Graphical Nadaraya Watson estimator

Martin Gjorgjevski

ENS Lyon
M2 Advanced Mathematics

July 2022

Introduction: Nonparametric regression and Integral operators

The classical nonparametric regression framework:

$$Y_i = f(X_i) + \epsilon_i \ , \ i = 1, 2, ..., n$$

where

- $X_1, ..., X_n \in \mathbb{R}^d$ i.i.d. variables called features or covariates with density p
- $Y_1, ..., Y_n \in \mathbb{R}$ are called response variables
- $\epsilon_1, ..., \epsilon_n$ Gaussian noise independent of $X_1, ..., X_n$.

The goal is to estimate the regression function $f(x) = E(Y|X = x)$.

NW estimator

One popular estimator for f in this context is the Nadaraya Watson estimator given by

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

Here $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is an arbitrary symmetric function (i.e. $k(x, z) = k(z, x)$) also known also called kernel

We also assume that

- k is stationary kernel, i.e. $k(x, y) = k(x - y)$
- k is bounded kernel, more precisely $0 \leq k(x, z) \leq 1$

With such a kernel one can associate an integral operator $T_k : L^1(\mathbb{R}^d, p) \to L^\infty(\mathbb{R}^d, p)$ with

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

Introduction: LPMs

Latent Postion Model

Given n, a kernel k on $\mathbb{R}^d$ 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:

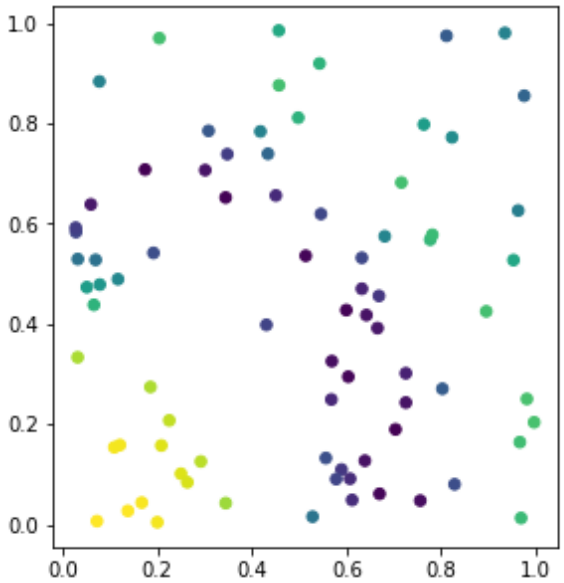
- For each vertex i, $1 \leq i \leq n$, a sample X_i is drawn with distribution p. This variable is known as the latent position of node i
- For each pair (i, j) with $i < j$,

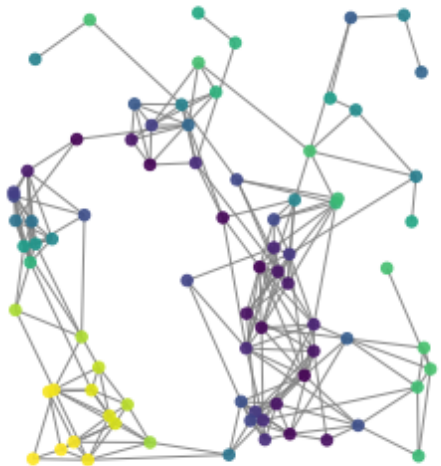
$$a(i, j) = a(X_i, X_j) = I(U_{i,j} \leq k(X_i, X_j))$$

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

Here,

- The samples $X_1, ..., X_n$ are not observed and are assumed to be independent
- The variables $U_{i,j}$, $1 \leq i < j \leq n$ are uniformly distributed on $[0, 1]$, independent among themselves and from X_i, $1 \leq i \leq n$.





