Adaptive Design for Locating the Maxima of a Regression Function : A Nonparametric Approach[1].

Partha Sarathi Dey
Indian Statistical Institute

March 17, 2007

[1] This dissertation was done under the supervision of Prof. Probal Chaudhury, ISI, Kolkata

Abstract

Consider a nonparametric regression setup where the regression function $M(x) = E(Y|X = x)$ is an unknown smooth function. In this setup several methods are known for the classic problem of locating the maxima of the regression function. However, almost all of these methods work only if the regression function has a unique local maximum. How to design an experiment so that we can solve the above problem when the regression function has more than one global maxima and possibly several local maxima is a question that has not been addressed adequately in the available literature. Here we propose two adaptive sequential methods for the above problem, which are applicable to regression functions of several variables. We derive results on the asymptotic behavior of the proposed procedures and report some simulation results that demonstrate their finite sample behavior.

Contents

Chapter 1

Introduction

1.1 Problem description

Let us consider the model

$$Y = M(X) + \sigma(x)\varepsilon, \tag{1.1}$$

where $E(\varepsilon) = 0, Var(\varepsilon) = 1$, $M(x)$ is a smooth function from $\mathcal{X}$ to $\mathbb{R}$, and $\sigma(x)$ is a continuous function on $\mathcal{X}$ bounded away from zero. Here, we consider a nonparametric setup, where no parametric form of $M(x)$ has been assumed.

Let us assume that $\mathcal{X}$ is a connected and compact subset of $\mathbb{R}^d$ for some $d \geq 1$, and $M(x)$ has only finitely many global maxima. Define

$$m = \sup_{x \in \mathcal{X}} M(x) \tag{1.2}$$

$$\text{and } \mathcal{X}_{opt} = \{x \in \mathcal{X} : M(x) = m\}. \tag{1.3}$$

Now, to estimate $M(x)$ over $\mathcal{X}$, in some parametric cases, we can find optimal fixed designs (D-optimal, A-optimal etc.) which minimizes appropriate criteria functions (see *e.g.,* Fedorov [9]). But if our objective is to find a fixed design, that is optimal with respect to the *minimum integrated mean square error* criteria in nonparametric setup, the optimal design turns out to be uniform design. In the case of kernel regression, Müller [18] proved that the optimal design w.r.t. the minimum IMSE criteria is the design having density $\sigma(x)/\int_{t \in \mathcal{X}} \sigma(t)\, dt$. In particular, uniform design is the best design in the homoscedastic case w.r.t. minimum IMSE criterion. This result is not unexpected, since the regression function belongs to a large class of functions and to estimate it “nicely” at all places no design can beat the uniform design. Moreover, if our objective is to estimate $M(x)$ efficiently at one point, clearly the design degenerate at that point would be the best design in that case.

However, sometimes it is more important to identify and locate some features of the regression function such as local or global maxima or minima of the function efficiently than estimating the whole regression function. For example, in a production precess, there are several factors influencing the production, and one important problem there is to find levels of explanatory variables, like temperature, pressure, concentration of chemicals etc., under which the output is maximum. Adaptive design, where the design for each trial depends on the data from previous trials, is naturally more suitable in these cases. In addition to precisely locating the global maxima of the unknown regression function, we want to keep the output to a high level throughout the experiment(*i.e.*, less number of "unproductive" trials). Moreover if the regression function has more than one maximum, full knowledge of the set of maxima can decrease the production cost of the process significantly while achieving the best level of the production. So, we consider criteria which reflect these ideas. Here, we consider the case of finding maxima. We also assume that the regression function $M(.)$ is nonnegative and $m = \sup_{x \in \mathcal{X}} M(x) > 0$. Suppose that our total resource (total sample size that can be used in the experiment) is n and the observations are $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$. Our criterion functions are

$$C_1 = \frac{1}{m}\left(m - \min(\bar{y}, m)\right) = 1 - \min\left(\frac{\sum_{i=1}^n y_i}{nm}, 1\right), \tag{1.4}$$

$$C_2 = \frac{1}{nm}\sum_{i=1}^n \Big(m - \min(y_i, m)\Big) = 1 - \frac{1}{n}\sum_{i=1}^n \min\left(\frac{y_i}{m}, 1\right) \tag{1.5}$$

$$\text{and } C_3 = \frac{1}{nm}\sum_{i=1}^n \Big(m - M(x_i)\Big) = 1 - \frac{1}{n}\sum_{i=1}^n \frac{M(x_i)}{m}. \tag{1.6}$$

1.2 Locating the Global Maxima of Regression Function : Literature survey

In the literature, several fixed design methods are available for estimating the maxima of a regression function in a nonparametric setup, such as best-r-points-average method based on order statistics (Chen [2]), kernel method with adaptive bandwidth selector (Müller [17]) etc. But one unavoidable shortcoming of fixed design methods is that many unproductive trials may be performed in the sense that a large fraction of the design points may come from regions where the value of the regression function is quite small.

In the case of adaptive designs, there are mainly two well known methods for locating the maximum of a regression function. They are the Kiefer-Wolfowitz recursive stochastic approximation procedure (see *e.g.,* Kiefer & Wolfowitz [12]) in a nonparametric setup and the response surface method (RSM, see *e.g.,* Draper & Smith [7]) in a parametric setup. Both these methods can be viewed as extensions of the commonly used gradient methods in numerical analysis to the case when the objective function is observed with random error. And for maximizing a known function, it is well-known that the gradient method may fail to locate the global maximum if the objective function has some saddle points. Hotelling [11] has proposed a two-stage procedure where one conducts a pilot survey to get an initial estimate of the maximum and uses design based on quadratic modeling near the estimated maximum to estimate the maximum more accurately.

Response surface method is currently a popular method to find conditions that maximize yields. But the use of RSM to find conditions that maximize the response has some limitations. First, the data are assumed to follow normal distribution, although GLMs are recently being discussed (see e.g. McCullagh & Nelder [16]). Second, the surface and the peak are determined by a parametric equation, usually of quadratic type. This implies that all interactions between the predictors are assumed to be of product type. Moreover parametric models have their own limitations of not being flexible in the sense that one parametric equation is assumed to relate the response to the predictors over the entire range of the predictors considered.

The Kiefer-Wolfowitz procedure has been briefly discussed in section 1.3. Sacks [19] proved the asymptotic normality of the estimator obtained by this procedure under general setup though here also the conditions on $M(x)$ are restrictive in nature and not always satisfied in reality. He has also shown that by choosing appropriate parameters in the procedure one can make the convergence rate very close to $O_p(n^{-1/2})$ without ever achieving it. Chen [3] has proved that the minimax rate of convergence for estimating the maximum of a regression function over a class which contains enough functions with p-th derivative bounded by a constant K, is $O_p(n^{-(p-1)/2p})$ for fixed designs and stochastic approximation procedures(see Stone [20], [21], Chen [3]).

1.3 Kiefer-Wolfowitz Procedure

Suppose that $\mathcal{X} = \mathbb{R}$ and the function $M(x)$ has a unique maximum at θ and is strictly increasing for $x < \theta$ and strictly decreasing for $x > \theta$.

1.3.1 Algorithm :

Let a_n, c_n be two infinite sequences of positive numbers such that,

$$\sum a_n = \infty, \quad c_n \to 0, \quad \sum a_n c_n < \infty \text{ and } \sum a_n^2 c_n^{-2} < \infty \tag{1.7}$$

(for example $a_n = 1/n, c_n = 1/n^{1/3}$).

Let X_1 be any point (fixed or random) from $\mathcal{X}$. Define $\{X_n : n \geq 2\}$ by the recursion

$$X_{n+1} = X_n - a_n c_n^{-1}[Y_{1,n} - Y_{2,n}], \tag{1.8}$$

where $Y_{1,n}$ and $Y_{2,n}$ are outcomes of independent noisy measurement of $M(X_n - c_n)$ and $M(X_n + c_n)$ respectively. (*i.e.* $Y_{1,n} = M(X_n - c_n) + \varepsilon_1, Y_{2,n} = M(X_n + c_n) + \varepsilon_2$ and $\varepsilon_1, \varepsilon_2$ are independent.)

Then under certain regularity conditions on $M(x)$, X_n converges to θ in probability (see *e.g.,* Kiefer & Wolfowitz [12]).

1.3.2 Asymptotic normality of X_n

Suppose that $a_n = An^{-1}, \forall n \geq 1$ for some $A > 0$ and $\{c_n\}$ be a sequence of positive real numbers such that

$$c_n \to 0, \quad n(c_n c_{n+1}^{-1} - 1) \to 0 \text{ as } n \to \infty \text{ and } \sum (nc_n)^{-2} < \infty \tag{1.9}$$

(for example, $c_n = \prod_{i=1}^{n}(1 - 1/\log(i+1))$).

Then under certain conditions on A and $M(x)$, $\sqrt{n}c_n(X_n - \theta)$ is asymptotically normal with mean 0 and variance $K\sigma^2$, where K is a constant depending on A and $M(x)$(see *e.g.,* Sacks [19]).

1.4 Limitations of Kiefer-Wolfowitz method:

1. This procedure is applicable only when the regression function $M(x)$ has a unique maximum. If $M(x)$ has multiple maxima, this procedure may fail to converge to any one of them.

2. The conditions on $M(x)$ are very restrictive. Conditions such as strict monotonicity of $M(x)$ in both $[\theta, \infty)$ and $(-\infty, \theta]$, bounded away-ness of right and left derivatives from zero and infinity in absolute value etc. restrict the applicability of the method to general regression functions (even for polynomial functions).

3. In this procedure, one always does the experiment at $X_n \pm c_n$, where X_n is the estimate of the maximum at the n-th stage. Hence, even if we get a good estimate of the true maximum, throughout the experiment many "unproductive" trials are performed.

4. This procedure cannot be easily generalized to higher dimensional design space $\mathcal{X}$.

Chapter 2

Proposed Methodology

We have proposed two methods based on kernel regression estimation to efficiently estimate the set $\mathcal{X}_{opt}$ as well as the maximum value m. Both methods use a fraction of the total sample to get an initial estimate of the regression function and then sequentially update the regression estimate using the subsequent data and thereby estimate $\mathcal{X}_{opt}$.

2.1 First Method

This method is motivated by the "simulated annealing algorithm" (see Kirkpatrick et.al. [13], [14]), which is used to find the global optima of a function when the function is known but the set $\mathcal{X}$ is a very large finite set. Here we discretized $\mathcal{X}$ making it a fine grid. This is not really a restriction since if the distance between two points in $\mathcal{X}$ is very small, we cannot distinguish them in terms of the values of the function M in practice, and the grid can be made as fine as we want. Let $\mathcal{X}^*$ be the set consisting of the grid points. Suppose $|\mathcal{X}^*| = N$. Also suppose that the total sample size, that can be used in the experiment is n.

Given total sample size n, we shall generate the design points in (t_n+1) stages and after each stage we shall update the estimate of the regression function, hence the estimate of $\mathcal{X}_{opt}$. We shall use a sample of size $n_0 = n_0(n)$ to get an initial estimate of the function $M(x)$. Assume that

$$n_0(n) \to \infty \text{ and } \frac{n_0(n)}{n} \to 0 \text{ as } n \to \infty. \tag{2.1}$$

The above conditions on n_0 is needed so that for large n we have enough data to estimate $M(x)$ efficiently even at the initial stage, but the fraction of the total sample used at the initial stage is small. Let $n_i(n)$ be the sample size that will be used in the i-th stage, $i = 1, 2, \ldots, t_n$. Define $n_i(n) = 0$ for $i > t_n$. So we have a sequence of nonnegative integers depending on n, $(n_0(n), n_1(n), \ldots,)$ such that $n_i(n) > 0$ for $i = 0, 1, 2, \ldots, t_n$, and $n_i(n) = 0$ for $i > t_n$.

In what follows, the following assumptions and notations will be used,

- Y_i is the outcome of a noisy measurement at X_i for $i = 1, 2, \ldots, n$, and
$$Y_i = M(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, 2, \ldots, n$$
where the ε_i's are independent with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = 1$.

- $K(.)$ is a positive symmetric kernel function on $\mathbb{R}^d$ such that
$$\int_{\mathbb{R}^d} K(x)dx = 1, \int_{\mathbb{R}^d} \|x\| \, K(x)dx < \infty$$

- $H_n(X_1, Y_1, \ldots, X_n, Y_n)$ is a consistent bandwidth based on the sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ (see Wand & Jones [22], Fan & Gijbels [8]). Denote $H_{n_1+\cdots+n_k}(X_1, Y_1, \ldots, X_{n_1+\cdots+n_k}, Y_{n_1+\cdots+n_k})$ by h_k.

- Given a sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{n_0+n_1+\cdots+n_{k-1}}, Y_{n_0+n_1+\cdots+n_{k-1}})$ of size $n_0 + n_1 + \cdots + n_{k-1}$, let $\hat{M}_k$ be a kernel regression estimate of $M(x)$ based on the available samples, using bandwidth h_k and kernel $K(.)$.

- $T_k^n = \Delta^*(\hat{M}_p)/(c_n \log(k + 1 - n_0)), k > n_0(n)$ is a random sequence of real numbers, where $c_n \in (0, 1) \; \forall \; n \geq 1$, p is such that $\sum_{i=0}^{p-1} n_i < k \leq \sum_{i=0}^{p} n_i$ and
$$\Delta^*(g) = \sup_{x \in \mathcal{X}^*} g(x) - \inf_{x \in \mathcal{X}^*} g(x).$$

2.1.1 Algorithm

1. Let $X_1, X_2, \ldots, X_{n_1}$ be n_1 random points chosen uniformly from $\mathcal{X}^*$.

2. At the k-th step, define recursively for $i = 1, 2, \ldots, n_k$ $Z_{i,k}$ as a random variable on $\mathcal{X}^*$, such that
$$Z_{i,k} | X_1, X_2, \ldots, X_{S(k-1)+i-1} \sim \text{Uniform}(\mathcal{X}^*).$$

and

$$X_{S(k-1)+i} = \begin{cases} Z_{i,k}, \text{ w. p. } \exp\,[-\frac{(\hat{M}_k(X_{S(k-1)+i-1}) - \hat{M}_k(Z_{i,k}))^+}{T_{S(k-1)+i}}] \\ X_{S(k-1)+i}, \text{ w. p. } 1 - \exp\,[-\frac{(\hat{M}_k(X_{S(k-1)+i-1}) - \hat{M}_k(Z_{i,k}))^+}{T_{S(k-1)+i}}] \end{cases}$$

where $S(k) = \sum_{j=0}^{k} n_j, j \geq 1$.

3. Continue this procedure for $k = 1, 2, \ldots, t_n$.

Fact

If, at any point x, we can evaluate $M(x)$ without any error (*i.e.,* no ε is there in that case), the above procedure with $n_0 = 1$ and $M(x)$ in place of $\hat{M}_k(x)$ gives an inhomogeneous Markov chain $\{X_n\}$ such that

$$\lim_{n \to \infty} P(X_n \in \mathcal{X}^*_{opt}) = 1$$

where $\mathcal{X}^*_{opt} = \{x \in \mathcal{X}^* : M(x) = \sup_{z \in \mathcal{X}^*} M(z)\}$. Moreover in this case the limiting stationary distribution for $\{X_n\}$ exists, and it is uniform over the set $\mathcal{X}_{opt}$(see *e.g.,* Winkler [23]).

2.2 Second Method

This method is based on one simple observation that $\mathcal{X}^k_{opt} \downarrow \mathcal{X}_{opt}$, where $\mathcal{X}^k_{opt} = \{x : M(x) \geq m(1 - c_k)\}$ and $\{c_k\}$ is a decreasing sequence of positive real numbers converging to 0. Here, the basic strategy would be to estimate the set $\mathcal{X}^k_{opt}$ at the k-th stage.

