^c \cap N_\delta^c$ we have

$$|\hat{f}_{GNW,2}(X) - \frac{T_k^2(f)(X)}{c_2(X)}| \leq \frac{4\delta}{r} + \frac{2\delta}{r^2}$$

We conclude with a union bound

$$P(C_r^c \cup A_\delta(1) \cup A_\delta(f) \cup N_\delta) \leq P(C_r^c) + c_1 n \exp(-H(B,\sigma^2)\delta^2(n-1))$$

□

Corollary 4 If $r = \inf_{x\in \operatorname{supp} p} c_2(x) > 0$ then

$$P(|\hat{f}_{GNW,2}(X) - \frac{T_k^2(f)(X)}{c_2(X)}| \geq \frac{(4r+2)\delta}{r^2}) \leq c_1 n \exp(-H(B,\sigma^2)\delta^2(n-1))$$

Proof. Follows immediately from Theorem 3. □

Remarks

Remark 8 (Higher order GNW Estimators) 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**:

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

The case $m = 2$ which was discussed in the previous paragraph is can be used as an inductive step in proving a concentration inequality for $\hat{f}_{GNW,m}$. It can be shown in a simmilar manner in which Lemma 6 was shown that

$$P(|\frac{(n-m)!}{n!}\sum_{i=1}^{n} f(X_i)w_m(X_i,X) - \frac{(n-(m-1))!}{n!}\sum_{i=1}^{n} T_k(f)(X_i)w_{m-1}(X_i,X)| \geq \delta) \leq 2n^{m-1}\exp(-\frac{2\delta^2(n-(m))}{5B})$$

Remark 9 (Application to the Stochastic Block Model) The stochastic block model $SBM(n, W, p)$ (where n is a positive integer, W a $k \times k$ symmetric matrix with entries in $[0, 1]$ and $p = (p_1, ..., p_k)$ is such that $p_1 + ... + p_k = 1$ and $p_i > 0$, $i = 1, ..., n$) is a random graph model that can be defined as follows: Each node i belongs to one of k (disjoint) sets $B_1, ..., B_k$ independently from the other nodes $j \neq i$ and with $P(i \in B_l) = p_k$, $1 \leq l \leq k$. Then for any two blocks B_l, B_s, we have

$$P(i \sim j | i \in B_l, j \in B_s) = W_{l,s}$$

The nodes $1, 2, ..., n$ may be thought of as individuals, the sets B_l may be thought of as different blocks or communities and the parameters $W_{l,s}$ as the probability of connection between blocks B_l and B_s. The Stochastic block model is a special case of Latent position model. Indeed, p to be a uniform $[0, 1]$ distribution with kernel

$$k(x_i, x_j) = \sum_{l \leq s} W_{l,s} I(x_i \in B_l, x_j \in B_s)$$

where B_l is the semi-open interval with endpoints $\sum_{j=1}^{l-1} p_j$ and $\sum_{j=1}^{l} p_j$. Then it is easy to see that the probability of edge between two vertices i, j in this Latent position model is equal to the probability of edge between the vertices i, j in a Stochastic block model $SBM(n, W, p)$. From here and the previous remark it follows that edge related statistics such as $w_m(i, j)$ can identyify the individuals in blocks with high probability, under suitable assumptions that these statistics differ among blocks. We omit the details.

6 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 $[0, 1]^d$ and we compare $\hat{f}_{GNW}(x)$, $\hat{f}_{NW}(x)$, $T_k(f)(x)$ and $f(x)$. We will study by simulations how the sample size, the dimension of the data and the noise level affects the estimator. We will also study how the kernel and the function f itself influence the performance.

6.1 Error plots

In this subsection we investigate various errors $|\hat{f}_{GNW} - f|$ and $|\hat{f}_{GNW} - T_k(f)|$ by simulations. We consider a grid G of 100 equally spaced points in $[0, 1]$. We will consider the following quantities:

$$\textit{True maximum error: } TME(f, x) = \max_{x \in G} |\hat{f}_{GNW}(x) - f(x)|$$

$$\textit{True average error: } TAE(f, x) = \frac{1}{|G|} \sum_{x \in G} |\hat{f}_{GNW}(x) - f(x)|$$

$$\textit{True square error: } TSE(f, x) = \frac{1}{|G|} \sum_{x \in G} |\hat{f}_{GNW}(x) - f(x)|^2$$

$$\textit{Biased maximum error: } BME(f, x) = \max_{x \in G} |\hat{f}_{GNW}(x) - \frac{T_k(x)}{c(x)}|$$

$$\textit{Biased average error: } BAE(f, x) = \frac{1}{|G|} \sum_{x \in G} |\hat{f}_{GNW}(x) - \frac{T_k(x)}{c(x)}|$$

$$\textit{Biased square error: } BSE(f, x) = \frac{1}{|G|} \sum_{x \in G} |\hat{f}_{GNW}(x) - \frac{T_k(x)}{c(x)}|^2$$

On Figure 2 we plot these errors against the logarithm of the sample size. The top left and bottom left images on Figure 2 show empirically that with fixed bandwith of the kernel the estimator will converge towards $\frac{T_k(f)(x)}{c(x)}$, which in general is at a fixed distance away from $f(x)$ (i.e. in L^∞ norm). On the other hand the other images illustrate that in average, these errors decrease as sample size increases. The true errors are lower bounded by the bias term, while the biased errors go to zero.

