Graphical Nadaraya Watson estimator

Martin Gjorgjevski

May 2022

Abstract

The topic of this report is a mixture of nonparametric staitstics and random graph theory. We analyze a simple and intuitive estimator in a graph regression setting. Given observations associated to a subset of the nodes, the estimator simply averages the observations over the neighbours of the node in question. We consider the latent position random graph model where each node i is associated to an unobserved random point X_i, these points being i.i.d. variables which take values in $\mathbb{R}^d$. Edges occur independently (conditionally on the latent positions), with the probability that nodes i and j are connected by an edge equal to $k(X_i, X_j)$ where k is a kernel on $\mathbb{R}^d$. While such an assumption on the data generating process may be an oversimplification for practical applications, it is a useful playground for a theoretical understanding quantities such as sample complexities and generalization bounds. Due to the connection with the classical Nadaraya Watson kernel based estimator, we call the proposed estimator the Graphical Nadaraya Watson estimator, denoted by $\hat{f}_{GNW}(x)$. In the nonparametric estimation literature it is well known that convolutional kernels need bandwith adaptation in order to achieve optimal performance, while in the graph learning literature it is well known that large social networks observed in practice are sparse (in terms of the adjacency matrix). In our setting we show that under certain conditions on k, these two phenomena are closely related. We show that for bounded functions f, $Var(\hat{f}_{GNW}(x)) = \Theta(1/\deg(x))$ and for bounded and sufficiently smooth functions f the mean square error of $E(\hat{f}_{GNW}(x) - f(x))^2 = O(\max(1/deg(x), deg(x)/n))$. As a consequence we conclude that the Graphical Nadaraya Watson estimator is consistent for arbitrarily sparse graphs. On the other hand, $\hat{f}_{GNW}$ suffers from the problems common in statstics such as bias-variance trade-off and the curse of dimensionality. Despite the simplicity of $\hat{f}_{GNW}$, we believe that these results will contribute to improve the theoretical understanding of more sophisticated graph learning architectures such as Graph Neural Networks (GNNs).

Contents

1 Introduction

1.1 Nonparametric regression and the Nadaraya-Watson estimator

Nonparametric regression In the classical nonparametric regression problem we are given data points $X_1, ..., X_n \in \mathbb{R}^d$ which are either fixed[1] or independent samples with common density p and noisy observations $Y_i = f(X_i) + \epsilon_i$. Here, $f\colon \mathbb{R}^d \to \mathbb{R}$ is an unknown function and in some suitable function class $\mathcal{F}$ and $\epsilon_1, ..., \epsilon_n$ are assumed to be i.i.d. centered variables 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 $m \in \mathbb{N}$.

The Nadaraya Watson estimator 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 the $n \times n$ matrix with (i,j)-th entry $k(x_i, x_j)$ is positive semi definite. If there exists a function $K\colon \mathbb{R}^d \to \mathbb{R}$ such that for all $x, y \in \mathbb{R}^d$, $k(x,y) = K(x-y)$ then k is said to be a stationary kernel. If, in addition, for all $x, y \in \mathbb{R}^d$, $k(x,y) = K(||x-y||)$ then k is said to be a radial basis kernel. The idea behind the Nadaraya Watson estimator is to *choose* a stationary kernel k and a bandwith parameter $h > 0$ and to estimate f by

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

The quality of the estimator in terms of the L^2-risk $E(\hat{f}_{NW}(x) - f(x))^2$ at a fixed point $x \in \mathbb{R}^d$ will depend on the many factors, such as the regularity of f, the choice of the kernel k but most importantly on the choice of bandwith $h > 0$. The L^2-risk admits the bias-variance decomposition

$$E(\hat{f}_{NW}(x) - f(x))^2 = Var(\hat{f}_{NW}(x)) + (E\hat{f}_{NW(x)} - f(x))^2 \tag{2}$$

The quantity $E(\hat{f}_{NW}(x) - f(x))$ is known as bias and under regularity assumptions on f such as α-Holder continuity, this term is upper bounded by Ch^α where C depends on k and α but not on the sample size n. We will give a more thorough discussion of the bias in Chapter 4. On the other hand, the variance $Var(\hat{f}_{NW}(x)$ tends to increase as h decreases, this is the famous bias-variance tradeoff phenomenon. In the fixed design setting optimal rates for h are available, and they are dependent on the sample size n (see **Tsybakov** section 1.5 and section 1.6, Proposition 1.13). The main takeaway is that the statistician should adapt h with respect to the sample size n.

[1] For example equally spaced points on the unit cube $[0,1]^d$

Linear estimators A general class of estimators have been extensively studied in the literature both theoretically and empirically is the class of linear estimators. 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 $Y_1, ..., Y_n$. For this class of estimators one typically takes $\mathcal{F}$ to be a Holder or a Sobolev space over a region $G \subseteq \mathbb{R}^d$. One generalization of the Nadaraya Watson estimator when $d = 1$ are the local polynomial estimators which aim to estimate not only f but also several of its derivatives $f^{(1)}, ..., f^{(l)}$. Another popular type of linear estimators are the projection estimators. They assume that f belongs in a the span[2] of a certain basis and then try to estimate coefficients with respect to said basis (e.g. trigonometric basis, wavelets and splines among others).

1.2 Random graphs and Latent Position Models

The Erdos-Renyi Model The most well known random graph model is the Erdos-Renyi random graph $G(n, p)$, for positive integer n and $0 \leq p \leq 1$. This model samples a random graph on n vertices with edges between vertices appearing independently with probability p. Results about $G(n, p)$ are stated asymptotically, that is as $n \to \infty$, for a certain range of values of p (more often than not depending on n), a graph sampled from $G(n, p)$ has a certain property with overwhelming probability. In their pioneering work, Erdos and Renyi show that for many graphical properties there is a sharp treshold p_c in the sense that for $p > p_c$ almost all $G(n, p)$ graphs have the property, while for $p < p_c$ almost none of them have the property. Classical examples are $p_c = 1/n$ for the emergence of the giant component and $p_c = \log(n)/n$ for connectedness.

Large Graphs from Real World Data Many large complex networks in the real world such as the World Wide Web, Movie actor collaboration networks, Citation Networks to name a few, have properties which are not present in the Erdos-Renyi graph. When such networks were compared to Erdos Renyi Random graph with same number of nodes and same average degree, it is observed that cliques in the Real World Networks form more often than in their Erdos-Renyi have counterpart, Similarly, Real World Networks have degree distributions which typically obey a power law, i.e. the proportion of vertices which have degree k is of the order $k^{-\alpha}$, while for their Erdos-Renyi model counterpart this should be a Poisson distribution. Such observations prompted the scientific community to consider different models for the data generating process which can better explain these phenomena (**Albert**, Section 2).

The Stochastic Block Model The Stochastic Block Model of Holland and Leindhart (**Holland1983StochasticBF**) naturally includes clustering of vertices. It assumes that n vertices are randomly sampled from K communities, and that conditionally on these communities edges form independently with probability depending only on the communities. More formally, $SBM(n, p, W)$ where n is a positive integer, $p = (p_1, ..., p_K)$ is K dimensional vector with $0 \leq p_l \leq 1$ and $\sum_{l=1}^{K} p_l = 1$ and W is a $K \times K$ symmetric matrix with entries $0 \leq w_{i,j} \leq 1$ generates edges on vertices $[n]$ by first randomly assigning a community $C_1, ..., C_K$ to each node with $P(i \in C_l) = p_l$ (these assignments are independent over distinct vertices) and then generating edges depending on the community of the endpoints of the edge, that is $P(i \sim j | i \in C_l, j \in C_s) = w_{ls}$. There are several questions in the SBM model such as deciding if an observed graph is indeed sampled by an SBM, under which conditions is there a way to fully or partially recover communities based on a single observed graph. The Stochastic Block Model has a long history in the statistics literature (**Snijders**). For recent developments we refer to (**Abbe**).