As before let $(n_0(n), n_1(n), \ldots,)$ be a sequence of positive integers depending on n such that $n_i(n) > 0$ for $i = 0, 1, 2, \ldots, t_n$, $n_i(n) = 0$ for $i > t_n$ and $\sum_{i=1}^{\infty} n_i(n) = n$. At the k-th step, we shall use a sample of size $n_k = n_k(n)$. Also, assume that

$$n_0(n) \to \infty \text{ and } \frac{n_0(n)}{n} \to 0 \text{ as } n \to \infty. \tag{2.2}$$

For a set $S \subset \mathcal{X}$, define

$$B(S, \varepsilon) = \{x \in \mathcal{X} : d(x, S) \leq \varepsilon\}.$$

At the k-th stage, the design points used in the experiment are denoted by $x_{k,1}, x_{k,2}, \ldots, x_{k,n_k}$, and these points are generated from the random design

$\mathcal{D}_k$, where $\mathcal{D}_0$ is the random uniform design over $\mathcal{X}$ and we shall update $\mathcal{D}_k$ recursively. Also, let $y_{k,i}$ be the outcome of a noisy measurement at $x_{k,i}$ for $i = 1, 2, \ldots, n$, and

$$y_{k,i} = M(x_{k,i}) + \sigma(x_{k,i})\varepsilon_{k,i}, \text{ for } i = 1, 2, \ldots, n_k, k = 1, 2, \ldots, k(n),$$

where $\varepsilon_{k,i}$'s are independent with

$$E(\varepsilon_{k,i}) = 0, Var(\varepsilon_{k,i}) = 1.$$

Let $K(.)$ be a positive symmetric kernel function on $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} K(x)dx = 1, \int_{\mathbb{R}^d} \|x\| K(x)dx < \infty.$$

Let H_n be a consistent bandwidth selector based on a sample of size n. Denote $H_{n_0+n_1+\ldots+n_k}(x_{1,1}, y_{1,1}, \ldots, x_{k,n_k}, y_{k,n_k})$ by h_k.

Given a sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{n_0+n_1+\cdots+n_{k-1}}, Y_{n_0+n_1+\cdots+n_{k-1}})$ of size $n_0 + n_1 + \cdots + n_{k-1}$, let $\hat{M}_k$ be a kernel regression estimate of $M(x)$ based on the available samples, using bandwidth h_k and kernel $K(.)$.

Let $\{\alpha_k^n\}, \{c_k^n\}$ and $\{r_k^n\}$ be sequences of real numbers in $(0, 1)$ such that for fixed n, they are decreasing in k and for fixed k, they converge to zero.

2.2.1 Algorithm

1. At the k-th step, generate $n_k(n)$ points $x_{k,1}, x_{k,2}, \ldots, x_{k,n_k}$ from the random design $\mathcal{D}_k^n$, where

$$\begin{aligned}
\mathcal{D}_0^n &= \text{ Uniform design distribution over } \mathcal{X}, && (2.3)\\
S_k^n &= \{x \in \mathcal{X} : \hat{M}_k(x) \geq (1 - c_k^n) \sup_{y \in \mathcal{X}} \hat{M}_k(y)\}, && (2.4)\\
U_k^n &= B(S_k^n, r_k^n), && (2.5)\\
\mathcal{C}_k^n &= \text{ Uniform design distribution over } U_k^n, \text{ and} && (2.6)\\
\mathcal{D}_k^n &= \alpha_k^n \mathcal{D}_{k-1}^n + (1 - \alpha_k^n)\mathcal{C}_k^n && (2.7)\\
&= \prod_{i=1}^{k} \alpha_i^n \mathcal{D}_0 + \sum_{i=1}^{k} (1 - \alpha_i^n) \prod_{j=i+1}^{k} \alpha_j^n \mathcal{C}_i^n. && (2.8)
\end{aligned}$$

2. Repeat the above step for $k = 1, 2, \ldots, t_n$.

Chapter 3

Simulation Studies on the performance of the proposed methods

3.1 Simulation Plan

In the numerical experiment, we have considered three functions of three different types

1. a function having a unique global maximum which is also the only local maximum,

2. a function having a unique global maximum and some local maxima, and

3. a function having several global as well as local maxima.

All the functions are defined on $[0, 1]$. For each function, we have considered σ with values such that the noise to signal ratio (*i.e.*, $\sigma/\int_0^1 M(x)dx$)is 1 and 2. For each combination, two sample of sizes 50 and 100 are considered. For the Kiefer-Wolwowitz procedure, we have considered the sequences

$$a_n = n^{-1}, c_n = n^{-\frac{1}{3}}. \tag{3.1}$$

In both of our proposed methods, we have considered

$$\begin{aligned} n_1(n) = 2\lceil\log(n)\rceil, n_i(n) = \lceil\log(n - \sum_{k=1}^{i-1} n_k(n))\rceil, i = 2, 3, \ldots \\ c_n = (n+1)^{-4}, r_n = [log(n+1)]^{-2}, \alpha_n = (n+1)^{-2}. \end{aligned} \tag{3.2}$$

For both of our proposed methods, Nadaraya-Watson kernel regression estimator with normal kernel and cross-validation bandwidth was used.

3.2 Function with a unique maximum

Here we consider the regression function (figure 3.1),

$$M(x) = 6x(1-x). \tag{3.3}$$

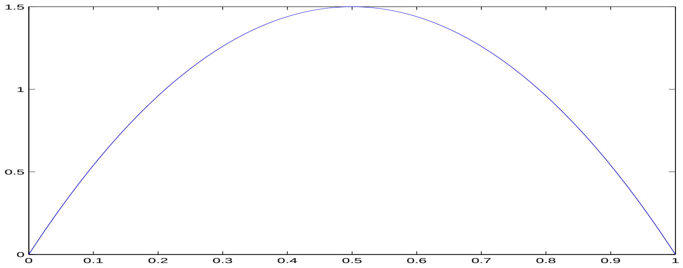

Figure 3.1: Plot of $M(x)$ given by (3.3).

$M(x)$ satisfies all the assumptions in Kiefer-Wolwowitz procedure, it has unique maximum at $x = 0.5$ and the range of the function is 1. Also $\int_0^1 M(x) = 1$.

3.2.1 Results for Kiefer-Wolwowitz procedure

The histograms of the x_i's for sample sizes 50 and 100 are shown in figure 3.2 and figure 3.3 respectively. Note that the x_i's are sampled around the true maximum but not at the maximum. So C_1 and C_2 criteria values are high, though the method gives a 'good' estimate of the true maximum.

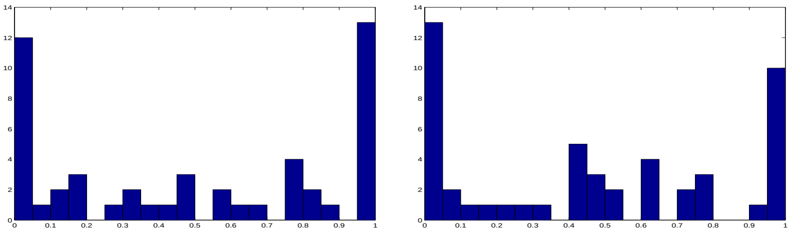

Figure 3.2: Histograms of the x_n's for Kiefer-Wolwowitz procedure with sample size 50 and, noise to signal ratio 1 and 2 respectively.

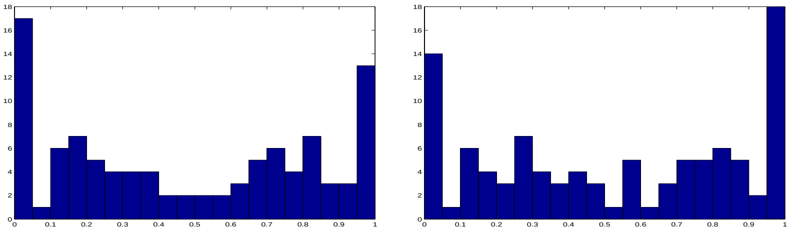

Figure 3.3: Histograms of the x_n's for Kiefer-Wolwowitz procedure with sample size 100 and, noise to signal ratio 1 and 2 respectively.

Sample size	50		100	
Noise to signal ratio	1	2	1	2
Estimate of maximum	0.4994	0.5437	0.4871	0.5445
C_1	63.2042	59.1585	48.3130	54.0282
C_2	74.8308	87.1079	61.8069	81.1042

Table 3.1: Results of K-W procedure for $M(x)$ given by (3.3)

3.2.2 Results for the first proposed method

The histograms of the sampled x_i's for sample sizes 50 and 100 are shown in figure 3.4 and figure 3.5 respectively. Note here the x_i's are clustered at the true maximum and the C_1, C_2 criteria values are smaller than that for the Kiefer-Wolwowitz procedure.

Sample size	50		100	
Noise to signal ratio	1	2	1	2
C_1	26.8417	13.6503	10.7209	15.1867
C_2	45.6964	60.7539	30.5762	57.6062

Table 3.2: Results of the first proposed method for $M(x)$ given by (3.3).

3.2.3 Results for the second proposed method

1) **Sample size 50 and noise to signal ratio = 1**

Here the initial sample size is $n_1 = 8$. The estimate of the regression function and the histograms of the x_i's sampled at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.6. The final density $\mathcal{D}_n$ and the scatter plot of the observations are shown in figure 3.7.

2) **Sample size 50 and noise to signal ratio = 2**

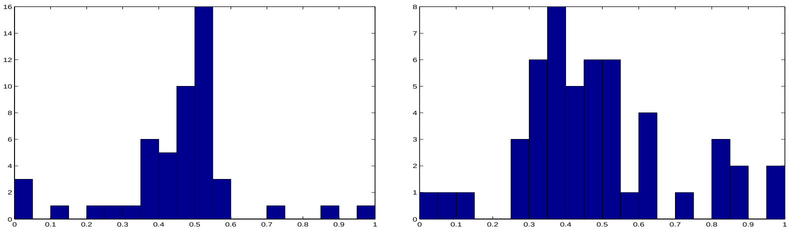

Figure 3.4: Histograms of the x_n's for the first proposed method with sample size 50 and, noise to signal ratio 1 and 2 respectively.

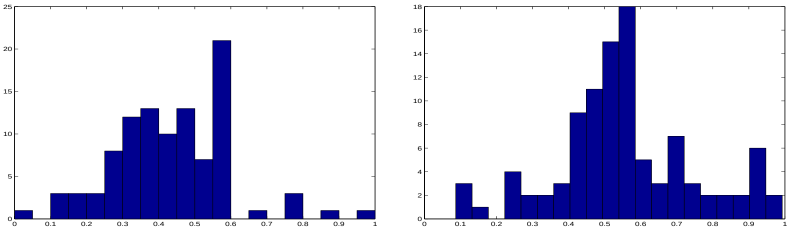

Figure 3.5: Histograms of the x_n's for the first proposed method with sample size 100 and, noise to signal ratio 1 and 2 respectively.

Here the initial sample size is $n_1 = 8$. The estimate of the regression function and the histograms of the x_i's sampled at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.8. The final density $\mathcal{D}n$ and the scatter plot of the observations are shown in figure 3.9.

3) **Sample size 100 and noise to signal ratio = 1**

Here the initial sample size is $n_1 = 10$. The estimate of the regression function and the histograms of the x_i's sampled at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.10. The final density $\mathcal{D}_n$ and the scatter plot of the observations are shown in figure 3.11.

4) **Sample size 100 and noise to signal ratio = 2**

Here the initial sample size is $n_1 = 10$. The estimate of the regression function and the histograms of the x_i's sampled at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.12. The final density $\mathcal{D}_n$ and the scatter plot of the observations are shown in figure 3.13.

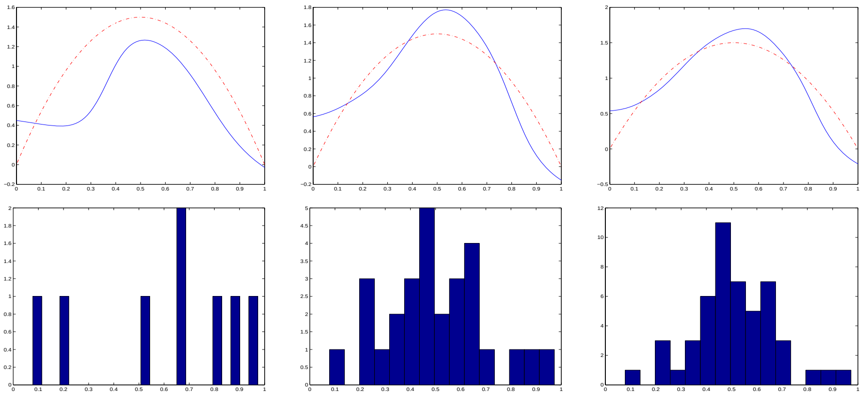

Figure 3.6: Estimate of $M(x)$ and the histograms of the x_i's for the second proposed method with sample size 50 and noise to signal ratio 1, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

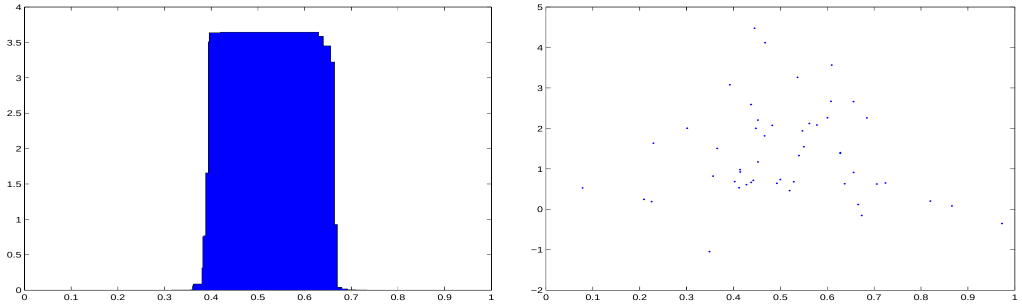

Figure 3.7: Density of $\mathcal{D}_n$ and the scatter plot of the observations in the second proposed method with sample size 50 and noise to signal ratio 1.

Sample size	50		100	
Noise to signal ratio	1	2	1	2
Estimate of $\mathcal{X}_{opt}$	[0.41,0.67]	[0.36,0.71]	[0.37,0.64]	[0.42,0.61]
C_1	9.0437	20.4619	13.6940	8.0167
C_2	36.1963	53.9820	37.5392	57.9038

Table 3.3: Results of the second proposed method for $M(x)$ as given in (3.3)

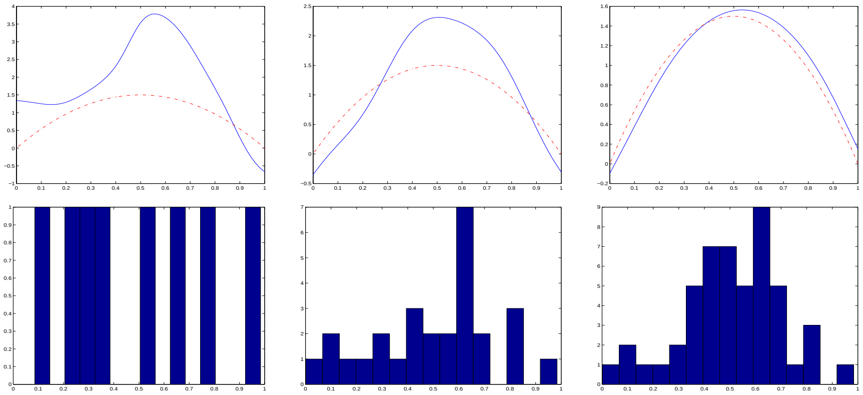

Figure 3.8: Estimate of $M(x)$ and the histograms of the x_i's for the second proposed method with sample size 50 and noise to signal ratio 2, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

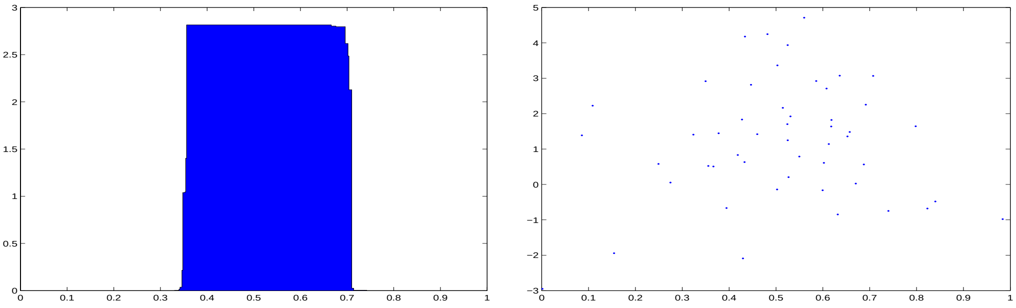

Figure 3.9: Density of $\mathcal{D}_n$ and scatter plot of the observations in the second proposed method with sample size 50 and noise to signal ratio 2.