The Basic GNW estimator

Given a LPM graph on $n+1$ vertices with **latent** positions $X_1, ..., X_n, X$ and additional data $Y_1, ..., Y_n$ on vertices $1, ..., n$, with $Y_i = f(X_i) + \epsilon_i$ we are interested in estimating the regression function $f(x)$ and the data at node X, $f(X)$. Inspired by the classical Nadaraya Watson estimator, we introduce

Graphical Nadaraya Watson estimator

The Graphical Nadaraya Watson estimator is defined as

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

Concentration results-Variance bounds

By the SLLN we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i a(x, X_i) = \int f(z)k(x, z)p(z)dz = T_k(f)(x)$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a(x, X_i) = \int k(x, z)p(z)dz = T_k(1)(x) = c(x, k)$$

Hence, we have that $\hat{f}_{GNW}(x) \to \frac{T_k(f)(x)}{T_k(1)(x)}$ a.s. In fact we can give a nonasymptotic quantification of this convergence

Theorem 1

Suppose that $||f(X_1)||_\infty \leq B$ and $c(x, k) > 0$. Then for $0 < \delta < 3B$ and $H(B, \sigma^2) = \min\{\frac{1}{90B^2}, \frac{C}{\sigma^2}\}$ we have

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

Concentration results - Variance bounds

We see that $c(x,k)$ plays an import role in the control of variance. Heuristically, $c(x,k)$ is the average degree at node associated with latent position x. Hence if $c(x,k) \geq r > 0$ we can prove a concentration result for $\hat{f}_{GNW}(X)$ where X is random variable i.i.d. with the latent data $X_1, ..., X_n$. More precisely we have the following result:

Theorem 2

Suppose that $X, X_1, ..., X_n$ are i.i.d. with density p and[a] c(x,k)>0 for all $x \in \text{supp}(p)$. Then for any $r > 0$, $0 < \delta < 3B$ and $H(B,\sigma^2) = \min(\frac{c_1}{\sigma^2}, \frac{1}{90B^2})$ we have

$$P(|\hat{f}_{GNW}(X) - \frac{T_k(f)(X)}{T_k(1)(X)}| \geq \delta) \leq 6e^{-H(B,\sigma^2)r^2\delta^2 n} + P(c(X,k) < r)$$

[a] for the sake of simplicity

Bias

In the random graph and nonparametric statistics literature it is common to assume that the kernel varies with sample size. We assume that

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

where

- $k : \mathbb{R}^d \to [0, M]$ is bounded and nonnegative
- $0 < \lambda_n \leq \frac{1}{M}$ for all $n \in \mathbb{N}$
- $h_n > 0$, for all $n \in \mathbb{N}$ and $h_n \to 0$

Contrary to the classical NW estimator, in our framework we do not get to choose the kernel. Hence we do not have control of the quantity $|\frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)} - f(x)|$. Despite this we can give conditions on k_n under which $\hat{f}_{GNW}$ is a consistent estimator

Lemma

Suppose that $0 \in \text{supp}(k)$ and f is α-Holder continuous with $0 < \alpha \leq 1$. Then there exist constants $C, c > 0$ such that

- If k has compact support and X is a random variable with distribution p, then

$$||\frac{T_{k_n}(f)(X)}{T_{k_n}(1)(X)} - f(X)||_\infty = \sup_{x \in \text{supp}(p)} |\frac{T_{k_n}(f)(x)}{T_{k_n}(1)(x)} - f(x)| \leq Ch_n^\alpha$$

- 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 $n \geq n(x)$,

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

Weak and strong consistency

Informally speaking under regularity assumptions such as Holder continuity on f and p, it can be shown that

$$c(x, k_n) \approx \lambda_n h_n^d p(x)$$

Theorem

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

$$|f(x) - f(z)| \leq L||x - z||^{\alpha} \text{ and } |p(x) - p(z)| \leq L||x - z||^{\beta}$$

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

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

Remarks

Clearly shrinking λ_n forces h_n to increase so as to preserve the rate $\omega(\frac{1}{\sqrt{n}})$ and hence the bound for the bias term is increasing. Thus in some sense the optimal setting is $\lambda_n = \lambda > 0$.
The GNW estimator has similar properties to the common estimators in classical and high dimensional statistics:

- **Bias-variance tradeoff** Small h_n mean small bias and large variance and vice versa, large h_n means large bias and small variance. In theory good tradeoff is possible when $h_n = \omega(\frac{1}{n^{1/2d}})$ and $h_n = o(1)$
- **Curse of dimensionality** It can be shown that the sample complexity $n(\epsilon, \delta) = c\epsilon^{-2\frac{d+\alpha}{\alpha}} \log(\frac{1}{\delta})$. Hence the sample complexity grows exponentially with the dimension of the latent data $X_1, ..., X_n$

Denote $E_*(Z) = E(ZI(\sum_{i=1}^{n} a(x, X_i) > 0))$.

Theorem

Suppose that $||f(X_1)||_\infty \leq B$ and $c(x) = c(x,k) > 0$. Then

$$E_*(\hat{f}_{GNW}(x) - \frac{T_k(f)(x)}{T_k(1)(x)})^2 \leq \frac{c_1 B^2}{nc^2(x)} + c_2 n^2 e^{-c_3 c^2(x) n} + \frac{2\sigma^2}{nc(x)}(1 + (nc(x)+1)e^{-\frac{nc^2(x)}{2}})$$

In particular, if $\lambda_n h_n^d = \omega(\sqrt{\frac{\log(n)}{n}})$, then the mean squared error goes to zero.

Generalizations

The proposed estimator $\hat{f}_{GNW}$ does not take advantage of the graph structure of the data. In order to account for the potential influence of vertices which are not direct neighbours of v, 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, $1 \leq m \leq n$ and $J_i = (i, i_1, ..., i_{m-1})$ 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**:

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

Generalizations

Definition

For $m \geq 1$ we define $T_k^m : L^1(\mathbb{R}^d, p) \to L^\infty(\mathbb{R}^d, p)$ with

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

for $f \in L^1(\mathbb{R}^d, p)$.

The function $T_k^m(1)$ is important for concentration of $\hat{f}_{GNW,m}$:

Theorem

Suppose $X_1, ..., X_n, X$ are i.i.d. with distribution p. There exist absolute constants $c_1, c_2 > 0$ such that

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

Plans/ To Do...

- $w_m(X_i, X)$ is a computationally challenging statistic
- Instead one can consider the powers of the adjacency matrix $A = [a(X_i, X_j)]_{i,j} \in \mathbb{R}^{(n+1)\times(n+1)}$, this is achieved by replacing $w_m(X_i, X_j)$ with $w_m^{'}(X_i, X_j) = [A^m]_{i,j}$
- $w_m^{'}$ is a lot faster to compute, but probably worse concentration rate compared to w_m

We propose the following estimators:

- Linear regression with design $G = [\hat{f}_{GNW,j}(X_i)]_{i,j} \in \mathbb{R}^{n\times m}$
- $\hat{\alpha_{LR}} = \text{argmin}_{\alpha\in\mathbb{R}^m} ||Y - G\alpha||^2$ or equivalently $\hat{\alpha_{LR}} = \text{argmin}_{\alpha\in\mathbb{R}^m} \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{m} \alpha_j \hat{f}_{GNW,m}(X_i))^2$
- Explicit solution $\hat{\alpha} = (G^T G)^{-1} G^T Y$, set $\hat{\alpha}_j = [\hat{\alpha}_{LR}]_j$
- $\hat{f}_{LR}(X) = \sum_{j=1}^{m} \hat{\alpha}_j \hat{f}_{GNW,j}(X)$

Plans/ To do...