The Random Geometric Graph Another popular random graph model is the Geometric Random Graph, which is generated by sampling n independent random variables $X_1, ..., X_n \in \mathbb{R}^d$ and placing edges between nodes if the sampled points are sufficiently close, that is there is $h > 0$ such that $P(i \sim j) = I(||X_i - X_j|| \leq h)$. The value h controls how well connected the graph is, in the sense that smaller values of h give rise to sparser graphs. It is common to study how the behavior of certain graph statistics (such as degree distributions, average degrees and subgraph counts) changes with respect to h. A classical treatment of this topic is the Monograph of Penrose (**Penrose2003RandomGG**) .

[2] potentially closed linear span

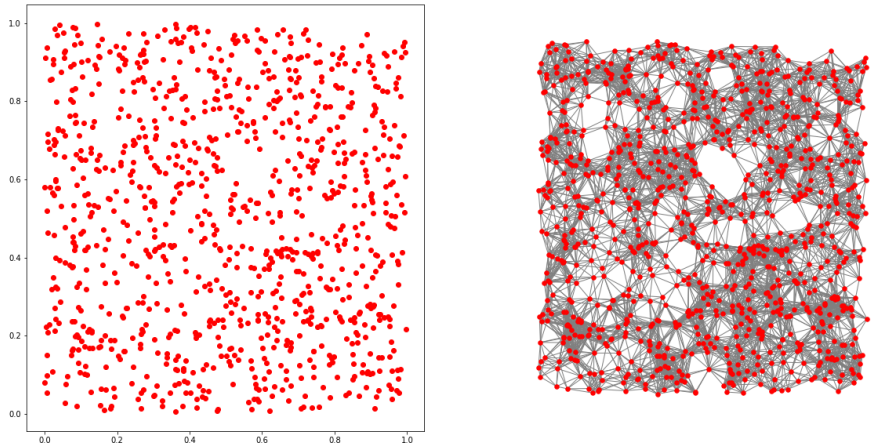

Figure 1: Random Geometric Graph with $n = 1000$ uniformly sampled points, $h = \sqrt{\frac{\log(n)}{n}}$

Latent Position Models The Latent Position Model was introduced by (**Hoff**). For a positive integer n, a 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) generates edges between vertices of $[n]$ in two stages; first a sample of size n of i.i.d. variables $X_1, ..., X_n$ with density p is drawn. The variable X_i can be thought of as the position of node i in the latent space. Next, given the sample $(X_1, ..., X_n)$ edges are drawn independently with $P(i \sim j | X_i, X_j) = k(X_i, X_j)$. 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. Both the Stochastic Block Model and the Random Geometric Graph can be thought of as a Latent Position Models with a suitable kernel k. If the kernel k is stationary then the graph structure of the LPM is in many respects similar to the graph structure of the Random Geometric Graph. On the other hand, The Stochastic Block Model can be represented as a Latent Position Model with $d = 1$, with a kernel k which can classify points that are arbitrarily close as being dissimilar. For the purposes of regression, the case where the kernel k is stationary is a much more natural setup.

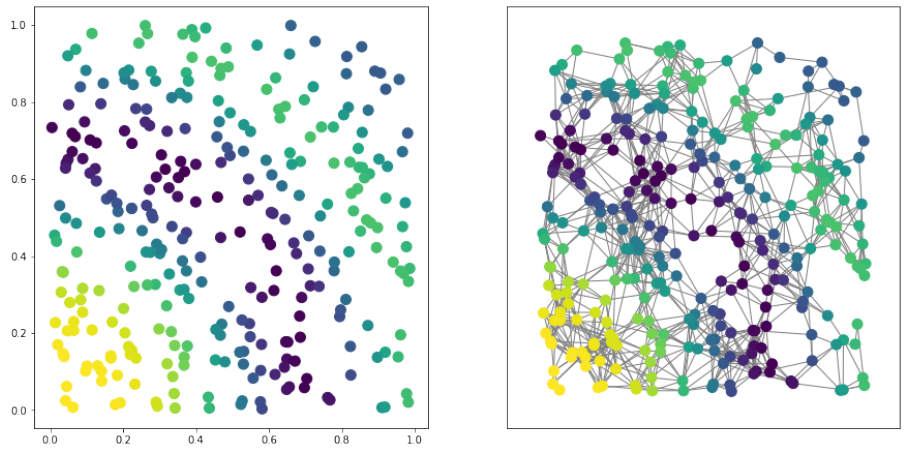

Figure 2: Latent position model with $n = 300$ nodes, Gaussian kernel, and bandwith $h = 0.75(\frac{\log(n)}{n})^{1/2}$

1.3 Framework

Framework We observe a random graph with vertex set $[n+1]$ sampled according to an LPM$(n+1, k_n, p)$ and assume that for nodes $i = 1, ...n$ (all but the last node) there is a label of the form $Y_i = f(X_i) + \epsilon_i$ with $f : \mathbb{R}^d \to \mathbb{R}$. Besides the graph itself and the explanatory variables $Y_1, ..., Y_n$ no other quantities

are assumed to be known. We write X in place of X_{n+1} and for $i = 1, ..., n$ we write $a(X, X_i)$ for the indicator that there is an edge between the node $n+1$ and node i. To describe edge generation more precisely, we assume that for $i = 1, ..., n$ the indicator of an edge between X and X_i is given by

$$a(X, X_i) = I(U_i \leq k(X, X_i))$$

where $U_1, ..., U_n$ are uniform variables on $[0, 1]$, such that $(U_1, ..., U_n, X_1, ..., X_n, X, \epsilon_1, ..., \epsilon_n)$ are jointly independent variables.

For $x \in \mathbb{R}^d$, we define

$$a(x, X_i) = I(U_i \leq k(x, X_i))$$

We introduce the **Graphical Nadaraya Watson** estimator given by

$$\hat{f}_{GNW}(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i a(x,X_i)}{\sum_{i=1}^n a(x,X_i)} & \text{if } \sum_{i=1}^n a(x, X_i) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

We are interested in two types of risk which we consider as the main measure of performance of $\hat{f}_{GNW}(x)$.

- For a fixed point $x \in \operatorname{supp} p$, we are interested in a bound on the risk at the point x given by

$$E(\hat{f}_{GNW}(x) - f(x))^2 \tag{4}$$

- For the random variable X which represents node $n+1$ we define the risk at the random point X

$$E(\hat{f}_{GNW}(X) - f(X))^2 \tag{5}$$

To our knowledge there are no results in on graph regression in this context.