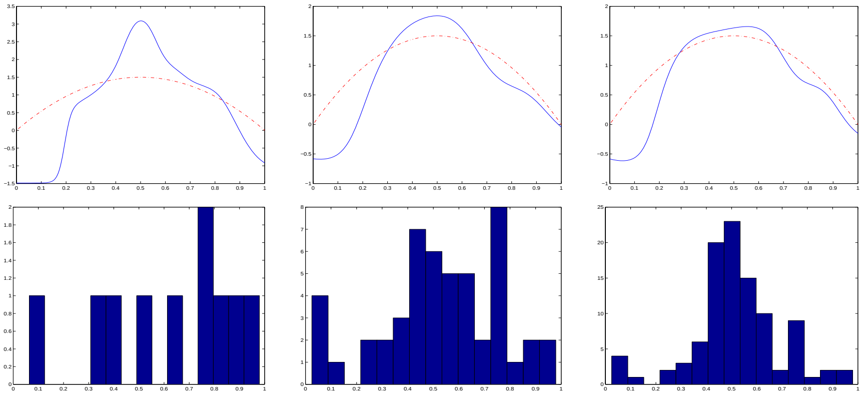

Figure 3.10: Estimate of $M(x)$ and histograms of x_i's for the second proposed method with sample size 100 and noise to signal ratio 1, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

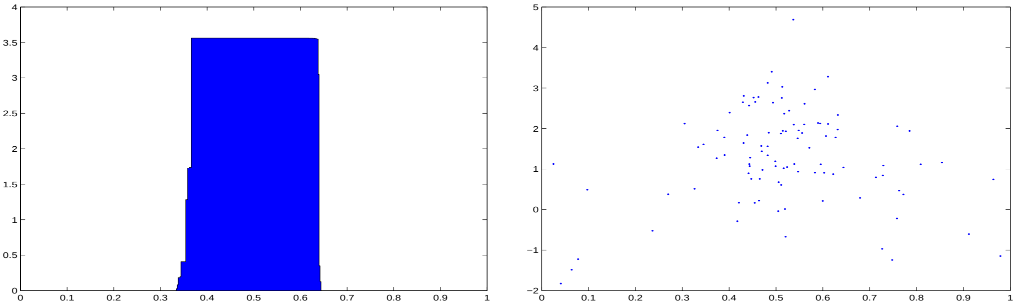

Figure 3.11: Density of $\mathcal{D}_n$ and scatter plot of the observations in the second proposed method with sample size 50 and noise to signal ratio 1.

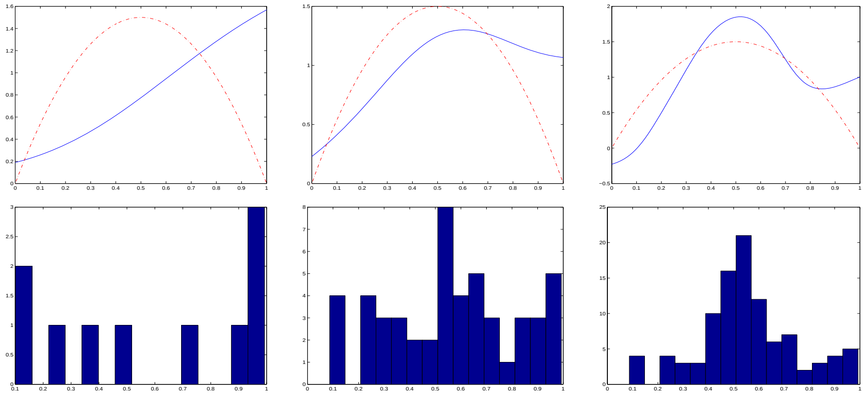

Figure 3.12: Estimate of $M(x)$ and histograms of x_i's for the second proposed method with sample size 100 and noise to signal ratio 2, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

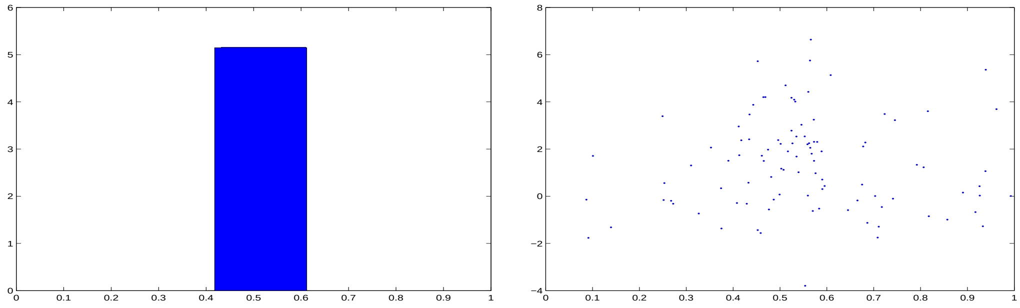

Figure 3.13: Density of $\mathcal{D}_n$ and the scatter plot of the observations in the second proposed method with sample size 100 and noise to signal ratio 2.

3.3 Function with a unique global maximum and two local maxima

Here we consider the regression function(figure 3.14),

$$\begin{aligned} M(x) = 1.42(&\exp[-50(x-.1)^2] + 2\exp[-200(x-.3)^2] \\ &+ \exp[-50(x-.8)^2]). \end{aligned} \tag{3.4}$$

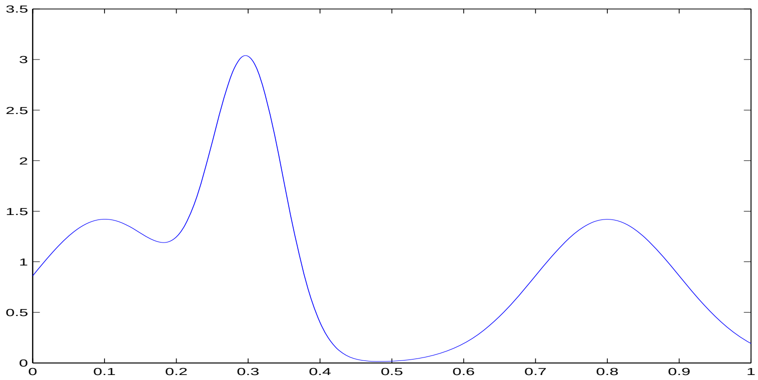

Figure 3.14: Plot of $M(x)$ given by (3.4).

$M(x)$ has unique global maximum at $x = 0.3$ and two local maxima at $x = 0.1$ and $x = 0.8$. The Range of the function is 3.03 and $\int_0^1 M(x) = 0.999$.

3.3.1 Results for Kiefer-Wolwowitz procedure

The histograms of the x_i's for sample sizes 50 and 100 are shown in figure 3.15 and figure 3.16 respectively.

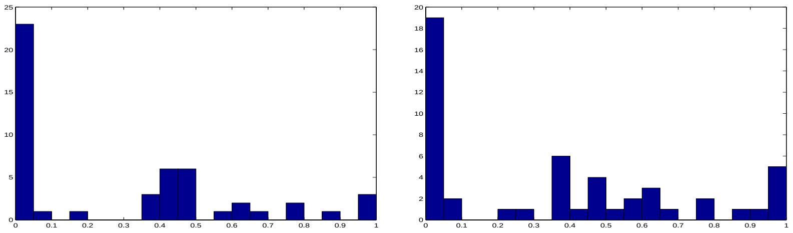

Figure 3.15: Histograms of the x_n's for Kiefer-Wolwowitz procedure with sample size 50 and, noise to signal ratio 1 and 2 respectively.

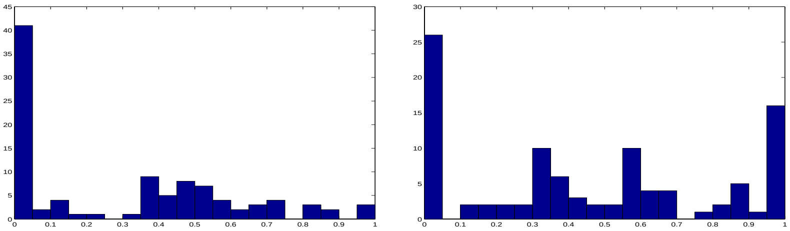

Figure 3.16: Histograms of the x_n's for Kiefer-Wolwowitz procedure with sample size 100 and, noise to signal ratio 1 and 2 respectively.

Sample size	50		100	
Noise to signal ratio	1	2	1	2
Estimate of maximum	0.50877	0.81947	0.12629	0.88406
C_1	92.536	75.981	77.315	82.725
C_2	92.536	79.838	77.372	84.942

Table 3.4: Results of K-W procedure for $M(x)$ given by (3.4).

3.3.2 Results for the first proposed method

The histograms of the sampled x_i's for sample sizes 50 and 100 are shown in figure 3.17 and figure 3.18 respectively. Note here the x_i's are clustered at the true maximum and the C_1, C_2 criteria values are smaller than the that corresponding to Kiefer-Wolwowitz procedure.

Sample size	50		100	
Noise to signal ratio	1	2	1	2
C_1	31.548	8.3149	27.819	33.285
C_2	39.141	30.305	35.441	47.567

Table 3.5: Results of the first proposed method for $M(x)$ given by (3.4).

3.3.3 Results for the second proposed method

1) **Sample size 50 and noise to signal ratio =1**

Here the initial sample size is $n_1 = 8$. The estimate of the regression function and the histograms of the x_i's sample at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.19. The final density $\mathcal{D}_n$ and scatter plot of the observations are shown in figure 3.20.

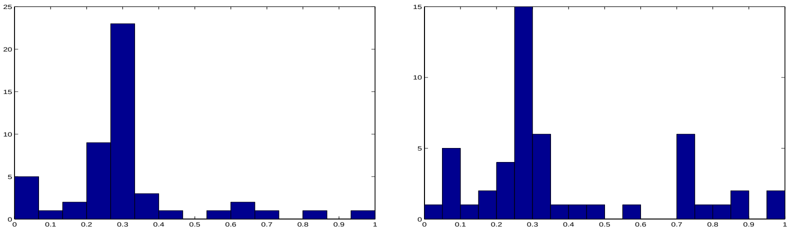

Figure 3.17: Histograms of the x_n's for the first proposed method with sample size 50 and, noise to signal ratio 1 and 2 respectively.

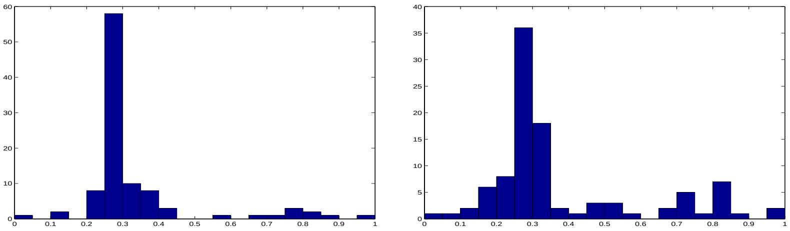

Figure 3.18: Histograms of the x_n's for the first proposed method with sample size 100 and, noise to signal ratio 1 and 2 respectively.

2) **Sample size 50 and noise to signal ratio= 2**

Here the initial sample size is $n_1 = 8$. The estimate of the regression function and the histograms of the x_i's sample at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.21. The final density $\mathcal{D}_n$ and scatter plot of the observations are shown in figure 3.22.

3) **Sample size 100 and noise to signal ratio= 1**

Here the initial sample size is $n_1 = 10$. The estimate of the regression function and the histograms of the x_i's sample at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.23. The final density $\mathcal{D}_n$ and scatter plot of the observations are shown in figure 3.24.

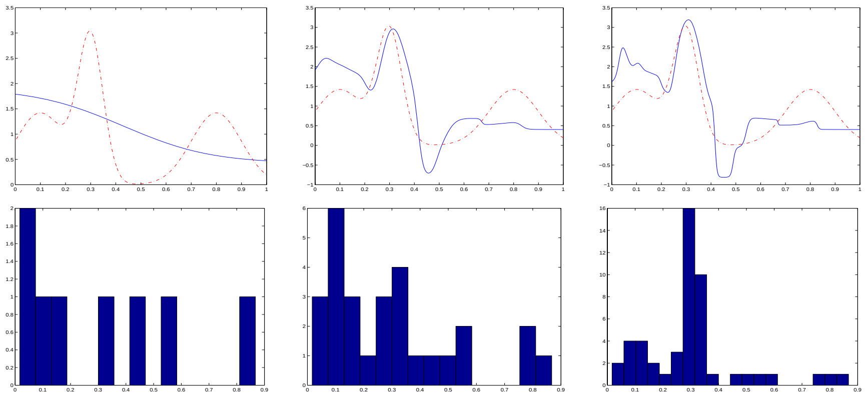

Figure 3.19: Estimate of $M(x)$ and the histograms of the x_i's for the second proposed method with sample size 50 and noise to signal ratio 1, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

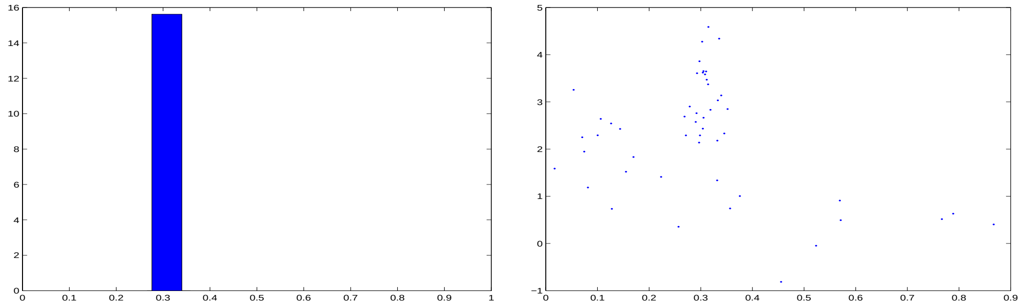

Figure 3.20: Density of $\mathcal{D}_n$ and the scatter plot of the observations in the second proposed method with sample size 50 and noise to signal ratio 1.

4) **Sample size 100 and noise to signal ratio= 2**

Here the initial sample size is $n_1 = 10$. The estimate of the regression function and the histograms of the x_i's sample at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.25. The final density $\mathcal{D}_n$ and scatter plot of the observations are shown in figure 3.26.

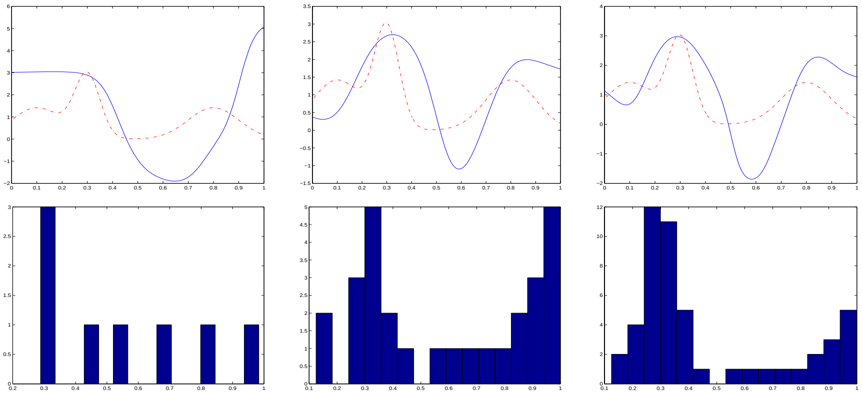

Figure 3.21: Estimate of $M(x)$ and the histograms of the x_i's for the second proposed method with sample size 50 and noise to signal ratio 2, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

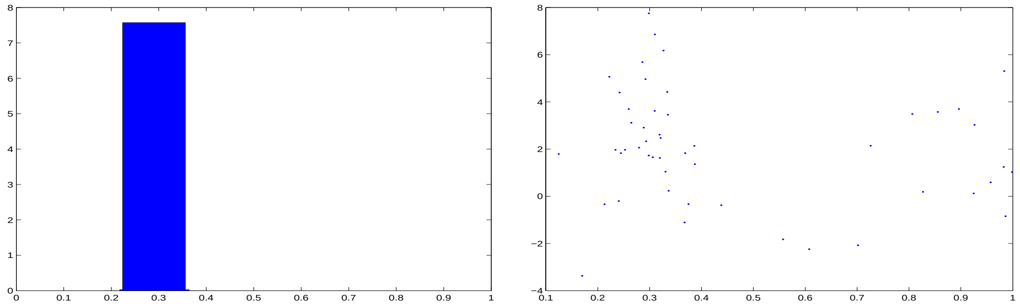

Figure 3.22: Density of $\mathcal{D}_n$ and the scatter plot of the observations in the second proposed method with sample size 50 and noise to signal ratio 2.