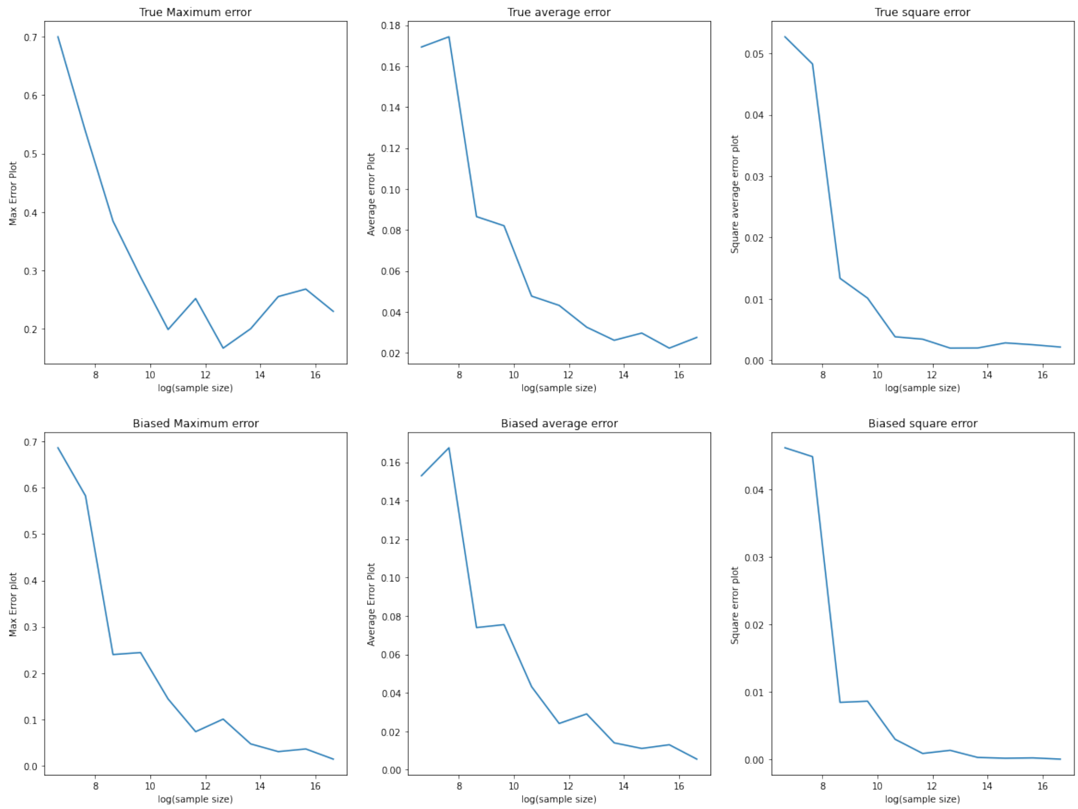

Figure 2: Error decay with sample size. The kernel is Gaussian $k(x, y) = \exp(-(\frac{x-y}{h})^2)$. The bandwith is set to $h = 0.11$, the noise is set to $\sigma^2 = 1$, and the function to be estimated is $f(x) = 2 + x - e^{-x^2}$. The experiment is repeated 20 times and the average error curves are plotted.

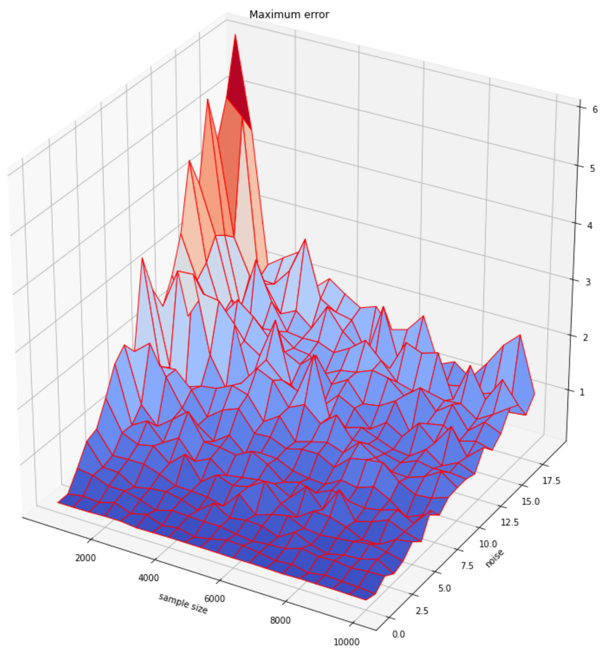

Figure 3: True maximum error plotted against sample size and noise level. The parameters are as in Figure 2.

6.2 Estimating functions

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.

[Ari+18] Ery Arias-Castro et al. *On the Estimation of Latent Distances Using Graph Distances*. 2018. DOI: 10.48550/ARXIV.1804.10611. URL: https://arxiv.org/abs/1804.10611.

[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).

[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.

[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.