1.4 Expected local degrees

We introduce the quantity

$$c_n(x) = \int_{\mathbb{R}^d} k_n(x, z)p(z)dz \tag{6}$$

Heuristically speaking, when conditioned on $X = x$ the degree of X is $\sum_{i=1}^n a(x, X_i)$. Hence we define the expected local degree at x with

$$d_n(x) = nc_n(x) \tag{7}$$

1.5 Classical and Graphical Nadaraya Watson estimators: similarities and differences

Because of the kernel based data generating process in the Latent position model, the Graphical Nadaraya Watson Estimator has strong connections with the classical Nadaraya Watson estimator. We emphasize the fact that while in classical nonparametric estimation the choice of the kernel is up to the statistician, in our setup the kernel k_n is given with the Latent Position Model, and hence it represents an oracle quantity. We define the integral operator $T_{k_n}(\cdot, x)$ on the set of bounded functions $f \colon \mathbb{R}^d \to \mathbb{R}$ by

$$T_k(f, x) == \int_{\mathbb{R}^d} f(z)k_n(x, z)p(z)dz \tag{8}$$

It is easy to see that

$$T_{k_n}(f, x) = Ef(X_i)a(x, X_i) \tag{9}$$

Note also that $c_n(x) = T_{k_n}(1, x)$. As a rough heuristic[3] to motivate further discussion, we expect that $\frac{1}{n}\sum_{i=1} f(X_i)a(x, X_i) - T_{k_n}(f, x) \to 0$. As $\hat{f}_{GNW}(x)$ is a quotient of $\frac{1}{n}\sum_{i=1}^n Y_i a(x, X_i)$ and

[3] If k_n did not depend on n this heuristic would be a fact, due to the law of large numbers

$\frac{1}{n}\sum_{i=1}^{n} a(x, X_i)$ which are close to $T_{k_n}(f, x)$ and $c_n(x)$ respectively, it is reasonable to expect that $\hat{f}_{GNW}(x)$ will be close to

$$b_n(f, x) = \frac{T_{k_n}(f, x)}{c_n(x)} \tag{10}$$

Note that the same rough heuristic applies to the classical Nadaraya Watson estimator by replacing $a(x, X_i)$ by $k_n(x, X_i)$.

1.6 Strategy

The risk at the point x (4) is much easier to control compared to the risk at the random point X (5), so first we give the main ideas on how to bound (4) from above. The strategy is to decompose Equation (4) into a variance term

$$E(\hat{f}_{GNW}(x) - b_n(f, x))^2 \tag{11}$$

and a bias term

$$b_n(f, x) - f(x) \tag{12}$$

and bound those terms separately. The tools used to the variance term (11) are probabilistic in nature such as decoupling arguments and concentration inequalities. On the other hand the tools needed to bound the bias term (12) are of geometric nature, as they basically ammount to studying convergence of certain integral operators towards the identity. We will show that in a general Latent position model the variance term (11) satisfies

$$\frac{C_1(\sigma^2)(1 - e^{-d_n(x)})}{d_n(x))} \leq E(\hat{f}_{GNW}(x) - b_n(f, x))^2 \leq \frac{C_2(B, \sigma^2)}{d_n(x}))$$

where $d_n(x)$ is the local average degree given by Equation (7). On the other hand the bias in a general Latent position model may behave poorly. We will restrict our attention to the geometric setting where the kernel k_n is of the form $k_n(x, z) = \rho_n \frac{K(x-z}{h_n}$ which will mimic the properties of the Random Geometric Graph. Pointwise results in this setting are still easy to obtain. Finally, once we establish upper bounds on Equation (12) and (11) we may integrate them against the p to obtain bounds for the risk at the random point X. This strategy requires strong assumption on the distribution in order to work.

Outline

2 Bounding the variance at a point

2.1 Computing $E\hat{f}_{GNW}(x)$

The risk at the point x (4) admits the decomposition

$$E(\hat{f}_{GNW}(x) - f(x))^2 = Var(\hat{f}_{GNW}(x)) + (E\hat{f}_{GNW}(x) - f(x))^2 \tag{13}$$

Equation (13) is known as the bias-variance decomposition of risk. If we want asymptotically vanishing risk at the point x, i.e. $E(\hat{f}_{GNW}(x) - f(x))^2 \to 0$ as $n \to \infty$, we need $E(\hat{f}_{GNW}(x)) \to f(x)$ as $n \to \infty$. Thus it is of basic interest to compute $E(\hat{f}_{GNW}(x))$, at least asymptotically. Being a quotient of two random variables, the exact value of $E\hat{f}_{GNW}(x)$ may seem difficult to compute. In this section we compute explicitly $E\hat{f}_{GNW}(x)$ for all x such that $c_n(x) > 0$. This is done via a decoupling argument. The method used here will be used again to bound the risk at the point x.

Decoupling argument Let $I \subseteq [n]$. For $I = \emptyset$ we define

$$R_\emptyset(x) = \begin{cases} \frac{1}{\sum_{i=1}^n a(x,X_i)} & \text{if } \sum_{i=1}^n a(x,X_i) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

and for $I \subseteq [n]$, $I \neq \emptyset$ we define

$$R_I(x) = \frac{1}{|I| + \sum_{j \notin I} a(x, X_j)}$$

For convenience of notation we write $R_i(x) = R_{\{i\}}(x)$. Note that for all pairs of disjoint subsets $I, J \subseteq [n]$ we have

$$R_J(x) \prod_{i \in I} a(x, X_i) = R_{I \cup J}(x) \prod_{i \in I} a(x, X_i) \tag{15}$$

and $R_{I\cup J}(x)$ is independent from $\{a(x, X_i) | i \in I\}$. In the special case when $I = \{i\}$ and $J = \emptyset$, Equation (15) becomes

$$R_\emptyset(x) a(x, X_i) = R_i(x) a(x, X_i) \tag{16}$$

This simple observation makes the computation of $E\hat{f}_{GNW}(x)$ possible.

Lemma 2.1. *For all $i - 1, 2, ..., n$ we have*

$$ER_i(x) = \frac{1 - (1 - \frac{d_n(x)}{n})^n}{d_n(x)}$$

Proof. Note that $R_i(x)$, $i = 1, 2, ..., n$ are identically distributed, hence $ER_i(x) = ER_1(x)$ for $i = 2, ..., n$. By summing Equations (16) for $i = 1, 2, 3, ..., n$ we have

$$\sum_{i=1}^n a(x, X_i) R_i(x) = R_\emptyset(x) \sum a(x, X_i) = I(\sum_{i=1}^n a(x, X_i) > 0) \tag{17}$$

Taking expectation and using the fact that $R_i(x)$ and $a(x, X_i)$ are independent, we get

$$\begin{aligned} E(\sum_{i=1}^n a(x, X_i) R_i(x)) &= \sum_{i=1}^n E(a(x, X_i) R_i(x)) \\ &= \sum_{i=1}^n E(a(x, X_i)) ER_i(x) \\ &= n c_n(x) ER_1(x) \end{aligned} \tag{18}$$

On the other hand,

$$P(\sum_{i=1}^n a(x, X_i) > 0) = 1 - P(\sum_{i=1}^n a(x, X_i) = 0) = 1 - (1 - c_n(x))^n \tag{19}$$

The result follows by combining Equations (17), (18) and (19).