Sample size	50		100	
Noise to signal ratio	1	2	1	2
Estimate of $\mathcal{X}_{opt}$	[0.267,0.332]	[0.225,0.355]	[0.285,0.325]	[0.246,0.332]
C_1	26.118	32.584	29.972	27.737
C_2	31.973	50.116	35.115	44.921

Table 3.6: Results of the second proposed method for $M(x)$ as given in (3.4).

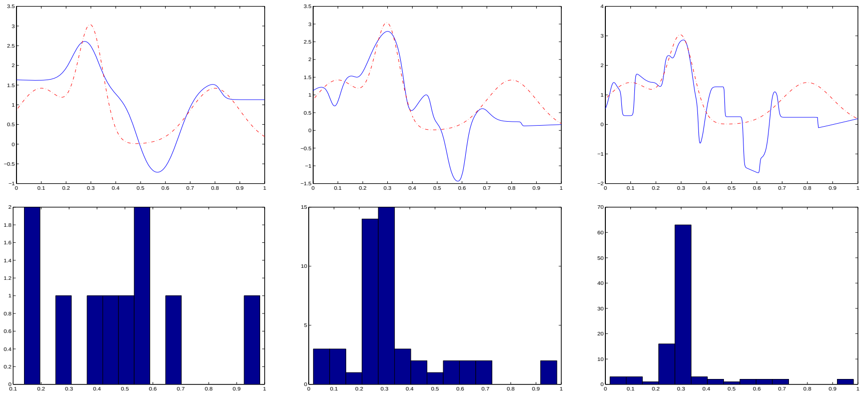

Figure 3.23: Estimate of $M(x)$ and the histograms of the x_i's for the second proposed method with sample size 100 and noise to signal ratio 1, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

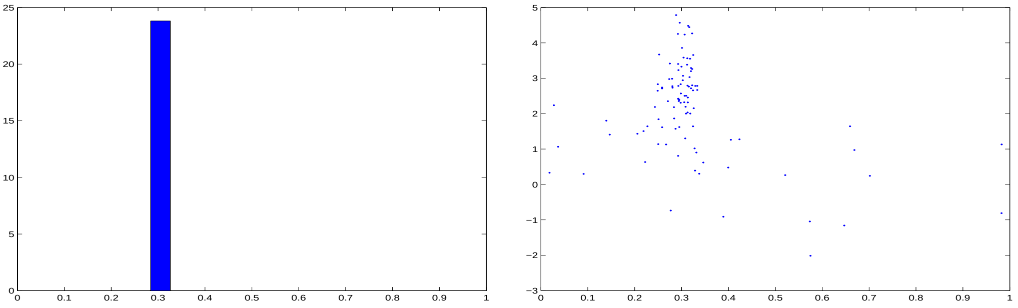

Figure 3.24: Density of $\mathcal{D}_n$ and the scatter plot of the observations in the second proposed method with sample size 50 and noise to signal ratio 1.

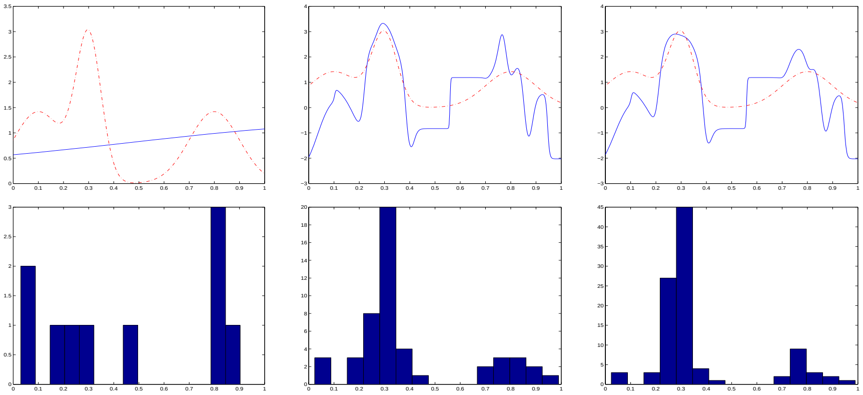

Figure 3.25: Estimate of $M(x)$ and the histograms of the x_i's for the second proposed method with sample size 100 and noise to signal ratio 2, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

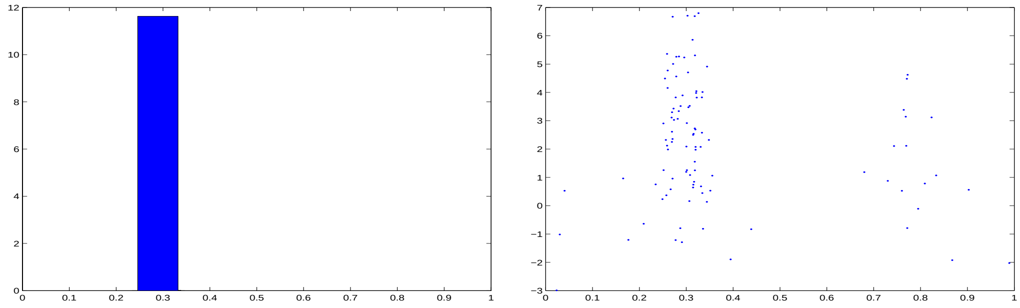

Figure 3.26: Density of $\mathcal{D}_n$ and the scatter plot of the observations in the second proposed method with sample size 100 and noise to signal ratio 2.

3.4 Function with two global maxima and three local maxima

Here we consider the regression function(figure 3.27),

$$\begin{aligned} M(x) = \exp(-50(x-0.1)^2) + 1.3\exp(-200(x-0.3)^2) \\ + \exp(-50(x-0.5)^2) + 1.3\exp(-200(x-0.7)^2) \\ + \exp(-50(x-0.9)^2). \end{aligned} \tag{3.5}$$

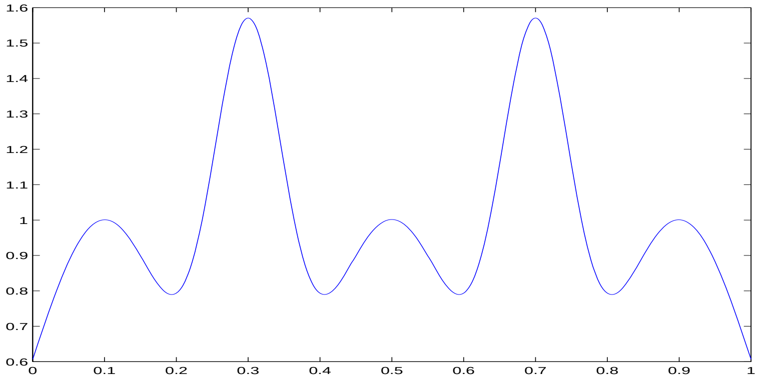

Figure 3.27: Plot of $M(x)$ given by (3.5).

$M(x)$ has global maxima at $x = 0.3$ and $x = 0.7$ and three local maxima at $x = 0.1$, $x = 0.5$ and $x = 0.9$. The Range of the function is 1.57 and $\int_0^1 M(x) = 0.999$.

3.4.1 Results for Kiefer-Wolwowitz procedure

The histograms of the x_i's for sample sizes 50 and 100 are shown in figure 3.28 and figure 3.29 respectively.

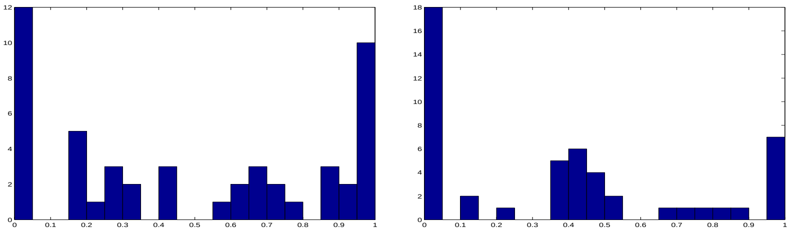

Figure 3.28: Histograms of the x_n's for Kiefer-Wolwowitz procedure with sample size 50 and, noise to signal ratio 1 and 2 respectively.

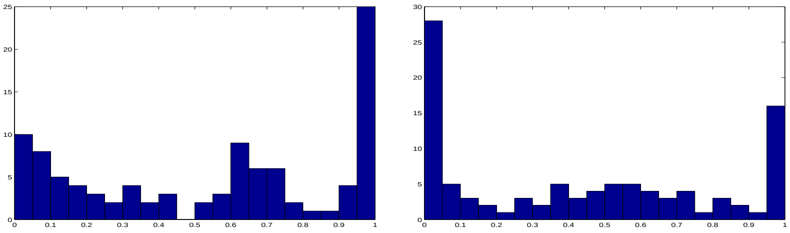

Figure 3.29: Histograms of the x_n's for Kiefer-Wolwowitz procedure with sample size 100 and, noise to signal ratio 1 and 2 respectively.

Sample size	50		100	
Noise to signal ratio	1	2	1	2
Estimate of maximum	0.51004	0.3169	0.45883	0.46019
C_1	60.58	35.868	39.33	46.455
C_2	66.809	68.692	49.314	74.206

Table 3.7: Results of K-W procedure for $M(x)$ given by (3.5)

3.4.2 Results for the first proposed method

The histograms of the sampled x_i's for sample sizes 50 and 100 are shown in figure 3.30 and figure 3.31 respectively. Note here the x_i's are clustered at the true maxima and the C_1, C_2 criteria values are smaller than that corresponding to the Kiefer-Wolfowitz procedure..

Sample size	50		100	
Noise to signal ratio	1	2	1	2
C_1	24.991	47.531	29.196	34.867
C_2	41.466	84.867	44.057	65.387

Table 3.8: Results of the first proposed method for $M(x)$ given by (3.5).

3.4.3 Results for the second proposed method

1) **Sample size 50 and noise to signal ratio= 1**

Here the initial sample size is $n_1 = 8$. The estimate of the regression function and the histograms of the x_i's sample at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.32. The final density $\mathcal{D}_n$ and the scatter plot of the observations are shown in figure 3.33.

2) **Sample size 50 and noise to signal ratio =2**

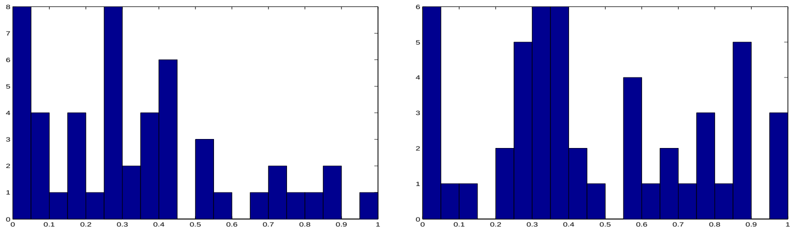

Figure 3.30: Histograms of the x_n's for the first proposed method with sample size 50 and, noise to signal ratio 1 and 2 respectively.

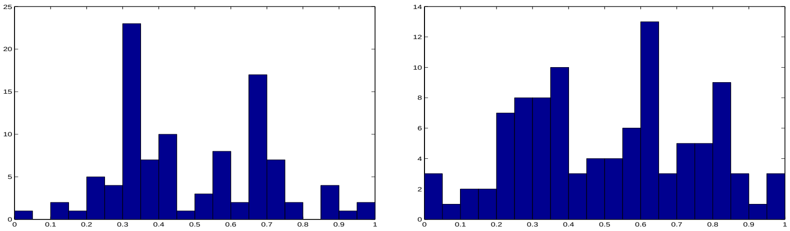

Figure 3.31: Histograms of the x_n's for the first proposed method with sample size 100 and, noise to signal ratio 1 and 2 respectively.

Here the initial sample size is $n_1 = 8$. The estimate of the regression function and the histograms of the x_i's sample at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.34. The final density $\mathcal{D}_n$ and the scatter plot of the observations are shown in figure 3.35.

3) **Sample size 100 and noise to signal ratio =1**

Here the initial sample size is $n_1 = 10$. The estimate of the regression function and the histograms of the x_i's sample at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.36. The final density $\mathcal{D}_n$ and the scatter plot of the observations are shown in figure 3.37.

4) **Sample size 100 and noise to signal ratio= 2**

Here the initial sample size is $n_1 = 10$. The estimate of the regression function and the histograms of the x_i's sample at the initial stage, the middle stage when half of the total sample has been used and the final stage are shown in figure 3.38. The final density $\mathcal{D}_n$ and the scatter plot of the observations are shown in figure 3.39.

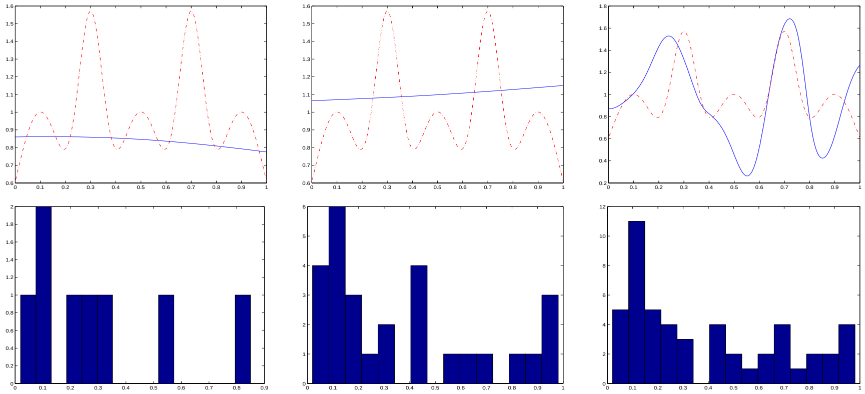

Figure 3.32: Estimate of $M(x)$ and histograms of x_i's for the second proposed method with sample size 50 and noise to signal ratio 1, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

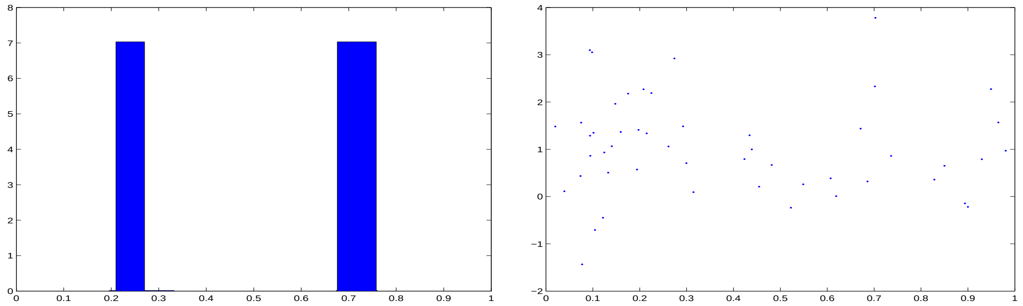

Figure 3.33: Density of $\mathcal{D}_n$ and the scatter plot of the observations in the second proposed method with sample size 50 and noise to signal ratio 1.

Sample size	50		100	
Noise to signal ratio	1	2	1	2
Estimate of $\mathcal{X}_{opt}$	[0.211,0.267] $\cup$ [0.673,0.756]	[0.01,0.65]	[0.28,0.314] $\cup$ [0.675,0.74]	[0.258,0.33] $\cup$ [0.645,0.732]
C_1	33.686	37.705	23.669	44.836
C_2	46.865	76.133	36.774	82.229

Table 3.9: Results of the second proposed method for $M(x)$ given in (3.5).