- For $\alpha \in \mathbb{R}_+^m$ define $w(\alpha)(X_i, X) = \sum_{j=1}^m \alpha_j w_j(X_i, X)$
- $\hat{\alpha} = \text{argmin}_{\alpha \in \mathbb{R}^m} \sum_{i=1}^n (Y_i - \frac{\sum_{j \neq i} Y_j w(\alpha)(X_j, X_i)}{\sum_{j \neq i} w(\alpha)(X_j, X_i)})^2$
- $\hat{f}(X) = \frac{\sum_{i=1}^n Y_i w(\alpha)(X_i, X)}{\sum_{i=1}^n w(\alpha)(X_i, X)}$
- No explicit solution, estimate for $\hat{\alpha}$ should be obtained through gradient descent methods

Simulations

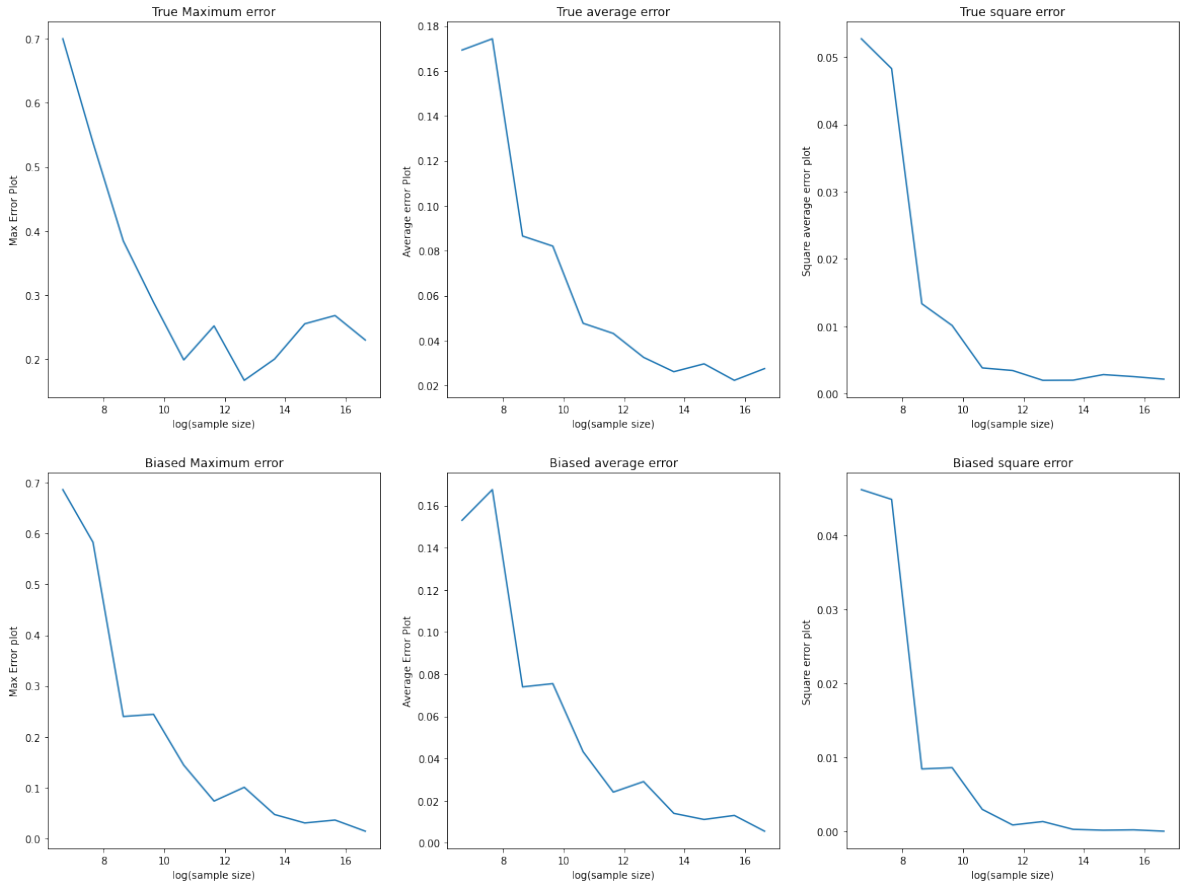
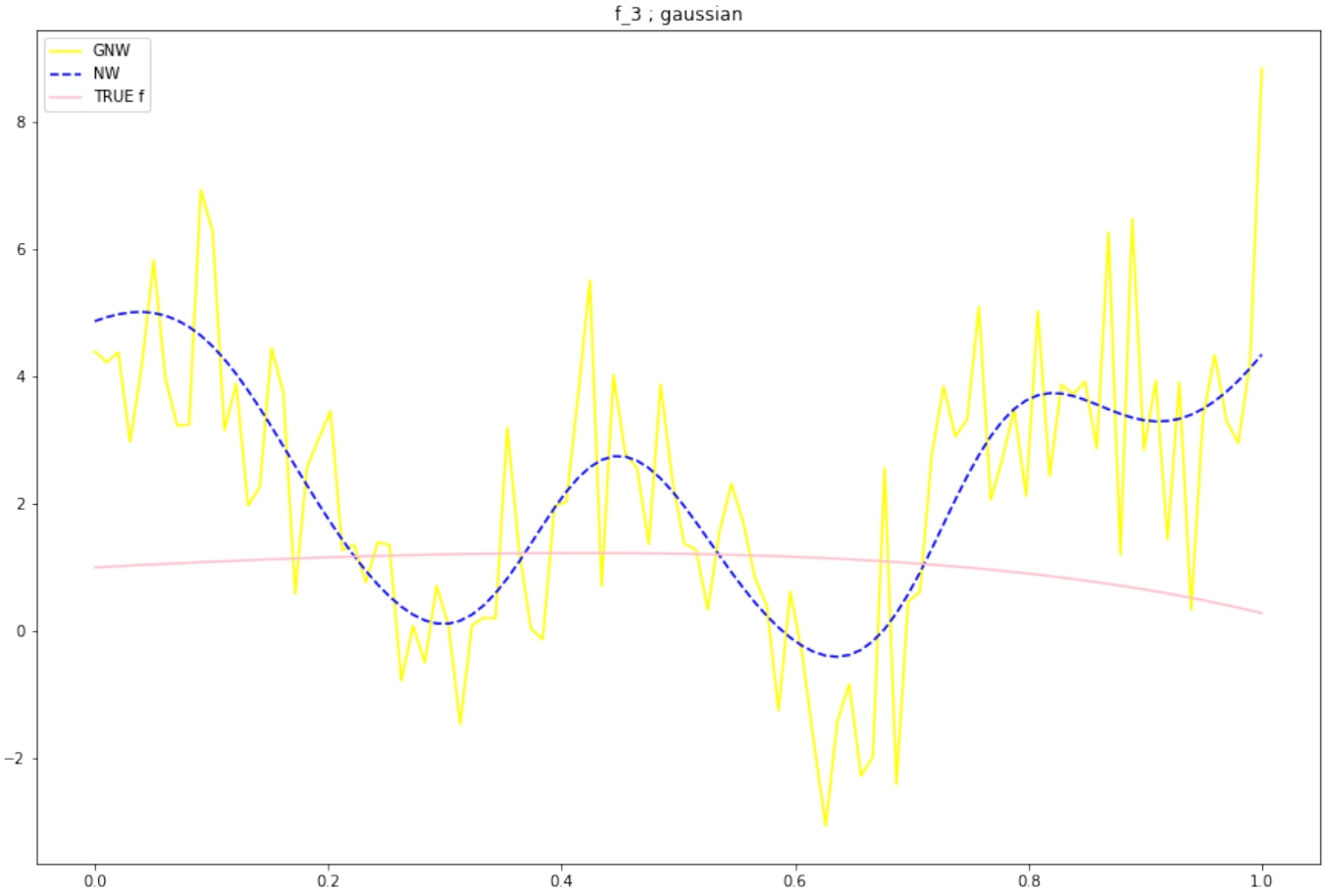
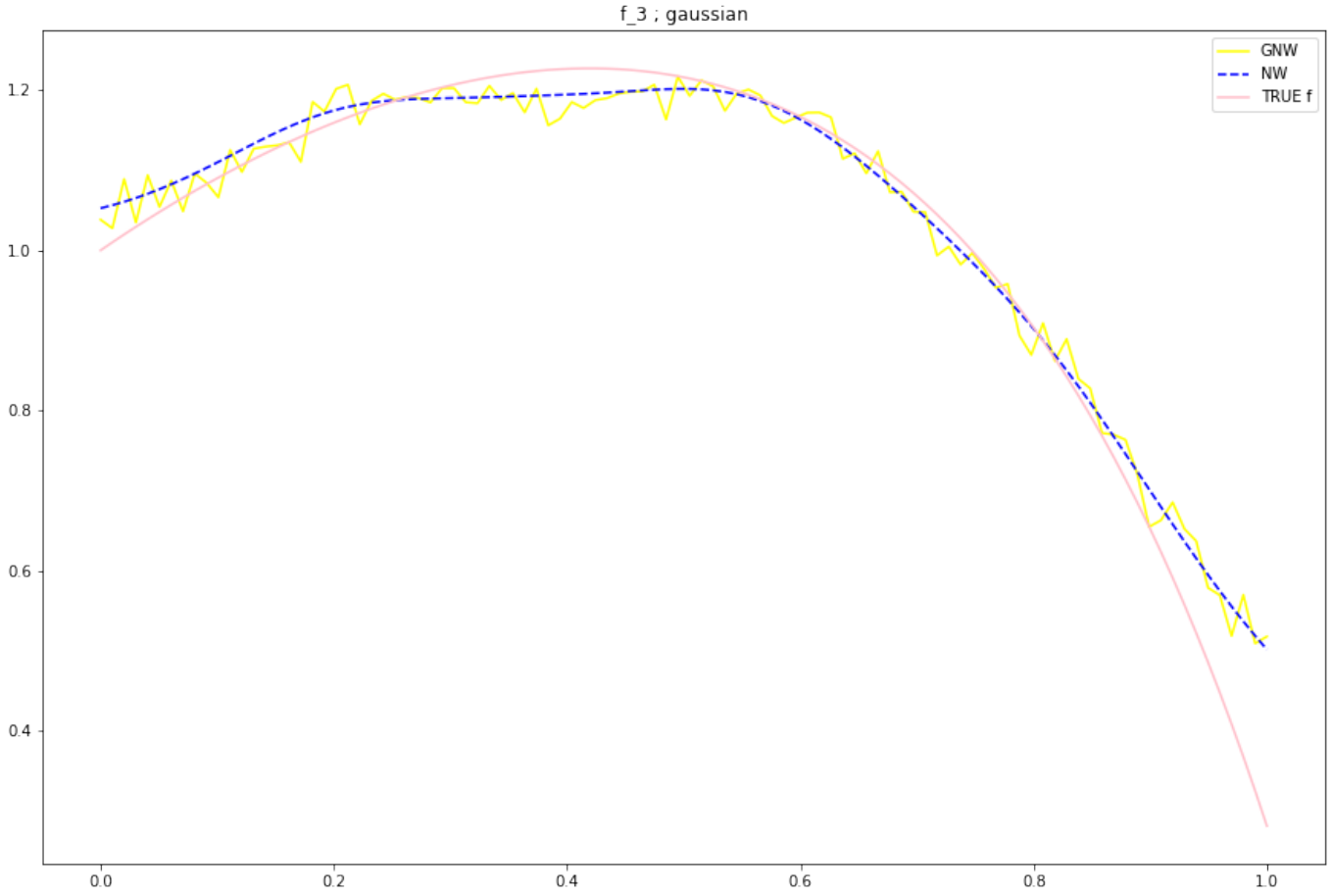


Figure: Error decay with sample size, Gaussian kernel $k(x, y) = \exp(-(\frac{x-y}{h})^2)$,bandwith $h = 0.11$, noise $\sigma^2 = 1$, and $f(x) = 2 + x - e^{-x^2}$. The experiment is repeated 20 times and the average error curves are plotted.

Simulations



Simulations



Simulations