□

Corollary 2.2.

$$E\hat{f}_{GNW}(x) = b_n(f, x)(1 - (1 - c_n(x))^n)$$

Proof. By Equation (15) we have

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

Hence, taking expectation and using Lemma 2.1, we get

$$
\begin{aligned}
E\hat{f}_{GNW}(x) &= \sum_{i=1}^{n} EY_i a(x, X_i) R_i(x) \\
&= \sum_{i=1}^{n} EY_i a(x, X_i) ER_i(x) \\
&= nEY_1 a(x, X_1) ER_1(x) \\
&= \frac{T_{k_n}(f)(x)(1-(1-c_n(x))^n)}{c_n(x)} \\
&= b_n(f,x)(1-(1-c_n(x))^n)
\end{aligned}
$$

□

2.2 The decoupling argument

In the previous section we found an explicit expression for $E\hat{f}_{GNW}(x)$ (Corollary 2.2). While this makes computation of $Var(\hat{f}_{GNW}(x)) = E(\hat{f}_{GNW}(x) - E\hat{f}_{GNW}(x))^2$ possible, we find that it is much simpler to analyse $E(\hat{f}_{GNW}(x) - b_n(f,x))^2$ instead. Note that by definition

$$
I(\sum_{i=1}^{n} a(x, X_i) = 0)\hat{f}_{GNW}(x) = 0 \tag{20}
$$

or equivalently

$$
I(\sum_{i=1}^{n} a(x, X_i) > 0)\hat{f}_{GNW}(x) = \hat{f}_{GNW}(x) \tag{21}
$$

Thus

$$
\begin{aligned}
E(\hat{f}_{GNW}(x) - b_n(f,x))^2 &= E[(\hat{f}_{GNW}(x) - b_n(f,x))^2 I(\sum_{i=1}^{n} a(x, X_i) > 0))] \\
&\quad + E[(\hat{f}_{GNW}(x) - b_n(f,x))^2 I(\sum_{i=1}^{n} a(x, X_i) = 0)] \\
&= E(\hat{f}_{GNW}(x) - b_n(f,x) I(\sum_{i=1}^{n} a(x, X_i) > 0))^2 + b_n^2(f,x) P(\sum_{i=1}^{n} a(x, X_i) = 0)
\end{aligned} \tag{22}
$$

Using Equation (15) and Equation (17) we have

$$
\hat{f}_{GNW}(x) - b_n(f.x) I(\sum_{i=1}^{n} a(x, X_i) > 0) = \sum_{i=1}^{n} (Y_i - b_n(f,x)) a(x, X_i) R_i(x) \tag{23}
$$

In Lemma 2.3 we show that the summands in this representation are uncorrelated and consequently obtain tractable expression for $E(\hat{f}_{GNW}(x) - b_n(f.x) I(\sum_{i=1}^{n} a(x, X_i) > 0))^2$. In contrast, computing the variance $Var(\hat{f}_{GNW}(x))$ directly leaves a much more complicated expression.

Lemma 2.3.

$$
E(\hat{f}_{GNW}(x) - b_n(f,x) I(\sum_{i=1}^{n} a(x, X_i) > 0))^2 = n[E(f(X_1) - b_n(f,x))^2 a(x, X_i) + \sigma^2 c_n(x)] ER_1^2(x)
$$

Proof. Using Equation (17), we have

$$E(\hat{f}_{GNW}(x) - b_n(f,x)I(\sum_{i=1}^{n} a(x,X_i) > 0))^2 = E(\sum_{i=1}^{n}(Y_i - b_n(f,x))a(x,X_i)R_i(x))^2$$
$$= \sum_{i=1}^{n} E(Y_i - b_n(f,x))a(x,X_i)R_i(x))^2$$
$$+ \sum_{i\neq j} E((Y_i - b_n(f,x))(Y_j - b_n(f,x))a(x,X_i)a(x,X_j)R_i(x)R_j(x)) \tag{24}$$

For $i \neq j$, using Equation (15), together with the fact that $R_{i,j}(x)$ is independent from $Y_i, Y_j, a(x,X_i)$ and $a(x,X_j)$, as well as the fact that the pairs $(Y_i, a(x,X_i)$ and $(Y_j, a(x,X_j)$ are independent, we have

$$\begin{aligned} &E[(Y_i - b_n(f,x))(Y_j - b_n(f,x))a(x,X_i)a(x,X_j)R_i(x)R_j(x)] = \\ &E[(Y_i - b_n(f,x))(Y_j - b_n(f,x))a(x,X_i)a(x,X_j)R_{i,j}^2(x)] = \\ &E[(Y_i - b_n(f,x))a(x,X_i)]E[(Y_j - b_n(f,x)a(x,X_j))]ER_{i,j}^2(x) = 0 \end{aligned} \tag{25}$$

Furthermore,

$$\begin{aligned} \sum_{i=1}^{n} E[(Y_i - b_n(f,x)^2 a(x,X_i)R_i^2(x)] &= \sum_{i=1}^{n} E((Y_i - b_n(f,x))^2 a(x,X_i)ER_i^2(x) \\ &= nE[((Y_1 - b_n(f,x))^2 a(x,X_i)]ER_1^2(x) \\ &= n[E(f(X_1) - b_n(f,x))^2 a(x,X_i) + \sigma^2 c_n(x)]ER_1^2(x) \end{aligned} \tag{26}$$

□

2.3 Lower bounds

We show that the presence of noise alone is sufficient for a lower bound of $E(\hat{f}_{GNW}(x) - b_n(f,x))^2$ that is of order $\frac{1}{d_n(x)}$.

Lemma 2.4.

$$E(\hat{f}_{GNW}(x) - b_n(f,x))^2 \geq \frac{\sigma^2(1 - e^{-d_n(x)})^2}{d_n(x)}$$

Proof. By Equation (22), Lemma 2.2, Lemma 2.3 and the basic inequality $1 - t \leq e^{-t}$ valid for all $t \geq 0$, we have

$$\begin{aligned} E(\hat{f}_{GNW}(x) - b_n(f,x))^2 &\geq E(\hat{f}_{GNW}(x) - b_n(f,x)I(\sum_{i=1}^{n} a(x,X_i) > 0))^2 \\ &= n[E(f(X_1) - b_n(f,x))^2 a(x,X_i) + \sigma^2 c_n(x)]ER_1^2(x) \\ &\geq \sigma^2 n c_n(x) ER_1^2(x) \\ &\geq \frac{\sigma^2(1 - (1 - c_n(x))^n)^2}{n c_n(x)} \\ &\geq \frac{\sigma^2(1 - e^{-n c_n(x)})^2}{n c_n(x)} \end{aligned} \tag{27}$$

□

2.4 Upper bounds

Lemma 2.5. *For $n \geq 3$*

$$E(\hat{f}_{GNW}(x) - b_n(f,x)I(\sum_{i=1}^{n} a(x,X_i) > 0))^2 \leq (4B^2 + \sigma^2)(\frac{65}{d_n(x)})$$

Proof. Recalling Lemma 2.3 and using the fact that $||f||_\infty \leq B$, we have

$$n[E(f(X_1) - b_n(f,x))^2 a(x,X_i) + \sigma^2 c_n(x)]ER_1^2(x) \leq (4B^2 + \sigma^2)nc_n(x)ER_1^2(x) \tag{28}$$

Hence it suffices to control $ER_1^2(x)$. We do this by splitting the expectation on the event that we observe at least $\frac{1}{2}(n-1)c_n(x)$ edges from $a(x,X_i)$, $i = 2,...,n$ and on it's complement. Let

$$A(x) = \{\sum_{i=2}^{n} a(x,X_i) \geq \frac{1}{2}(n-1)c_n(x)\} \tag{29}$$

For $n \geq 2$ we have

$$ER_1^2(x)I(A(x)) \leq \frac{1}{(1+\frac{1}{2}(n-1)c_n(x))^2}P(A(x)) \leq \frac{16}{n^2c_n^2(x)} \tag{30}$$

An application of Bernstein's inequality for bounded distributions (**vershynin** Theorem 2.8.4, page 39) with $a(x,X_i) - c_n(x)$, $i = 2,3,...n$ as the bounded, centered and independent variables yields

$$P(|\sum_{i=2}^{n} a(x,X_i) - (n-1)c_n(x)| \geq t) \leq 2\exp(-\frac{t^2}{(n-1)c_n(x)(1-c_n(x)) + \frac{t}{3}}) \tag{31}$$