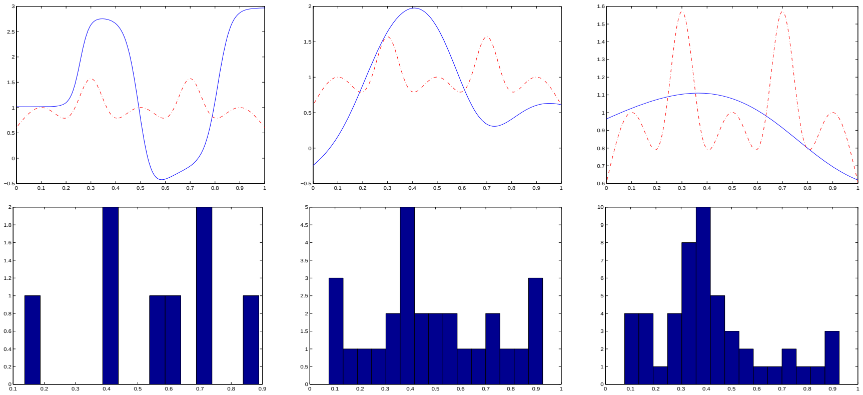

Figure 3.34: Estimate of $M(x)$ and histograms of x_i's for the second proposed method with sample size 50 and noise to signal ratio 2, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

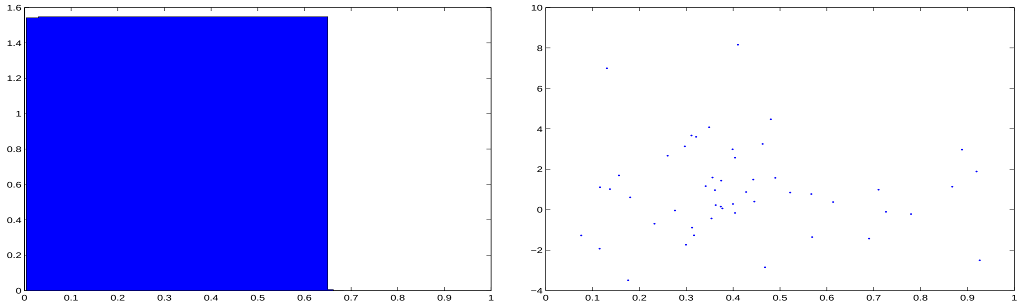

Figure 3.35: Density of $\mathcal{D}_n$ and the scatter plot of the observations in the second proposed method with sample size 50 and noise to signal ratio 2.

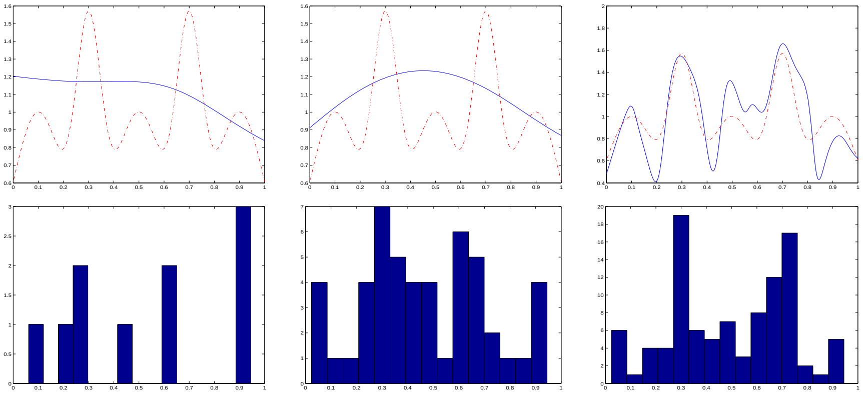

Figure 3.36: Estimate of $M(x)$ and the histograms of the x_i's for the second proposed method with sample size 100 and noise to signal ratio 1, at the initial step, middle step and final step respectively.

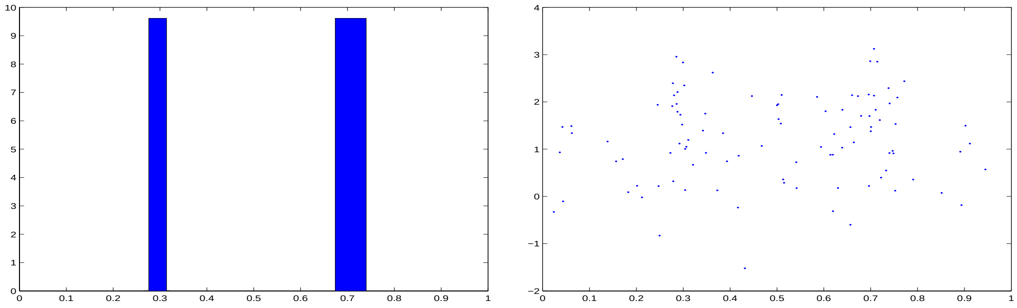

Figure 3.37: Density of $\mathcal{D}_n$ and the scatter plot of the observations in the second proposed method with sample size 50 and noise to signal ratio 1.

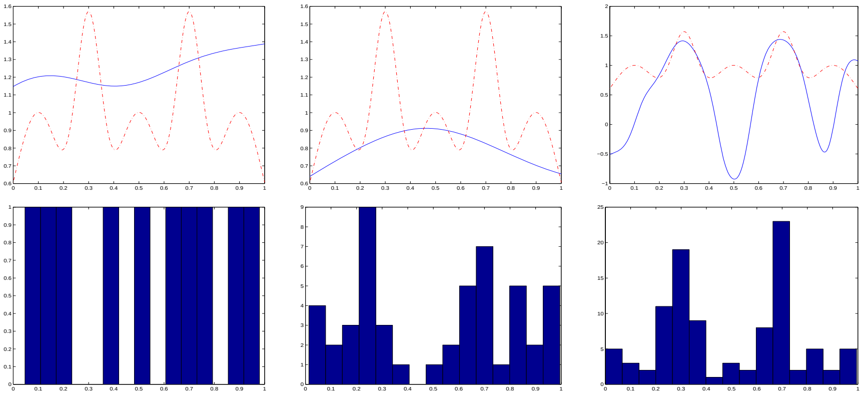

Figure 3.38: Estimate of $M(x)$ and histograms of x_i's for the second proposed method with sample size 100 and noise to signal ratio 2, at the initial stage, the middle stage when half of the total sample has been used and the final stage respectively.

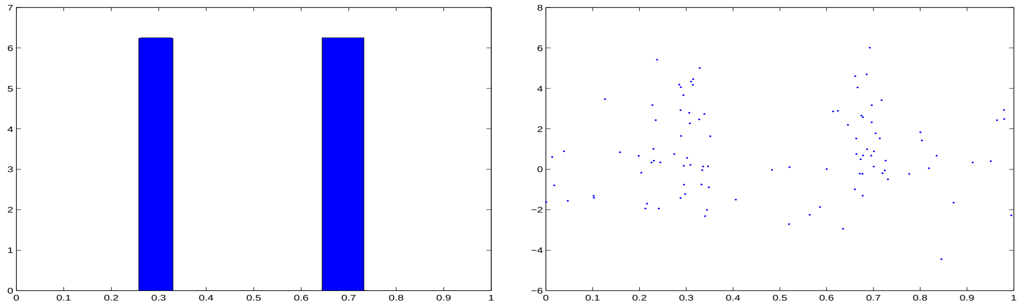

Figure 3.39: Density of $\mathcal{D}_n$ and the scatter plot of the observations in the second proposed method with sample size 100 and noise to signal ratio 2.

Chapter 4

Asymptotic results and their implications

Let $\mathcal{X}$ be a compact connected subset of $\mathbb{R}^d$ for some $d > 0$ and $\mathcal{X}^*$ be a finite subset of $\mathcal{X}$ with $|\mathcal{X}^*| = N$. For a function g on $\mathcal{X}$, define

$$m(g) = \sup_{x \in \mathcal{X}} g(x), \tag{4.1}$$

$$m^*(g) = \sup_{x \in \mathcal{X}^*} g(x), \tag{4.2}$$

$$\mathcal{X}_{opt}(g) = \{x \in \mathcal{X} : g(x) = m(g)\}, \tag{4.3}$$

$$\mathcal{X}^*_{opt}(g) = \{x \in \mathcal{X}^* : g(x) = m^*(g)\}, \tag{4.4}$$

$$\Delta(g) = \sup_{x \in \mathcal{X}} g(x) - \inf_{x \in \mathcal{X}} g(x), \tag{4.5}$$

$$\Delta^*(g) = \sup_{x \in \mathcal{X}^*} g(x) - \inf_{x \in \mathcal{X}^*} g(x), \tag{4.6}$$

$$\mu_{opt}(g) = \text{Uniform distribution over } \mathcal{X}_{opt}(g) \tag{4.7}$$

$$\text{and } \mu^*_{opt}(g) = \text{Uniform distribution over } \mathcal{X}^*_{opt}(g). \tag{4.8}$$

Suppose that M is a continuous function on $\mathcal{X}$. Assume that $|\mathcal{X}_{opt}(M)| = K < \infty$ and $|\mathcal{X}_{opt}(M)| = K^*$. For simplicity, we write $m, \Delta, \mu_{opt}, \mathcal{X}_{opt}$ in place of $m(M), \Delta(M), \mu_{opt}(M), \mathcal{X}_{opt}(M)$ respectively. Similarly for $m^*(M), \Delta^*(M)$, $\mu^*_{opt}(M)$, $\mathcal{X}^*_{opt}(M)$.

For asymptotic analysis we introduce a triangular array setup,

$$\begin{array}{llllll}
(X_1^{(1)}, Y_1^{(1)}) & & & & & \\
(X_1^{(2)}, Y_1^{(2)}) & (X_2^{(2)}, Y_2^{(2)}) & & & & \\
\quad\cdots & \quad\cdots & \ddots & & & \\
\quad\cdots & \quad\cdots & \cdots & \quad\cdots & & \\
(X_1^{(n)}, Y_1^{(n)}) & (X_2^{(n)}, Y_2^{(n)}) & \cdots & (X_{n-1}^{(n)}, Y_{n-1}^{(n)}) & (X_n^{(n)}, Y_n^{(n)}) & \\
\quad\cdots & \quad\cdots & \cdots & \quad\cdots & \quad\cdots & \cdots
\end{array}$$

where the n^{th} row corresponds to the data generated according to our algorithm when the total sample size is n.

4.1 Asymptotic Results for the first method

Suppose that given total sample size n, we shall first generate $n_0(n)$ design points using the initial distribution ν(which could be uniform) on $\mathcal{X}$. Based on the available sample $\{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \cdots, (X_{n_0}^{(n)}, Y_{n_0}^{(n)})\}$ we construct an estimate $\hat{M_1}^{(n)}$ of M. Then remaining $n - n_0(n)$ design points will be generated as follows. We shall use $\hat{M_1}^{(n)}$ for the first $n_1(n)$ iterations, $\hat{M_2}^{(n)}$ for the next $n_2(n)$ iterations, $\ldots$ and $\hat{M_{t_n}}^{(n)}$ for the last $n_{t_n}(n)$ iterations with $T_i^{(n)} = \Delta^*(\hat{M_k}^{(n)})/(c_n \log(i + 1 - n_0)), i > n_0$ where k is such that $\sum_0^{k-1} n_s(n) < i \leq \sum_0^k n_s(n)$ and $\hat{M_k}^{(n)}$ is an estimate of M based on $\{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \cdots, (X_{n_0+n_1+\ldots+n_{k-1}}^{(n)}, Y_{n_0+n_1+\ldots+n_{k-1}}^{(n)})\}$, for $i = 1, 2, 3, \cdots, t_n$.

Theorem 4.1. *Consider the above setup. Assume that, $a \leq c_n \leq b$ for all large n for some constants $a, b \in (0, 1)$.*

a) Assume that,

$$\sup_{x \in \mathcal{X}^*} \left| \hat{M_{t_n}}^{(n)}(x) - M(x) \right| \xrightarrow{\text{P}} 0,$$

$$n - n_0(n) - n_{t_n}(n) \longrightarrow \infty$$

$$\textit{and } (n - n_0(n))^{1-b} - (n - n_0(n) - n_{t_n}(n))^{1-b} \longrightarrow \infty \textit{ as } n \longrightarrow \infty.$$

Then,

$$M\left(X_{k(n)}^{(n)}\right) \longrightarrow m^* \textit{ in probability}$$

$$\textit{and } P\left(X_{k(n)}^{(n)} \in \mathcal{X}_{opt}^*\right) \longrightarrow 1, \textit{ as } n \longrightarrow \infty.$$

b) Moreover if there exists an $s \geq 0$ such that

$$n_0(n) \longrightarrow \infty, \tag{4.9}$$

$$\frac{\sum_{i=0}^{s} n_i(n)}{n} \longrightarrow 0 \textit{ as } n \longrightarrow \infty, \tag{4.10}$$

$$\frac{1}{n}\sum_{i=s}^{t_n-1}(s_i+1)^b\left(1-\exp\left[\frac{(s_i+1)^{1-b}-(s_{i+1}+1)^{1-b}}{1-b}\right]\right) \longrightarrow 0 \tag{4.11}$$

$$\textit{and } \max_{s<i\leq t_n}\sup_{x\in\mathcal{X}^*}\left|\hat{M}_i^{(n)}(x)-M(x)\right| \xrightarrow{\mathrm{P}} 0, \tag{4.12}$$

where $s_i(n) = \sum_{k=1}^{i} n_k(n)$, we have,

$$\zeta_n(\mathcal{X}^*_{opt}) \xrightarrow{\mathrm{P}} 1 \tag{4.13}$$

where ζ_n is the empirical distribution of $\{X_1^{(n)}, X_2^{(n)}, \ldots, X_n^{(n)}\}$.

c) Moreover, if $t_n = t$ for all n for some positive integer t and there exists an $s \geq 0$ such that

$$n_0(n) \longrightarrow \infty, \tag{4.14}$$

$$\frac{\sum_{i=0}^{s} n_i(n)}{n} \longrightarrow 0 \textit{ as } n \longrightarrow \infty, \tag{4.15}$$

$$(s_{i+1}+1)^{1-b}-(s_i+1)^{1-b} \longrightarrow \infty, \forall\ i = s,\ldots,t-1 \tag{4.16}$$

$$\textit{and } \max_{s<i\leq t_n}\sup_{x\in\mathcal{X}^*}\left|\hat{M}_i^{(n)}(x)-M(x)\right| \xrightarrow{\mathrm{P}} 0 \tag{4.17}$$

then

$$\left\|\zeta_n - \mu^*_{opt}\right\| \xrightarrow{\mathrm{P}} 0. \tag{4.18}$$

Proof. The proof is given in the appendix. □

Corollary 4.2. *Recall the setup as in part (b) of Theorem (4.1). Then,*

$$C_3 \xrightarrow{\mathrm{P}} \frac{m-m^*}{m} \textit{ as } n \longrightarrow \infty. \tag{4.19}$$

4.2 Asymptotic Results for the second method

Recall the setup as in section (4.1). For a set $S \subset \mathcal{X}$, define

$$B(S,\varepsilon) = \{x \in \mathcal{X} : d(x,S) \leq \varepsilon\}.$$

Suppose that $\{\hat{M}_i^{(1)}, \hat{M}_i^{(2)}, \ldots\}$ is a sequence of estimates of M for every $1 \leq i \leq t$, where t is a fixed positive integer. Given the total sample size n, we shall first generate $n_0(n)$ design points using some initial distribution ν (which could be uniform) on $\mathcal{X}$. Then remaining $n - n_0(n) = k(n)$ design points will be generated as follows. At the k^{th} step, n_k iid points will be generated from $\mathcal{D}_k^n$ for $k = 1, 2, \ldots, t$, where,

$$S_k^n = \{x \in \mathcal{X} : \hat{M}_k^{(n)}(x) \geq (1 - c_k^n) \sup_{y \in \mathcal{X}} \hat{M}_k^{(n)}(y)\}, \tag{4.20}$$

$$U_k^n = B(S_k^n, r_k^n), \tag{4.21}$$

$$\mathcal{C}_k^n = \text{ Uniform design over } U_k^n, \tag{4.22}$$

$$\mathcal{D}_k^n = \alpha_k^n \mathcal{D}_{k-1}^n + (1 - \alpha_k^n) \mathcal{C}_k^n \tag{4.23}$$

and $\{\alpha_k^n\}, \{c_k^n\}$ and $\{r_k^n\}$ are sequences of real numbers in $(0, 1)$ such that for fixed n, they are decreasing in k and for fixed k, they converge to zero.

Theorem 4.3. *Consider the above setup.*

a) Assume that,

$$\sup_{x \in \mathcal{X}} \left| \hat{M}_t^{(n)}(x) - M(x) \right| \xrightarrow{\text{P}} 0$$
$$\text{and } n_t \longrightarrow \infty \text{ as } n \longrightarrow \infty.$$

Then,

$$M\left(X_n^{(n)}\right) \longrightarrow m \text{ in probability}$$
$$\text{and } P\left(X_n^{(n)} \in B(\mathcal{X}_{opt}, \varepsilon)\right) \longrightarrow 1, \text{ as } n \longrightarrow \infty \text{ for any } \varepsilon > 0.$$

b) Moreover if there exists an $s \in \{0, 1, 2, \ldots, t\}$ such that

$$\frac{\sum_{i=0}^{s} n_i(n)}{n} \longrightarrow 0, \tag{4.24}$$

$$n_i(n) \longrightarrow \infty, \forall\ i = s+1, \ldots, t, \tag{4.25}$$

$$\text{and } \max_{s < i \leq t} \sup_{x \in \mathcal{X}} \left| \hat{M}_i^{(n)}(x) - M(x) \right| \xrightarrow{\text{P}} 0, \text{ as } n \longrightarrow \infty. \tag{4.26}$$

We have for any $\varepsilon > 0$,

$$\zeta_n(B(\mathcal{X}_{opt}, \varepsilon)) \xrightarrow{\text{P}} 1 \tag{4.27}$$

where ζ_n is the empirical distribution of $\{X_1^{(n)}, X_2^{(n)}, \ldots, X_n^{(n)}\}$.

Proof. The proof is given in the appendix. □

Corollary 4.4. *Consider the setup as given in part* (*b*) *of theorem* (4.3)*. Then,*

$$C_3 \xrightarrow{\mathrm{P}} 0 \text{ as } n \longrightarrow \infty. \tag{4.28}$$

Chapter 5

Concluding Remarks

Our proposed methods apply to cases when the regression function M is a real-valued continuous function defined on a compact subset of some metric space under the condition that we can estimate the regression function M by an estimate which is uniformly weakly consistent. For Euclidean spaces, we can show the existence of such a sequence of estimates using Nadaraya-Watson kernel regression estimates and appropriate choice of bandwidth.

5.1 Existence of an estimate of M satisfying assumption of Theorem 4.1

Suppose that M is a smooth function on a compact subset of $\mathbb{R}^d$ for some $d > 0$. Fix $t \geq 1$. Let $\{n_i(n)\}_{n>0}$ be a sequence of positive integers $\forall\, 0 \leq i \leq t$ such that $n = \sum_{i=0}^{t} n_i(n)$ and

$$n_0(n) \longrightarrow \infty,\ \frac{n_0(n)}{n} \longrightarrow 0.$$

For example, $n_0(n) = \sqrt{n}, n_i(n) = (n - n_0(n))/t, i = 1, 2, \ldots, t$ will satisfy the above conditions. Suppose that given sample size n, we take $n_0(n)$ samples $X_1^{(n)}, X_2^{(n)}, \ldots, X_{n_0(n)}^{(n)}$ uniformly from the design space $\mathcal{X}$.

Let $\hat{M}_1$ be the Nadaraya-Watson kernel estimate of M based on the sample $\{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \cdots, (X_{n_0}^{(n)}, Y_{n_0}^{(n)})\}$ with bandwidth h_n and kernel K, where

$$h_n \longrightarrow 0,\ \frac{n_0 h_n^d}{\log n_0} \longrightarrow \infty.$$

Then under appropriate conditions on M and K, $\hat{M}_1$ is a uniformly weakly consistent estimate of M(see *e.g.,* Devroye [5], Silverman [15]). Now suppose that we use the procedure according to our proposed methods to generate the design points. At each stage, we estimate the regression function using the Nadaraya-Watson kernel estimate based on available samples with the same bandwidth h_n and kernel K. Then the $\hat{M}_i$'s for $1 \leq i \leq t$ are uniformly weakly consistent estimate of M(see Devroye [5], Silverman [15]). Hence we have one such sequence. Note that the estimates presented above are not so practical as we are keeping the bandwidth in Nadaraya-Watson kernel estimate fixed throughout the experiment. For the estimates to be weakly consistent we need that for every point in $\mathcal{X}$ we have enough data in a small neighborhood. At every stage if we use bandwidth (which may depend on the point where we are estimating the function) satisfying the above condition then we believe that that sequence will also be weakly uniformly consistent under appropriate condition, but that needs a proof.

In simulation studies and in asymptotic analysis we have seen that our proposed methods are working quite satisfactorily. Now one natural question is why two methods? We have support for both the methods.

- First note that our first proposed method is easy to implement than the second one. In the second case, we have to work with design distributions which are not standard, while in the first case the Markov chain can easily be implemented.

- Also, for the same sample size, the second method takes longer time to generate the design points than the first method. In the first method we have to estimate the regression function only at one point per sample, while in the second method we have to find the maximum of the estimate over a large set and also to find the points where the value of the estimate is near the maximum value. Also generating design points from the non-standard distributions $\mathcal{D}_k$ is far more time consuming than generating design points from the Markov chain in the first method.

- On the other hand, the second method has its own advantages. First of all as seen in the simulation and the asymptotic analysis, the criteria values(see 1.4, 1.5, 1.6) are much smaller for the second method.

- The second method takes less samples to estimate the set $\mathcal{X}_{opt}$ accurately. For a given sample size, the first method may miss more points in $\mathcal{X}_{opt}$ than the second method.

5.2 Future Works

The following questions arise in the course of the current study and they will be addressed in a future study.

1. What is the rate of convergence for our proposed methods?
2. How to choose the parameters used in our proposed algorithm optimally?
3. Given a bound on the error in estimating the set $\mathcal{X}_{opt}$ and the maximum value m, what is the minimum sample size needed to attain that bound?

Appendix A

Proof of Theorem 4.1

A.1 Proof of part (a) of Theorem 4.1

Without loss of generality, going to a subsequence, we may assume that

$$\sup_{x\in\mathcal{X}} \left|\hat{M_{t_n}}^{(n)}(x) - M(x)\right| \longrightarrow 0 \text{ a.s.}$$

Remember that, $\mathcal{X}^*$ is a finite subset with $|\mathcal{X}^*| = N$ and M is a continuous function on $\mathcal{X}$ with $|\mathcal{X}_{opt}(M)| = K < \infty$. *From now on we shall work with only $\mathcal{X}^*$. Hence for simplicity we shall omit the stars in this section. Remember that $\mathcal{X}$ is originally $\mathcal{X}^*$.* For a function g on $\mathcal{X}$, define

$$\tilde{m}(g) = m(g) - \sup\{g(x) : x \in \mathcal{X}, g(x) \neq m(g)\} \tag{A.1}$$

Recall that the first n_0 design points are generated from the fixed design distribution ν. From then onwards we are generating the design points sequentially using a markov chain where the transition matrix is random and depends on the step number. Also, note that, the transition matrix at the i-th step, $i > n_0$, is $p(\widehat{M}_k^{(n)}, T_i^{(n)})$, where $T_i^{(n)} = \Delta^*(\widehat{M}_k^{(n)}/(c_n \log(i+1-n_0))$, k is such that $\sum_{s=0}^{k-1} n_s(n) < i \leq \sum_{s=0}^{k} n_s(n)$ and $p(M,T)$ is the transition matrix given by,

$$p(M,T)(x,y) = \begin{cases} \frac{1}{N}\exp\left[-\frac{(M(x)-M(y))^+}{T}\right], & \text{if } x \neq y \\ 1 - \frac{1}{N}\sum_{z\neq x}\exp\left[-\frac{(M(x)-M(z))^+}{T}\right], & \text{if } x = y, \end{cases} \tag{A.2}$$

and the stationary distribution of $p(M,T)$ is given by the Gibbs distribution

$$\mu(M,T)(x) = \frac{\exp(M(x)/T)}{\sum_{y\in\mathcal{X}}\exp(M(y)/T)}, \quad x \in \mathcal{X}. \tag{A.3}$$

We shall show that conditional on $\{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \cdots, (X_{n-n_{t_n}(n)}^{(n)}, Y_{n-n_{t_n}(n)}^{(n)})\}$,

$$P\left(X_n^{(n)} \in \mathcal{X}_{opt}\right) \longrightarrow 1, \text{ as } n \longrightarrow \infty. \tag{A.4}$$

Then we have conditional on $\{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \cdots, (X_{n-n_{t_n}(n)}^{(n)}, Y_{n-n_{t_n}(n)}^{(n)})\}$,

$$M\left(X_n^{(n)}\right) \xrightarrow{\mathrm{P}} m$$

by the fact that

$$\begin{aligned}&\mathrm{E}\left(\left|M\left(X_n^{(n)}\right) - m\right| \Big| \{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \cdots, (X_{n-n_{t_n}(n)}^{(n)}, Y_{n-n_{t_n}(n)}^{(n)})\}\right)\\ &\leq 2mP\left(X_n^{(n)} \notin \mathcal{X}_{opt} \Big| \{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \cdots, (X_{n-n_{t_n}(n)}^{(n)}, Y_{n-n_{t_n}(n)}^{(n)})\}\right)\end{aligned}$$

Then by DCT part (a) of Therem 4.1 will follow easily.

First, we show that A.4 is true for a nonrandom sequence of estimates $\{M_1, M_2, \ldots\}$ converging to M uniformly. For simplicity, we write m_n, Δ_n, $\mu_{opt}^n, \tilde{m}_n, \mathcal{X}_{opt}^n$ in place of $m(M_n), \Delta(M_n)$, $\mu_{opt}(M_n), \tilde{m}(M_n), \mathcal{X}_{opt}(M_n)$ respectively. To prove that, we use the following lemma,

Lemma A.1 (Rate of convergence)**.** *Let $\mathcal{X}, M, p$ be as above. Assume $T_n = \Delta(M)/(c\log(n+1)), \forall\ n > 1$, for some $c \in (0, b]$ where $0 < b < 1$. Then, for any initial distribution ν on $\mathcal{X}$, $n > m > 0$,*

$$\|\nu P_m P_{m+1} \cdots P_n - \mu_{opt}\| \leq 2\exp\left[\frac{m^{1-b} - n^{1-b}}{1-b}\right] + \frac{6N}{m^{\alpha}}$$

where $P_i = p(M, T_i)$ and $0 < \alpha \leq \frac{c\tilde{m}}{\Delta}$.

Proof. The proof is given later in this section. □

Now M_n's are converging to M uniformly. Hence, $\exists\ n_0$ such that for $n > n_0$,

$$\mathcal{X}_{opt}(M_n) = \mathcal{X}_{opt}(M),\ \Delta_n \leq \Delta + \tilde{m},\ \tilde{m}_n \geq \frac{\tilde{m}}{4}, \tag{A.5}$$

and hence,

$$\frac{c_n \tilde{m}_n}{\Delta_n} \geq \frac{a\tilde{m}}{4(\tilde{m} + \Delta)} = \alpha(\text{say}).$$

Hence for any initial distribution ν on $\mathcal{X}$, we have using lemma A.1,

$$\begin{aligned}
&\left\| \nu p(M_n, T^{(n)}_{n-n_{t_n}(n)+1}) p(M_n, T^{(n)}_{n-n_{t_n}(n)+2}) \cdots p(M_n, T^{(n)}_n) - \mu_{opt} \right\| \\
&\le 2\exp\left[\frac{(n-n_0(n)-n_{t_n}(n)+1)^{1-b} - (n-n_0(n))^{1-b}}{1-b} \right] \\
&\qquad + \frac{6N}{(n-n_0(n)-n_{t_n}(n)+1)^{\alpha}},
\end{aligned}$$

which implies that,

$$\begin{aligned}
&\mathrm{E}\left(\nu p(M_n, T^{(n)}_{n-n_{t_n}(n)+1}) p(M_n, T^{(n)}_{n-n_{t_n}(n)+2}) \cdots p(M_n, T^{(n)}_n) \left(\mathcal{X}^c_{opt}\right) \right) \\
&\le 2\exp\left[\frac{(n-n_0(n)-n_{t_n}(n)+1)^{1-b} - (n-n_0(n))^{1-b}}{1-b} \right] \\
&\qquad + \frac{6N}{(n-n_0(n)-n_{t_n}(n)+1)^{\alpha}}. \qquad \text{(A.6)}
\end{aligned}$$

Under the assumption in part (a) in Therem 4.1 RHS converges to zero as $n \longrightarrow \infty$. This completes the proof for the fixed sequence case.

Now let's go back to our case. Conditioning on $\{(X^{(n)}_1, Y^{(n)}_1), (X^{(n)}_2, Y^{(n)}_2), \cdots, (X^{(n)}_{n-n_{t_n}(n)}, Y^{(n)}_{n-n_{t_n}(n)})\}$, and taking ν to be the distribution degenerate at $X^{(n)}_{n-n_{t_n}(n)}$, it follows by A.6 that conditionally

$$\begin{aligned}
&\mathrm{E}\left(\nu p(\widehat{M}^{(n)}_{t_n}, T^{(n)}_{n-n_{t_n}(n)+1}) p(\widehat{M}^{(n)}_{t_n}, T^{(n)}_{n-n_{t_n}(n)+2}) \cdots p(\widehat{M}^{(n)}_{t_n}, T^{(n)}_n) \left(\mathcal{X}^c_{opt}\right) \right) \\
&= P\left(X^{(n)}_n \notin \mathcal{X}_{opt}\right) \longrightarrow 0
\end{aligned}$$

This completes the proof of part (a) of Theorem 4.1. □

A.2 Proof of part (b) of Theorem 4.1

Take $0 < \varepsilon < \frac{m-\tilde{m}}{4}$. Define,

$$A_{n,\varepsilon} = \{ \max_{s<i\le t_n} \sup_{x\in\mathcal{X}} \left| \hat{M}_i^{(n)}(x) - M(x) \right| \le \varepsilon \}.$$

Then, by assumption in part (b) of Theorem 4.1 we have,

$$P(A_{n,\varepsilon}) \longrightarrow 0, \text{ as } n \longrightarrow \infty.$$

Define, $s_i(n) = \sum_{k=1}^{i} n_k(n)$. For $w \in A_{n,\varepsilon}$, by lemma (A.1) we have,

$$\begin{aligned}
&\left\|\delta_{X_{s_i}(w)} p(\hat{M}_{i+1}^{(n)}, s_i+1) p(\hat{M}_{i+1}^{(n)}, s_i+2) \cdots p(\hat{M}_{i+1}^{(n)}, s_i+k+1) - \mu_{opt}\right\| \\
&\qquad \leq 2\exp\left[\frac{(s_i+1)^{1-b} - (s_i+k+1)^{1-b}}{1-b}\right] + \frac{6N}{(s_i+1)^{\alpha/(1+\alpha)}}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\delta_{X_{s_i}(w)} p(\hat{M}_{i+1}^{(n)}, s_i+1) p(\hat{M}_{i+1}^{(n)}, s_i+2) \cdots p(\hat{M}_{i+1}^{(n)}, s_i+k+1)(\mathcal{X}_{opt}^c) \\
&\qquad \leq 2\exp\left[\frac{(s_i+1)^{1-b} - (s_i+k+1)^{1-b}}{1-b}\right] + \frac{6N}{(s_i+1)^{\alpha/(1+\alpha)}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbf{1}_{A_{n,\varepsilon}} \frac{1}{n} \sum_{i=s}^{t_n-1} \sum_{k=1}^{n_{i+1}} \delta_{X_{s_i}(w)} p(\hat{M}_{i+1}^{(n)}, s_i+1) p(\hat{M}_{i+1}^{(n)}, s_i+2) \cdots p(\hat{M}_{i+1}^{(n)}, s_i+k+1)(\mathcal{X}_{opt}^c) \\
&\leq \frac{2}{n} \sum_{i=s}^{t_n-1} \sum_{k=1}^{n_{i+1}} \exp\left[\frac{(s_i+1)^{1-b} - (s_i+k+1)^{1-b}}{1-b}\right] + \frac{1}{n} \sum_{i=0}^{k_n-1} \frac{6n_{i+1}N}{(s_i+1)^{\alpha/(1+\alpha)}} \\
&\leq \frac{2}{n} \sum_{i=s}^{t_n-1} (s_i+1)^b \left(1 - \exp\left[\frac{(s_i+1)^{1-b} - (s_{i+1}+1)^{1-b}}{1-b}\right]\right) + \frac{6N}{(n_0+1)^{\alpha/(1+\alpha)}}.
\end{aligned} \tag{A.7}$$