Setting $t = \frac{1}{2}(n-1)c_n(x)$ in Equation (31) together with the observation that $A^c(x)$ implies

$$|\sum_{i=2}^{n}(a(x,X_i) - c_n(x))| \geq \frac{1}{2}(n-1)c_n(x)$$

For $n \geq 3$, we get

$$\begin{aligned} P(A^c(x)) &\leq P(|\sum_{i=2}^{n}[a(x,X_i) - c_n(x)]| \geq \frac{1}{2}(n-1)c_n(x)) \\ &\leq \exp(-\frac{(n-1)c_n(x)}{4(1-c_n(x)) + 2/3}) \\ &\leq \exp(-\frac{3(n-1)c_n(x)}{14}) \\ &\leq \exp(-\frac{nc_n(x)}{7}) \end{aligned} \tag{32}$$

Using the fact that $R_1 \leq 1$ along with Equation (32) we get

$$ER_1^2(x)I(A^c(x)) \leq P(A^c(x)) \leq \exp(-\frac{nc_n(x)}{7}) \tag{33}$$

Combining Equation (30) and Equation (33) gives

$$ER_1^2(x) \leq \frac{16}{n^2c_n^2(x)} + \exp(-\frac{nc_n(x)}{7}) \tag{34}$$

The conclusion follows by combining Equationms (28) and (34), together with the basic inequality which states that for all $x \geq 0$, $x^2e^{-x} \leq 1$. □

Lemma 2.6.

$$E(\hat{f}_{GNW}(x) - b_n(f,x))^2 \leq \frac{261B^2 + 65\sigma^2}{d_n(x)}$$

Proof. Using Equation (22), Lemma 2.5 and using the basic inequality $1-t \leq \exp(-t)$ valid for all $t \geq 0$, we get

$$\begin{aligned} E(\hat{f}_{GNW}(x) - b_n(f,x))^2 &= E(\hat{f}_{GNW}(x) - I(\sum_{i=1}^{n} a(x,X_i)b_n(f,x))^2 + b_n^2(f,x)P(\sum_{i=1}^{n} a(x,X_i) = 0) \\ &\leq (4B^2+\sigma^2)(\frac{16}{nc_n(x)} + nc_n(x)\exp(-\frac{nc_n(x)}{7})) + B^2(1-c_n(x))^n \\ &\leq (4B^2+\sigma^2)(\frac{16}{nc_n(x)} + nc_n(x)\exp(-\frac{nc_n(x)}{7})) + B^2\exp(-nc_n(x)) \end{aligned}$$

We conclude by using the basic inequalities: for all $t \geq 0$, $t^2e^{-t} \leq 1$ and $te^{-t} \leq 1$. □

Remark 2.7. *Suppose that $d_n(x) \to \infty$ as $n \to \infty$. Then Lemma (2.6) implies*

$$E(\hat{f}_{GNW}(x) - b_n(f,x))^2 \to 0$$

The growth of $d_n(x)$ can be arbitrarily slow, and the statement still holds. On the other hand if $d_n(x) \leq D$ for all $n \in \mathbb{N}$ and $\sigma^2 > 0$ then Lemma 2.4 gives

$$E(\hat{f}_{GNW}(x) - b_n(f,x))^2 \geq \frac{\sigma^2(1-e^{-D})}{D} > 0$$

3 Controlling the Bias at a point

3.1 Geometric concerns

Disconnected LPMs Since we want to consider the case where X is a random variable, we need to make sure that $b_n(f,X)$ is well defined. In a general Latent position model it could happen that with positive probability for all $n \in \mathbb{N}$, $c_n(X) = 0$, which means that with positive probability X will be an isolated node in the graph. To avoid such trivialities

Regularity

Kernel assumptions These assumptions on k_n in simple terms say that edge generation is dependent only on distances and thus make the model comparable to the Random Geometric Graph. It is an assumption so strong that it implies that $dp(x)$-almost surely $b_n(f,x) \to f(x)$ for a very general class of functions f. However, pointwise convergence of this form without uniform control of the error $|b_n(f,x) - f(x)|$ is not useful for our purposes. The following assumptions on f provide an easy way for bound on $\sup_{x \in \operatorname{supp} p} |b_n(f,x) - f(x)|$. Finally, in order to translate our results for fixed points $x \in \operatorname{supp} p$ into results for a random variable X with distribution p, we will need an assumption on the density p itself.

Density assumptions

- $p\colon \mathbb{R}^d \to \mathbb{R}$ is β-Holder continous on it's support $\operatorname{supp} p$ i.e. there exists $L_\beta > 0$ such that

$$\sup_{x,z \in \operatorname{supp} p} \frac{|p(x) - p(z)|}{||x-z||^\beta} \leq L_\beta$$

Lemma 3.1. *Suppose that **Kernel assumptions** hold. Then*

$$\operatorname{supp} p \subseteq \{x \in \mathbb{R}^d : c_n(x) > 0\}$$

Proof. Suppose that $c_n(x) = 0$. By continuity of K at 0, there is $r > 0$ such that for all z for which $||x - z|| \leq rh_n$, $K(\frac{x-z}{h_n}) \geq \frac{1}{2}$. Hence

$$\begin{aligned}\int I(||x-z|| \leq rh_n)p(z)dz &\leq 2\int I(||x-z|| \leq rh_n)K(\frac{x-z}{h_n})p(z)dz \\ &\leq 2\int K(\frac{x-z}{h_n})p(z)dz \\ &= 2c_n(x) \\ &= 0\end{aligned}$$

Hence $x \notin \operatorname{supp} p$, this proves the claim by contraposition.

□

Lemma 3.2. *Suppose that* ***Kerenl assumptions*** *and* ***Function assumptions*** *hold. Then*

$$\sup_{x \in \operatorname{supp}(p)} |b_n(f,x) - f(x)| \leq L_\alpha M^\alpha h_n^\alpha$$

Proof. For $x \in \operatorname{supp} p$ by Lemma 3.1, $c_n(x) > 0$. We have

$$\begin{aligned}|b_n(f,x) - f(x)| &= |\frac{\rho_n \int f(z)K(\frac{x-z}{h_n})p(z)dz}{\rho_n \int K(\frac{x-z}{h_n})p(z)dz} - f(x)| \\ &= |\frac{\int f(z)K(\frac{x-z}{h_n})p(z)dz}{\int K(\frac{x-z}{h_n})p(z)dz} - \frac{\int f(x)K(\frac{x-z}{h_n})p(z)dz}{\int K(\frac{x-z}{h_n})p(z)dz}| \\ &= |\frac{\int [f(z)-f(x)]K(\frac{x-z}{h_n})p(z)dz}{\int K(\frac{x-z}{h_n})p(z)dz}| \\ &\leq L_\alpha \frac{\int ||z-x||^\alpha K(\frac{x-z}{h_n})p(z)dz}{\int K(\frac{x-z}{h_n})p(z)dz} \\ &\leq L_\alpha M^\alpha h_n^\alpha\end{aligned}$$

Here, we used the fact that for any function $G \in L^1(dp(x))$, $\int_{\mathbb{R}^d} G(z)p(z)dz = \int_{\operatorname{supp} p} G(z)p(z)dz$ and crucially the facts that f is α-Holder continuous and that K is compactly supported in the last inequality. □

Corollary 3.3. *Suppose that* ***Kernel assumptions*** *and* ***Function assumptions*** *hold. Let X be a random variable with density p.*

$$P(|b_n(f,X) - f(X)| \leq L_\alpha M^\alpha h_n^\alpha) = 1$$

Proof. We have

$$\begin{aligned}P(|b_n(f,X) - f(X)| \leq L_\alpha M^\alpha h_n^\alpha) = &\int I(|b_n(f,x) - f(x)| \leq L_\alpha M^\alpha h_n^\alpha)p(x)dx \\ &\int_{\operatorname{supp} p} I(|b_n(f,x) - f(x)| \leq L_\alpha M^\alpha h_n^\alpha)p(x)dx \\ &= 1\end{aligned}$$

where we used Lemma 3.2 in the last line.