Taking expectations of LHS of (A.7) after some simplifications we have,

$$\begin{aligned}
&E\left[\frac{1}{n} \sum_{i=s}^{t_n-1} \sum_{k=1}^{n_{i+1}} \delta_{X_{s_i}(w)} p(\hat{M}_{i+1}^{(n)}, s_i+1) p(\hat{M}_{i+1}^{(n)}, s_i+2) \cdots p(\hat{M}_{i+1}^{(n)}, s_i+k+1)(\mathcal{X}_{opt}^c)\right] \\
&\leq \frac{2}{n} \sum_{i=s}^{t_n-1} (s_i+1)^b \left(1 - \exp\left[\frac{(s_i+1)^{1-b} - (s_{i+1}+1)^{1-b}}{1-b}\right]\right) + \frac{6N}{(n_0+1)^{\alpha/(1+\alpha)}} \\
&\qquad + P(A_{n,\varepsilon}^c) \\
&\longrightarrow 0 \text{ as } n \longrightarrow \infty.
\end{aligned}$$

Now note that,

$$\begin{aligned}
&\mathrm{E}\left(\zeta_n(\mathcal{X}_{opt}^c)\right) \\
&\leq E\left[\frac{1}{n} \sum_{i=s}^{t_n-1} \sum_{k=1}^{n_{i+1}} \delta_{X_{s_i}(w)} p(\hat{M}_{i+1}^{(n)}, s_i+1) p(\hat{M}_{i+1}^{(n)}, s_i+2) \cdots p(\hat{M}_{i+1}^{(n)}, s_i+k+1)(\mathcal{X}_{opt}^c)\right] \\
&\qquad + \frac{\sum_{i=0}^{s} n_i(n)}{n}.
\end{aligned}$$

This completes the proof of part (b) of Theorem (4.1). □

A.3 Proof of part (c) of Theorem 4.1

Under the assumptions of part (c) in Theorem 4.1, the conditions of part (b) in Theorem 4.1 can be easily verified. Hence, we have

$$\zeta_n((\mathcal{X}_{opt})^c) \xrightarrow{\mathrm{P}} 0 \tag{A.8}$$

Let $\zeta_n^{(i)}$ be the empirical distribution of $\{X^{(n)}_{n_0+\cdots+n_{i-1}+1}, X^{(n)}_{n_0+\cdots+n_{i-1}+2}, \ldots, X^{(n)}_{n_1+\cdots+n_i}\}$, for $i = 0, 1, \ldots, t$.

Then, for any $x \in \mathcal{X}$,

$$\begin{aligned} |\zeta_n(x) - \mu_{opt}(x)| &= \left| \frac{1}{n} \sum_{i=0}^{n} \mathbf{1}_{(X_i^{(n)}=x)} - \mu_{opt}(x) \right| \\ &\leq \frac{\sum_{i=0}^{s} n_i(n)}{n} + \sum_{i=s+1}^{t} \frac{n_i(n)}{n} \left| \zeta_n^i(x) - \mu_{opt}(x) \right| \end{aligned} \tag{A.9}$$

By assumption in part (c) in Therem (4.1),

$$\frac{\sum_{i=0}^{s} n_i(n)}{n} \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

Therefore, it is enough to show that for each $i \in \{s+1, \ldots, t\}, x \in \mathcal{X}_{opt}$,

$$\frac{n_i(n)}{n} \left| \zeta_n^i(x) - \mu_{opt}(x) \right| \xrightarrow{\mathrm{P}} 0 \text{ as } n \longrightarrow \infty.$$

We shall show that for each $i \in \{s+1, \ldots, t\}, x \in \mathcal{X}_{opt}$, conditional on $\{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \cdots, (X^{(n)}_{n_0+\cdots+n_{i-1}}, Y^{(n)}_{n_0+\cdots+n_{i-1}})\}$,

$$\mathrm{E}\left[\frac{n_i(n)}{n} \left| \zeta_n^i(x) - \mu_{opt}(x) \right| \right]^2 \longrightarrow 0, \text{ as } n \longrightarrow \infty. \tag{A.10}$$

Fix $i \in \{s+1, \ldots, t\}$. By our assumption $(n_1+\cdots+n_i+1)^{1-b} - (n_1+\cdots+n_{i-1}+1)^{1-b} \longrightarrow \infty$, where $c_n \leq b < 1 \;\forall\; n$. Hence (A.10) follows using the the following lemma (A.2), conditional on $\{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \cdots, (X^{(n)}_{n_0+\cdots+n_{i-1}}, Y^{(n)}_{n_0+\cdots+n_{i-1}})\}$ with $s_n = n_1 + \cdots + n_{i-1}$ and $a_n = n_i$. □

Lemma A.2. *Let $\{M_1, M_2, \ldots\}$ be a nonrandom sequence of estimates of M converging to M uniformly. Let $n = n_0 + s_n + a_n \;\forall\; n$, where $n_0 \longrightarrow \infty$ and $(s_n + a_n)^{1-b} - s_n^{1-b} \longrightarrow \infty$ as $n \longrightarrow \infty$ where $0 < b < 1$. For* *fixed n, let $\{X_1^{(n)}, \cdots, X_{a_n}^{(n)}\}$ be the nonhomogeneous Markov chain with initial distribution ν and transition matrix $P_i^{(n)} = p(M_n, T_i^n)$, where $T_i^n = \Delta(M_n)/(c_n \log(i + s_n))$, $0 < a \leq c_n \leq b\ \forall\ n$ and ζ_n be the empirical distribution of $\{X_1^{(n)}, \cdots, X_{a_n}^{(n)}\}$. Then for $x \in \mathcal{X}_{opt}$,*

$$\left(\frac{a_n}{n}\right)^2 \mathrm{E}\left(\zeta_n(x) - \mu_{opt}(x)\right)^2 \longrightarrow 0, \text{ as } n \longrightarrow \infty. \tag{A.11}$$

Proof. Fix $x \in \mathcal{X}_{opt}$. Note that

$$\begin{aligned}
&\mathrm{E}\left[\frac{a_n}{n}\left|\zeta_n(x) - \mu_{opt}(x)\right|\right]^2 = \frac{1}{n^2} E\left[\left(\sum_{i=1}^{a_n}[\mathbf{1}_{X_i^{(n)}=x} - \mu_{opt}(x)]\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{a_n}\sum_{j=1}^{a_n}\left[P(X_i^{(n)} = x, X_j^{(n)} = x) - 2\mu_{opt}(x)\mu_i^{(n)}(x) + \mu_{opt}(x)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{a_n}\sum_{j=1}^{a_n}\left[P(X_i^{(n)} = x, X_j^{(n)} = x) - \mu_{opt}(x)^2\right] + \frac{2}{n}\sum_{i=1}^{a_n}[\mu_i^{(n)}(x) - \mu_{opt}(x)] \\
&\leq \frac{2}{n^2}\sum_{i<j}\left|P(X_i^{(n)} = x, X_j^{(n)} = x) - \mu_{opt}(x)^2\right| + \frac{4}{n}\sum_{i=1}^{a_n}\left\|\mu_i^{(n)} - \mu_{opt}\right\|,
\end{aligned} \tag{A.12}$$

where $\mu_i^{(n)}(x) = P(X_i^{(n)} = x)$.

Now for $1 \leq i < j \leq a_n$,

$$\begin{aligned}
&\left|P(X_i^{(n)} = x, X_j^{(n)} = x) - \mu_{opt}(x)^2\right| \\
&= \left|\mu_i^{(n)}(x) \cdot \delta_x P_i^{(n)} P_{i+1}^{(n)} \cdots P_j^{(n)} - \mu_{opt}(x)^2\right| \\
&\leq \left|\delta_x P_i^{(n)} P_{i+1}^{(n)} \cdots P_j^{(n)} - \mu_{opt}(x)\right| + \left|\mu_i^{(n)}(x) - \mu_{opt}(x)\right|.
\end{aligned} \tag{A.13}$$

Define,

$$a_{ij}^{(n)} = \delta_x P_i^{(n)} P_{i+1}^{(n)} \cdots P_j^{(n)} - \mu_{opt}(x), \forall\ 1 \leq i < j \leq a_n.$$

Hence from (A.12), we have

$$\begin{aligned}
&\mathrm{E}\left[\frac{a_n}{n}\left|\zeta_n(x) - \mu_{opt}(x)\right|\right]^2 \\
&\leq \frac{2}{n^2}\sum_{i<j}\left|a_{ij}^{(n)}\right| + \frac{6}{n}\sum_{i=1}^{a_n}\left\|\mu_i^{(n)} - \mu_{opt}\right\|.
\end{aligned} \tag{A.14}$$

Now, by lemma (A.1), the second term in the RHS of (A.14) goes to zero as $n \longrightarrow \infty$. To show that the first term also goes to zero as $n \longrightarrow \infty$, note that for $i < j$, by lemma (A.1), uniform convergence of $\{M_n\}$ and the fact that $T_i^n = \Delta(M_n)/(c_n \log(i + s_n))$, we have for large enough n,

$$\begin{aligned} \left|a_{ij}^{(n)}\right| &\le \left\|\delta_x P_i^{(n)} P_{i+1}^{(n)} \cdots P_j^{(n)} - \mu_{opt}\right\| \\ &\le 2\exp\left[\frac{(s_n+i)^{1-b} - (s_n+j)^{1-b}}{1-b}\right] + \frac{6N}{i^\alpha} \end{aligned} \tag{A.15}$$

where $\alpha = a\tilde{m}/(4(\tilde{m} + \Delta))$.

Hence

$$\frac{2}{n^2} \sum_{1\le i<j\le a_n} \left|a_{ij}^{(n)}\right| \le \frac{4}{n^2} \sum_{1\le i<j\le a_n} \exp\left[\frac{(s_n+i)^{1-b} - (s_n+j)^{1-b}}{1-b}\right] + \frac{12N}{n} \sum_{i=1}^{a_n} \frac{1}{i^\alpha}$$

Now,

$$\begin{aligned} &\frac{1}{n^2} \sum_{1\le i<j\le a_n} \exp\left[\frac{(s_n+i)^{1-b} - (s_n+j)^{1-b}}{1-b}\right] \\ &= \frac{1}{n^2} \sum_{1\le i\le a_n} \exp\left[\frac{(s_n+i)^{1-b}}{1-b}\right] \sum_{i<j\le a_n} \exp\left[\frac{-(s_n+j)^{1-b}}{1-b}\right] \\ &\le \frac{1}{n^2} \sum_{1\le i\le a_n} \exp\left[\frac{(s_n+i)^{1-b}}{1-b}\right] \int_{s_n+i}^{s_n+a_n} \exp\left[\frac{-x^{1-b}}{1-b}\right]\, dx \\ &\le \frac{1}{n^2} \sum_{1\le i\le a_n} \exp\left[\frac{(s_n+i)^{1-b}}{1-b}\right] (s_n+a_n)^b \left(\exp\left[-\frac{(s_n+i)^{1-b}}{1-b}\right] - \exp\left[-\frac{(s_n+a_n)^{1-b}}{1-b}\right]\right) \\ &\le \frac{n^b}{n^2} \sum_{1\le i\le a_n} \left(1 - \exp\left[\frac{(s_n+i)^{1-b} - (s_n+a_n)^{1-b}}{1-b}\right]\right) \\ &\le \frac{n\cdot n^b}{n^2} + \frac{n^b}{n^2} \exp\left[\frac{-(s_n+a_n)^{1-b}}{1-b}\right] \sum_{1\le i\le a_n} \exp\left[\frac{(s_n+i)^{1-b}}{1-b}\right] \\ &\le \frac{2}{n^{1-b}} + \frac{n^b}{n^2} \exp\left[\frac{-(s_n+a_n)^{1-b}}{1-b}\right] \sum_{1\le i<a_n} \exp\left[\frac{(s_n+i)^{1-b}}{1-b}\right] \\ &\le \frac{2}{n^{1-b}} + \frac{n^b}{n^2} \exp\left[\frac{-(s_n+a_n)^{1-b}}{1-b}\right] \int_{s_n}^{s_n+a_n} \exp\left[\frac{x^{1-b}}{1-b}\right]\, dx \\ &\le \frac{2}{n^{1-b}} + \frac{n^b}{n^2} \exp\left[\frac{-(s_n+a_n)^{1-b}}{1-b}\right] (s_n+a_n)^b \left(\exp\left[\frac{(s_n+a_n)^{1-b}}{1-b}\right] - \exp\left[\frac{s_n^{1-b}}{1-b}\right]\right) \end{aligned}$$

$$\leq \frac{2}{n^{1-b}} + \frac{n^{2b}}{n^2}\left(1 - \exp\left[\frac{s_n^{1-b} - (s_n + a_n)^{1-b}}{1-b}\right]\right) \longrightarrow 0, \text{ as } n \longrightarrow \infty,$$

by the fact that $0 < b < 1$ and our assumption that $(s_n + a_n)^{1-b} - s_n^{1-b} \longrightarrow \infty$ as $n \longrightarrow \infty$. So we have, $n^{-2} \sum_{i<j} \left| a_{ij}^{(n)} \right| \longrightarrow 0$ as $n \longrightarrow \infty$. This completes the proof. □

A.4 Proof of lemma (A.1)

Recall that $P_i = p(M, T_i)$. Define $\mu_i = \mu(M, T_i)$, the stationary distribution corresponding to the homogeneous Markov chain with transition matrix P_i. Note that, using results from standard Markov chain theory (see Winkler [23]), we have,

$$\begin{aligned}
&\|\nu P_m P_{m+1} \cdots P_n - \mu_{opt}\| \\
&\leq \|\nu P_m P_{m+1} \cdots P_n - \mu_m P_m \cdots P_n\| + \|\mu_m P_m \cdots P_n - \mu_{opt}\| \\
&\leq 2 \prod_{k=m}^{n} c(P_k) + \|\mu_m P_m \cdots P_n - \mu_{opt}\|,
\end{aligned}$$

where $c(P) = \frac{1}{2} \sup_{x,y \in \mathcal{X}} \|P(x,.) - P(y,.)\|$.
Now,

$$\begin{aligned}
\prod_{k=m}^{n} c(P_k) \leq \prod_{k=m}^{n} (1 - \frac{1}{k^c}) &= \exp\left[\sum_{k=m}^{n} log(1 - \frac{1}{k^c})\right] \\
&\leq \exp\left[-\sum_{k=m}^{n} \frac{1}{k^c}\right] \leq \exp\left[-\int_m^n \frac{dx}{x^c}\right] \\
&= \exp\left[\frac{m^{1-c} - n^{1-c}}{1-c}\right].
\end{aligned}$$

and by a result on simulated annealing (see Winkler [23]), for $T_i = \frac{\Delta}{c \log i}$,

$$\|\mu_i P_i \cdots P_n - \mu_{opt}\| \leq \frac{6N}{i^{c\tilde{m}/\Delta}} \leq \frac{6N}{i^\alpha}, \tag{A.16}$$

where $0 < \alpha \leq \frac{c\tilde{m}}{\Delta}$.