□

Lemma 3.4 shows that under **Density assumptions** the .

Lemma 3.4.

4 Risk convergence for a fixed point

4.1 Main Results

Theorem 4.1. *Assume that* ***Kernel assumptions*** *and* ***Function assumptions*** *hold.*
If $mh_n \leq d(x, \partial \operatorname{supp} p)$

$$E(\hat{f}_{GNW}(x) - f(x))^2 \leq 2$$

4.2 Discussion

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.

Remark 10 (Non compact case) 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.

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 $[0, 1]^d$ 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.

5.1 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 3 we plot these errors against the logarithm of the sample size. The top left and bottom left images on Figure 3 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.

5.2 Estimating functions

$\mathbb{P}\ \mathbb{E}\ \supset$

6 Appendix

6.1 related work/future plans

Clustering algorithms on Stochastic block models
Oliviera
Lei'2015
Levina-Vershynin
Latent position model
Bickel
Tang
Arias-Castro
Chatterjee

6.2 Probabilistic proof of the Bernoulli inequality

The Bernoulli inequality[4] states that for all $0 < p < 1$ and $n \in \mathbb{N}$,

$$(1-p)^n \geq 1 - np \tag{35}$$

More generally, for $0 < p < 1$ and $n \in \mathbb{N}$ we are going to derive bounds for

$$b_j = (-1)^j[(1-p)^n - \sum_{l=0}^{j-1} \binom{n}{l}(-1)^l p^l] \tag{36}$$

[4]In fact the Bernoulli inequality states that for all $y > -1$ and $n \in \mathbb{N}$, $(1+y)^n \geq 1 + ny$ but we are only interested in the case $-1 < y < 0$

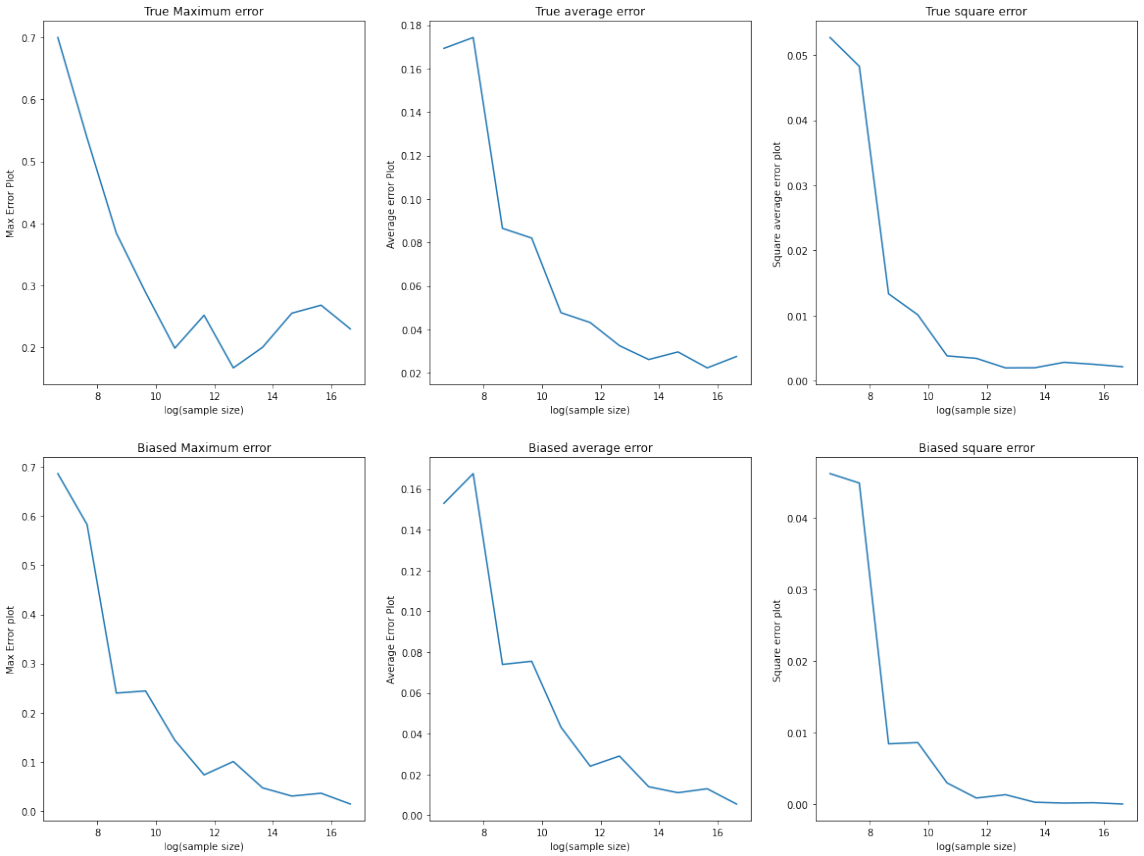

Figure 3: Error decay with sample size. The kernel is Gaussian $k(x, y) = \exp(-(\frac{x-y}{h})^2)$. The bandwith is set to $h = 0.11$, the noise is set to $\sigma^2 = 1$, and the function to be estimated is $f(x) = 2 + x - e^{-x^2}$. The experiment is repeated 20 times and the average error curves are plotted.

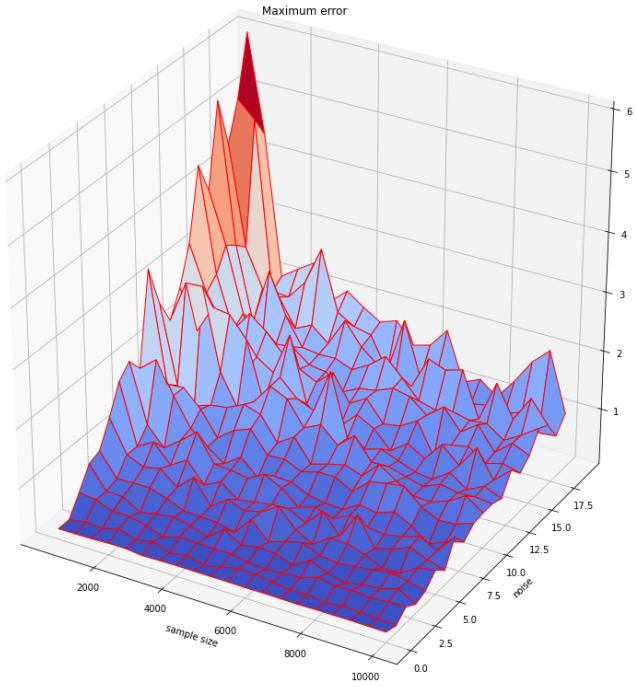

Figure 4: True maximum error plotted against sample size and noise level. The parameters are as in Figure 3.

The proof of Bernoulli's inequality follows easily from the observation that $R_1 \leq 1$. Indeed, by Lemma 2.1

$$\frac{1-(1-c(x))^n}{nc(x)} = ER_1 \leq 1$$

Setting $p = c(x)$ one easily derives inequality (35). The following lemma provides bounds on (36).

Lemma 6.1. *For* $j = 1, 2, ..., n-1$

$$ER_{[j+1]} = \frac{1 - jER_{[j]}}{(n-j)c(x)}$$

Proof. Using Equation (15) for $l = j+1, ...n$, we have

$$a(x, X_l)R_{[j]} = a(x, X_l)R_{[j]\cup\{l\}}$$

Summing these equations for $l = j+1, ..., n$ we get

$$\begin{aligned} 1 - jR_{[j]} &= 1 - \frac{j}{j + \sum_{l=j+1}^{n} a(x, X_l)} \\ &= \frac{\sum_{l=j+1}^{n} a(x, X_l)}{j + \sum_{l=j+1} a(x, X_l)} \\ &= \sum_{l=j+1}^{n} a(x, X_l)R_{[j]\cup\{l\}} \end{aligned}$$

Taking expectation and using the fact that $a(x, X_l)$ and $R_{[j]\cup\{l\}}$ are independent, and the fact that $R_{[j]\cup\{l\}}$ and $R_{[j+1]}$ are identically distributed, we get

$$
\begin{aligned}
1 - jER_{[j]} &= \sum_{l=j+1}^{n} Ea(x, X_l) R_{[j]\cup\{l\}} \\
&= \sum_{l=j+1}^{n} Ea(x, X_l) ER_{[j]\cup\{l\}} \\
&= (n-j)c(x) ER_{[j+1]}
\end{aligned}
$$

□

Corollary 6.2. *For $j = 1, 2, 3, ..., n-1$ let b_j be given by (36). Then*

$$
\frac{j}{n}\binom{n}{j}p^j < b_j < \binom{n}{j}p^j
$$

Proof. Consider the sequence $b_j^{'} = \frac{b_j}{j\binom{n}{j}p^j}$. It is easy to verify that $b_1^{'} = ER_1$ and that

$$
b_{j+1}^{'} = \frac{1}{1}
$$

□

6.3 Alternative approach to concentration properties

6.3.1 Motivation and main ideas

Motivation Given $x \in \mathbb{R}^d$, as soon as $c(x) > 0$, the strong law of large numbers states that

$$
\hat{f}_{GNW}(x) \to b(f, x) \text{ almost surely} \tag{37}
$$

Although statement (37) is good as a heuristic, it is asymptotic in nature and hence of limited importance for theoretical guarantees such as sample complexities. We use concentration inequalities to specify a rate at which this convergence occurs. This section contains two results, one about concentration properties of $\hat{f}_{GNW}(x)$ with $x \in \mathbb{R}^d$ fixed and such that $c(x) > 0$ and one about concentration properties of $\hat{f}_{GNW}(X)$ where $X, X_1, ..., X_n$ are i.i.d. random variables with density p. We make the following assumptions throughout this section:

- $f \colon \mathbb{R}^d \to \mathbb{R}$ is bounded and measurable with $||f||_\infty \leq B$
- $x \in \mathbb{R}^d$ is a point with $c(x) = \int k(x, z)p(z)dz > 0$
- $X, X_1, ..., X_n$ are i.i.d. random variables with density p

Concentration for a deterministic point The Graphical Nadaraya Watson estimator $\hat{f}_{GNW}(x)$ concentrates towards $b(f, x)$ for all bounded functions $f \colon \mathbb{R}^d \to \mathbb{R}$. 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 precise statement of this result is Theorem 6.5. The main idea in the proof is to show separately concentration for the statistics $\frac{1}{n}\sum_{i=1}^n Y_i a(x, X_i)$ and $\frac{1}{n}\sum_{i=1}^n a(x, X_i)$ towards $T_k(f)(x)$ and $c(x) = T_k(f)(x)$ respectively. The first step towards Theorem 6.5 is to show that $\frac{1}{n}\sum_{i=1}^n f(X_i)a(x, X_i)$ concentrates towards $\int f(z)k(x, z)p(z)dz$ using McDiarmid's concentration inequality. This is done in Lemma 6.3. Next, we need show that the noise term $\frac{1}{n}\sum_{i=1}^n \epsilon_i a(x, X_i)$ concentrates towards 0. This is done in Lemma 6.4 using Subgaussian concentration inequalities. Finally, using Lemma 6.3 and Lemma 6.4 we prove Theorem 6.5.

Concentration result for a random point The situation is more involved in the case where X is a random point. The first complications arise out of the fact that the variables $a(X, X_i)$ and $a(X, X_j)$ are no longer independent for $i \neq j$. The concentration result in this case can be stated informally as $\hat{f}_{GNW}(X)$ will concentrate towards $b(f, X)$ with a rate that is exponential in n provided that the probability that $c(X)$ takes small values is small. The precise statement of this result is Theorem 6.8. The first step towards the proof of Theorem 6.8 is to derive concentration statements about $\frac{1}{n}\sum_{i=1}^{n} Y_i a(X, X_i)$ and $\frac{1}{n}\sum_{i=1}^{n} a(X, X_i)$ towards $T_k(f)(X)$ and $c(X)$ respectively. This is done in Corollary 6.6 and Corollary 6.7, which can be considered as the random point analogues of Lemma 6.3 and Lemma 6.4 respectively. Informally speaking, one can think of Corollary 6.3 as integrating the inequality in Lemma 6.3 with respect to the density p.

6.4 Concentration for a deterministic point

Lemma 6.3. *Suppose that f is bounded, measurable function with $||f||_\infty \leq B$. Then*

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

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** 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 follows immediately from McDiarmid's inequality. □

In the following corollary we prove a concentration result for a variable X which is independent and identically distributed with $X_1, ..., X_n$.

Lemma 6.4. *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*

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

where $c > 0$ is an absolute constant.

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

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

We have

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

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

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

In particular this shows that

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

Using the General Hoeffding's inequality (**vershynin** Thm 2.6.3), we have

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

with $c > 0$ an absolute constant. □

Theorem 6.5. *Suppose that* $||f||_\infty \le B$ *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*

$$P(|\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{\int k(x,z)p(z)dz}| \ge \delta) \le 6\exp(-H(B,\sigma^2)c(x)^2\delta^2 n)$$

Proof. Let $\delta > 0$ and denote

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

Let $\delta_1, \delta_2, \delta_3 > 0$, we will specify them later. Choosing $\delta_2 \le \frac{1}{2}c(x)$, on $B_{\delta_2}^c$ we have $\frac{1}{n}\sum_{i=1}^n a(x,X_i) \ge \frac{1}{2}c(x)$ and in particular $\sum_{i=1}^n a(x,X_i) > 0$. Hence on $B_{\delta_2}^c$, we have

$$\begin{aligned}\hat{f}_{GNW}(x) - \frac{\int f(z)k(x,z)p(z)dz}{c(x)} &= \frac{\frac{1}{n}\sum_{i=1}^n Y_i a(x,X_i)}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} - \frac{\int f(z)k(x,z)p(z)dz}{c(x)} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n [f(X_i) - \frac{T_k(f)(x)}{c(x)}]a(x,X_i)}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)} + \frac{\frac{1}{n}\sum_{i=1}\epsilon_i a(x,X_i)}{\frac{1}{n}\sum_{i=1}^n a(x,X_i)}\end{aligned} \tag{38}$$

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

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

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 6.3, 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 6.4 we have $P(C_{\delta_3}) \leq 2\exp(-\frac{C\delta_3^2 n}{\sigma^2})$ where $C > 0$ is a constant.