Hence the lemma is proved. □

A.5 Proof of corollary (4.2)

Recall that,

$$\begin{aligned}
m &= \sup_{x \in \mathcal{X}} M(x), \\
m^* &= \sup_{x \in \mathcal{X}^*} M(x), \\
C_3 &= \frac{1}{m}\left[m - \frac{1}{n}\sum_{i=1}^{n} M(X_i^{(n)})\right] = 1 - \frac{1}{nm}\sum_{i=1}^{n} M(X_i^{(n)}).
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{1}{nm}\sum_{i=1}^{n} M(X_i^{(n)}) &= \frac{m^*}{m} \cdot \zeta_n(\boldsymbol{\mathcal{X}}_{opt}^*) + \frac{1}{nm}\sum_{i: X_i^{(n)} \notin \mathcal{X}_{opt}^*} M(X_i^{(n)}) \\
&\xrightarrow{\text{P}} \frac{m^*}{m}, \qquad \text{(A.17)}
\end{aligned}$$

as

$$\left|\frac{1}{n}\sum_{i: X_i^{(n)} \notin \mathcal{X}_{opt}^*} M(X_i^{(n)})\right| \leq m(1 - \zeta_n(\boldsymbol{\mathcal{X}}_{opt}^*)) \text{ and } \zeta_n(\boldsymbol{\mathcal{X}}_{opt}^*) \xrightarrow{\text{P}} 1.$$

Hence, $C_3 \xrightarrow{\text{P}} \frac{m-m^*}{m}.$ □

Appendix B

Proof of Theorem 4.3

Recall that M is a smooth function on $\mathcal{X}$. For a set $S \subset \mathcal{X}$, define the closed ε ball around S

$$B(S, \varepsilon) = \{x \in \mathcal{X} : d(x, S) \leq \varepsilon\}.$$

Suppose that $\{\hat{M}_i^{(1)}, \hat{M}_i^{(2)}, \ldots\}$ is a sequence of estimates of M for every $1 \leq i \leq t$, where t is a fixed positive integer. Let $\{\alpha_k^n\}, \{c_k^n\}$ and $\{r_k^n\}$ be sequences of real numbers in $(0, 1)$ such that for fixed n, they are decreasing in k and for fixed k, they converge to zero. Given total sample size n, we shall first generate $n_0(n)$ design points using some initial distribution ν(which could be uniform distribution) on $\mathcal{X}$. Then remaining $n - n_0(n) = k(n)$ design points will be generated as follows. At the k^{th} step, n_k i.i.d. points will be generated from $\mathcal{D}_k^n$ for $k = 1, 2, \ldots, t$, where

$$\begin{aligned} S_k^n &= \{x \in \mathcal{X} : \hat{M}_k^{(n)}(x) \geq (1 - c_k^n) \sup_{y \in \mathcal{X}} \hat{M}_k^{(n)}(y)\}, & \text{(B.1)} \\ U_k^n &= B(S_k^n, r_k^n), & \text{(B.2)} \\ \mathcal{C}_k^n &= \text{ Uniform design over } U_k^n & \text{(B.3)} \\ \text{and } \mathcal{D}_k^n &= \alpha_k^n \mathcal{D}_0 + (1 - \alpha_k^n) \mathcal{C}_k^n. & \text{(B.4)} \end{aligned}$$

B.1 Proof of part (a) of Theorem 4.3

Without loss of generality, going to a subsequence, we may assume that

$$\sup_{x \in \mathcal{X}} \left| \widehat{M}_t^{(n)}(x) - M(x) \right| \longrightarrow 0 \text{ a.s.}$$

Note that it is enough to prove that for any $\varepsilon > 0$,

$$P\left(X_n^{(n)} \in B(\mathcal{X}_{opt}, \varepsilon)\right) \longrightarrow 1. \tag{B.5}$$

We shall show that conditional on $\{X_1^{(n)}, X_2^{(n)}, \ldots, X_{n-n_t(n)}^{(n)}\}$,

$$P\left(X_n^{(n)} \in B(\mathcal{X}_{opt}, \varepsilon)\right) \longrightarrow 1, \text{ as } n \longrightarrow \infty. \tag{B.6}$$

Then, by DCT (B.5), follows easily. We use the following lemma to prove (B.6) and the proof of it ill be given later.

Lemma B.1. *Suppose that $\{M_1, M_2, \ldots\}$ is a sequence of functions on $\mathcal{X}$ converging to M uniformly,* i.e.,,

$$\sup_{x \in \mathcal{X}} |M(x) - M_n(x)| \longrightarrow 0, \text{ as } n \longrightarrow \infty. \tag{B.7}$$

Let $\{\alpha_n\}, \{c_n\}$ and $\{r_n\}$ be sequences of real numbers in $(0,1)$ decreasing to zero. Define

$$S_n = \{x \in \mathcal{X} : M_n(x) \geq (1 - c_n) m_n\}, \ n = 1, 2, \ldots \tag{B.8}$$

$$S = \mathcal{X}_{opt}(M) \tag{B.9}$$

Then for any $\varepsilon > 0, \exists\ N > 0$ such that

$$B(S_n, r_n) \subseteq B(S, \varepsilon), \forall\ n \geq N. \tag{B.10}$$

Recall that conditional on $\{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \ldots, (X_{n-n_t(n)}^{(n)}, Y_{n-n_t(n)}^{(n)})\}$, $X_n^{(n)} \sim \mathcal{D}_t^n = \alpha_t^n \mathcal{D}_{t-1}^n + (1 - \alpha_t^n)\mathcal{C}_t^n$. Therefore, conditionally

$$\begin{aligned} P(X_n^{(n)} \in B(\mathcal{X}_{opt}, \varepsilon)^c) &\leq \alpha_t^n + (1 - \alpha_t^n)\mathcal{C}_t^n(B(\mathcal{X}_{opt}, \varepsilon)^c)) \\ &\leq \alpha_t^n + \mathcal{C}_t^n(B(S_t^n, r_t^n)^c)) \text{ for large } n \\ &= \alpha_t^n \\ &\longrightarrow 0 \text{ as } n \longrightarrow \infty. \end{aligned} \tag{B.11}$$

This completes the proof. □

B.2 Proof of part (b) of Theorem 4.3

Let $\zeta_n^{(i)}$ be the empirical distribution of $\{X_{n_0+\cdots+n_{i-1}+1}^{(n)}, X_{n_0+\cdots+n_{i-1}+2}^{(n)}, \ldots, X_{n_0+n_1+\cdots+n_i}^{(n)}\}$ for $i = 0, 1, \ldots, t$.

So we have

$$
\begin{aligned}
|\zeta_n(B(\mathcal{X}_{opt},\varepsilon))-1| &= \zeta_n(B(\mathcal{X}_{opt},\varepsilon)^c) \\
&= \frac{1}{n}\sum_{i=0}^{n}\mathbf{1}_{(X_i^{(n)}\in B(\mathcal{X}_{opt},\varepsilon)^c)} \\
&\leq \frac{\sum_{i=0}^{s} n_i(n)}{n} + \sum_{i=s+1}^{t}\frac{n_i(n)}{n}\zeta_n^i(B(\mathcal{X}_{opt},\varepsilon)^c) \\
&\leq \frac{\sum_{i=1}^{s} n_i(n)}{n} + \sum_{i=s+1}^{t}\zeta_n^i(B(\mathcal{X}_{opt},\varepsilon)^c) \qquad \text{(B.12)}
\end{aligned}
$$

Therefore, it is enough to show that for $i = s+1, s+2, \ldots, t$,

$$E\left(\zeta_n^i(B(\mathcal{X}_{opt},\varepsilon)^c)\right) \xrightarrow{\text{P}} 0. \qquad \text{(B.13)}$$

Fix $i \in \{s+1, s+2, \ldots, t\}$. Conditional on $\{(X_1^{(n)}, Y_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}), \ldots, (X_{n_0+n_1+\cdots+n_{i-1}}^{(n)}, Y_{n_0+n_1+\cdots+n_{i-1}}^{(n)})\}$,

$$X_{n_0+\cdots+n_{i-1}+j}^{(n)} \overset{i.i.d.}{\sim} \mathcal{D}_i^n,\ j = 1, 2, \ldots, n_i.$$

Hence, by (B.11),

$$
\begin{aligned}
&E\left(\zeta_n^i(B(\mathcal{X}_{opt},\varepsilon)^c)|(X_1^{(n)}, Y_1^{(n)}), \ldots, (X_{n_0+n_1+\cdots+n_{i-1}}^{(n)}, Y_{n_0+n_1+\cdots+n_{i-1}}^{(n)})\right) \\
&= \mathcal{D}_i^n(B(\mathcal{X}_{opt},\varepsilon)^c) \\
&\longrightarrow 0 \text{ as } n \longrightarrow \infty.
\end{aligned}
$$

B.3 Proof of Lemma B.1

Fix $\varepsilon > 0$. There exists $\eta > 0$ such that

$$\{x \in \mathcal{X} : M(x) \geq m - \eta\} \subseteq B(S, \varepsilon/2). \qquad \text{(B.14)}$$

Now, choose $N > 0$ such that

$$
\begin{aligned}
\sup_{x\in\mathcal{X}} |M(x) - M_n(x)| &\leq \frac{\eta}{4}, \qquad &\text{(B.15)} \\
\left|c_n(m - \frac{\eta}{4})\right| &\leq \frac{\eta}{4} \qquad &\text{(B.16)} \\
\text{and } r_n &\leq \frac{\varepsilon}{2}, \forall\ n \geq N. \qquad &\text{(B.17)}
\end{aligned}
$$

Then, for $n \geq N, x \in S_n$, we have

$$M(x) \geq M_n(x) - \frac{\eta}{4} \geq (1-c_n)m_n - \frac{\eta}{4} \geq (1-c_n)(m - \frac{\eta}{4}) - \frac{\eta}{4} \geq m - \eta.$$

Hence, for $n \geq N$,

$$\begin{aligned} B(S_n, r_n) &= B(\{x \in \mathcal{X} : M_n(x) \geq (1-c_n)m_n\}, r_n) \\ &\subseteq B(\{x \in \mathcal{X} : M_n(x) \geq m - \eta\}, r_n) \\ &\subseteq B(B(S, \frac{\varepsilon}{2}), \frac{\varepsilon}{2}) \\ &\subseteq B(S, \varepsilon). \qquad \text{(B.18)} \end{aligned}$$

This completes the proof of lemma B.1. □

B.4 Proof of corollary (4.4)

Recall that,

$$m = \sup_{x \in \mathcal{X}} M(x) \text{ and } C_3 = 1 - \frac{1}{nm} \sum_{i=1}^{n} M(X_i^{(n)}).$$

Fix $\varepsilon \in (0,1)$. Since $M(.)$ is continuous and has finitely many maxima, *i.e.,* $|\mathcal{X}_{opt}| < \infty$, there exists $\delta > 0$ such that $d(x, \mathcal{X}_{opt}) \leq \delta$ implies $|M(x)/m - 1| < \varepsilon/2$. Now,

$$\begin{aligned} |C_3| &= \left| \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{M(X_i^{(n)})}{m} \right) \right| \\ &\leq \left| \frac{1}{n} \sum_{i: X_i^{(n)} \in B(\mathcal{X}_{opt}, \delta)} \left(1 - \frac{M(X_i^{(n)})}{m} \right) \right| + \left| \frac{1}{n} \sum_{i: X_i^{(n)} \notin B(\mathcal{X}_{opt}, \delta)} \left(1 - \frac{M(X_i^{(n)})}{m} \right) \right| \\ &\leq \frac{\varepsilon}{2} \zeta_n(B(\mathcal{X}_{opt}, \delta)) + \zeta_n(B(\mathcal{X}_{opt}, \delta)^c) \\ &= \frac{\varepsilon}{2} + (1 - \frac{\varepsilon}{2}) \zeta_n(B(\mathcal{X}_{opt}, \delta)^c). \end{aligned}$$

Hence,

$$\begin{aligned} P(|C_3| > \varepsilon) &\leq P\left(\frac{\varepsilon}{2} + (1 - \frac{\varepsilon}{2}) \zeta_n(B(\mathcal{X}_{opt}, \delta)^c) > \varepsilon \right) \\ &= P\left(\zeta_n(B(\mathcal{X}_{opt}, \delta)^c) > \frac{\varepsilon}{2 - \varepsilon} \right) \\ &\longrightarrow 0 \text{ as } n \longrightarrow \infty. \end{aligned}$$

This completes the proof of corollary (4.4). □

Bibliography

[1] Catoni O., Rough Large Deviation Estimates for Simulated Annealing: Application to exponential Schedules, *The Annals of Probability* **20**, (1992), 1109-1146.

[2] Chen H., Huang M. N. L. and Huang W. J., Estimation of the location of the maximum of a regression function using extreme order statistics, *Journal Of Multivariate Analysis* **57** (1996), 191–214.

[3] Chen H., Lower rate of convergence for locating a maximum of a function, *Annals of Statistics* **16** (1988), 1330–1334.

[4] Deuschel J. D., Mazza C., L^2 Convergence of Time Nonhomogeneous Markov Process: I. Special Estimates, *Annals of Applied Probability* **4** (1994), 1012–1056.

[5] Devroye L. P., The Uniform Covergence of the Nadaraya-Watson Regression Function Estimate, *Canadian Journal of Statistics* **6** (1978), 179–191.

[6] Dobrushin R. L., Central Limit Theorem for Non-Stationary Markov Chains I, II., *Theo. Prob. Appl.*, **1**, (1956), 65-80, 329-383.

[7] Draper N. R., Smith H., *Applied regression analysis*, 2nd edition, Wiley Publication, 1981.

[8] Fan J., Gijbels I., *Local Polynomial Modelling And its Applications*, Chapman & Hall New York, 1996.

[9] Fedorov V. V., *Theory of Optimal Experiments*, Academic Press, New York, 1972.

[10] Geman D. and Geman S., Stachastic Relaxation, Gibbs distributions, and the bayesian restoration of images, *IEEE Trans. PAMI*, **6**, (1984), 721-741.

[11] Hotelling H., Experimental determination of the maximum of a function, *Annals of Mathematical Statistics* **12** (1941), 20–45.

[12] Kiefer J. and Wolfowitz J., Stochastic estimation of the maximum of a regression function, *Annals of Mathematical Statistics* **23** (1952), 462–466.

[13] Kirkpatrick S., Gelatt C. D. Jr., Vecchi M.P., *Optimization by simulated annealing*, IBM T.J. Watson Research Center, Yorktown Heights, New York, 1982.

[14] Kirkpatrick S., Gelatt C. D. Jr., Vecchi M.P., Optimization by simulated annealing, *Science* **220**, (1982), 671-680.

[15] Mack Y. P., Silverman B. W., Weak and Strong Uniform Consistency of Kernel Regression Estimate, *Zeitschrift für Wahrscheinlichkeitsteorie und verwandte Gebiete* **51** (1982), 405–415.

[16] McCullagh P., Nelder J. A., *Generalised linear models*, 2nd edition, Chapman & Hall, New York, 1992.

[17] Müller H. G., Adaptive nonparametric peak estimation, *Annals of Statistics* **17** (1989), 1053-1069.

[18] Müller H. G., Optimal designs for nonparametric kernel regression, *Statistics And Probability Letters* **2** (1984), 285–290.

[19] Sacks J., Asymptotic distribution of stochastic approximation procedures, *Annals of Mathematical Statistics* **28** (1957), 373–405.

[20] Stone. C. J., Optimal Rates Of Convergence For Nonparametric Estimators, *Annals of Statistics* **8** (1980), 1348–1360.

[21] Stone. C. J., Optimal Global Rates Of Convergence For Nonparametric Regression, *Annals of Statistics* **10** (1982), 1040–1053.

[22] Wand M. P., Jones M. C., *Kernel Smoothing*, Chapman & Hall, New York, 1995.

[23] Winkler, G., *Image analysis, random fields and dynamic Monte Carlo methods: a mathematical introduction*, 2nd edition, Springer Verlag, Berlin, 1993.