The result follows from a union bound

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

□

6.5 Concentration for a random point

Corollary 6.6. *Suppose that f is bounded, measurable function with $||f||_\infty \leq B$ and that $X, X_1, ..., X_n$ are i.i.d. random variables with density p. Then*

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

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 6.3, we have

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

Finally, using Fubini's theorem, we have

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

□

Corollary 6.7. *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*

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

Proof. The result follows by Lemma 6.4 using the same method that was used to derive Corollary 6.6 from Lemma 6.3. □

Theorem 6.8. *Suppose that $||f||_\infty \leq B$ and $X, X_1, ..., X_n$ are i.i.d. random variables with density p. Set $H(B, \sigma^2) = \min(\frac{c_1}{\sigma^2}, \frac{1}{90B^2})$. Then for all $r > 0$ and $0 < \delta < 3B$ we have*

$$
P(c(X) > 0, |\hat{f}_{GNW}(X) - \frac{T_k(f)(X)}{c(X)}| \geq \delta) \leq 6\exp(-H(B, \sigma^2)r^2\delta^2 n) + P(c(X) < r)
$$

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

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

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 6.5 gives

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

Choosing $\delta_1 = \delta_3 = \frac{r\delta}{6}$ and $\delta_2 = \min(\frac{r}{2}, \frac{r\delta}{6B})$, on $C_r \cap A_{\delta_1}(f)^c \cap A_{\delta_2}(1)^c \cap N_\delta^c$ we have $c(X) > 0$ and

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

To conclude, when $\delta < 3B$, using Corollary 6.6 and Corollary 6.7 we have

$$P(c(X) > 0, |\hat{f}_{GNW}(X) - \frac{T_k(f)(X)}{c(X)}| \geq \delta) \leq P(C_r^c \cup A_{\delta_1}(f) \cup A_{\delta_2}(1) \cup N_\delta)$$
$$\leq P(c(X) < r) + 2\exp(-\frac{r^2\delta^2 n}{90B^2}) + 2\exp(-\frac{r^2\delta^2 n}{90B^2}) + 2\exp(-\frac{c_1 r^2\delta^2 n}{\sigma^2})$$
$$\leq P(c(X) < r) + 6\exp(-H(B,\sigma^2)r^2\delta^2 n)$$

□

6.6 Remarks

Remark 6.9. *The quantity $H(B,\sigma^2)$ is inversly proportional to the boundedness and noise parameters B and σ^2 respectively. In particular if either of the two parameters B and σ^2 increases while the other is fixed, the quantity $H(B,\sigma^2) = \min(\frac{C}{\sigma^2}, \frac{1}{90B^2})$ decreases and hence the concentration rate in Theorem 6.5 and Theorem 6.8 decreases.*

Remark 6.10. *The proof of Lemma 6.4 relies on sub-gaussian inequalities. These inequalities hold 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 6.11. *As long as $E|f(X_1)k(x,X_1)| = \int |f(z)|k(x,z)p(z)dz < \infty$ and $c(x) = \int k(x,z)p(z)dz > 0$, the strong law of large numbers states that*

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

This is the case if $E|f(X_1)| = \int f(z)p(z)dz < \infty$. 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 $t > 0$ such that

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

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 6.12. *A slight modification of the presented proofs shows that the classical Nadaraya Watson estimator $\hat{f}_{NW}$ given by 1 satisfies*

$$P(|\hat{f}_{GNW}(x) - \hat{f}_{NW}(x)| \geq \delta) \leq c_1\exp(-H_1(B,\sigma^2)c(x)^2\delta^2 n)$$

for some absolute constants $c_1 > 0$ and $H_1(B,\sigma^2) > 0$. Indeed, if we take

$$F_1(x_1, ...x_n) = \frac{1}{n}\sum_{i=1}^{n}[f(x_i)k(x,x_i) - \int f(z)k(x,z)p(z)dz]$$

then $EF_1(X_1, ..., X_n) = 0$ and similar ideas as in Lemma 6.3 apply. We omit the details. Note that the ambient space in which the data $X_1, ..., X_n$ is embedded does not a play a role in this section as long as the variables are independent and $||f||_\infty \leq B$. 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 as we show in the next section.

Remark 6.13. *Assuming that $\inf_{x \in \text{supp}\, p} c(x) \geq r > 0$ gives $P(\int k(X,z)p(z)dz < r) = 0$ so that by Theorem 6.8, $\hat{f}_{GNW}(X)$ concentrates around $\frac{\int f(z)k(X,z)p(z)dz}{\int k(X,z)p(z)dz}$ with exponentially small probability in the sample size n. An application of Borel-Cantelli's lemma in such a case gives that $\hat{f}_{GNW}(X) \to \frac{T_k(f)(X)}{T_k(1)(X)}$ almost surely. This is the case 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 $\frac{\int f(z)k(X,z)p(z)dz}{c(X)}$.*

6.7 A Generalization: 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

6.7.1 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[5]

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

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:

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

$$\begin{aligned}\frac{1}{n(n-1)}\sum_{i=1}^{n} f(X_i)w_2(X_i,X) &= \frac{1}{n(n-1)}\sum_{j=1}^{n}[\sum_{i\neq j} f(X_i)a(X_i,X_j)]a(X_j,X)\\ &= \frac{1}{n}\sum_{j=1}^{n}[\frac{1}{n-1}\sum_{i\neq j} f(X_i)a(X_i,X_j) - \int f(z)k(X_j,z)p(z)dz]a(X_j,X)\\ &+ \frac{1}{n}\sum_{j=1}^{n}[\int f(z)k(X_j,z)p(z)dz]a(X_j,X)\\ &= \frac{1}{n}\sum_{j=1}^{n}[S_j - T_k(f)(X_j)]a(X_j,X) + \frac{1}{n}\sum_{j=1}^{n} T_k(X_j)a(X_j,X)\end{aligned} \tag{39}$$

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

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

Hence, by a union bound we have

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

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

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

Finally, combining the last two displays together with (39), we have

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

□

Theorem 4 For any $r > 0$,

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

Proof. Denote

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

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

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

and

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

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[6]

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

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 $r = \inf_{x \in \text{supp}\, p} c_2(x) > 0$ 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

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

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.

7 maybe useful?

As pdx is a probability measure, compositions of T_k of any order $m \geq 1$ are well defined, and

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

remarks? 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 $n \geq n(x)$,

$$|\frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)} - f(x)| \leq ch_n^\alpha$$

with $c > 0$ an absolute constant depending on k and the Holder constants of f and p.

remarks..

- Suppose now that k is compactly supported and let $M > 0$ be a constant such that $k(x) = 0$ for $|x| > M$. Then, from (**??**) we get

$$|\frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)} - f(x)| \leq LM^{\alpha} h_n^{\alpha}$$

  This bound is independent of $x \in \operatorname{supp} p$ 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

$$\begin{aligned} |\int G(y)k(y)p(x+h_n y)dy - p(x)\int G(y)k(y)dy| &= |\int G(y)k(y)[p(x+h_n y) - p(x)]dy| \\ &\leq h_n^{\beta}\int ||y||^{\beta}|G(y)|k(y)dy \end{aligned} \tag{40}$$

  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 (**??**), we get

$$|\frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)} - f(x)| \leq ch_n^{\alpha}$$

  where $c > 0$ is an absolute constant depending on k and the Holder constants of f and p.

suboptimal stuff

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

$$\lim_{n\to\infty} c_n(x)^2 n = \infty \tag{41}$$

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

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

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),

$$c_1 p(x)\lambda_n h_n^d \leq c_n(x) \leq c_2 p(x)\lambda_n h_n^d \tag{42}$$

- Suppose that $\lambda_n h_n^d = \omega(\frac{1}{\sqrt{n}})$. Then using Theorem 1 and 42 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).

□