stats

MDPI

Article

An Upper Bound of the Bias of Nadaraya–Watson Kernel Regression under Lipschitz Assumptions

Samuele Tosatto [1,*], Riad Akrour [1] and Jan Peters [1,2]

[1] Computer Science Department, Technische Universität Darmstadt, 64289 Darmstadt, Germany; riad.akrour@tu-darmstadt.de (R.A.); mail@jan-peters.net (J.P.)
[2] Computer Science Department, Max Planck Institute for Intelligent Systems, 70569 Stuttgart, Germany
* Correspondence: samuele.tosatto@tu-darmstadt.de

Abstract: The Nadaraya–Watson kernel estimator is among the most popular nonparameteric regression technique thanks to its simplicity. Its asymptotic bias has been studied by Rosenblatt in 1969 and has been reported in several related literature. However, given its asymptotic nature, it gives no access to a hard bound. The increasing popularity of predictive tools for automated decision-making surges the need for hard (non-probabilistic) guarantees. To alleviate this issue, we propose an upper bound of the bias which holds for finite bandwidths using Lipschitz assumptions and mitigating some of the prerequisites of Rosenblatt's analysis. Our bound has potential applications in fields like surgical robots or self-driving cars, where some hard guarantees on the prediction-error are needed.

Keywords: nonparametric regression; Nadaraya-Watson kernel regression; bias

Citation: Tosatto, S.; Akrour, R.; Peters, J. An Upper Bound of the Bias of Nadaraya–Watson Kernel Regression under Lipschitz Assumptions. *Stats* **2021**, *4*, 1–17. https://doi.org/

Received: 30 October 2020
Accepted: 25 December 2020
Published: 30 December 2020

1. Introduction

Nonparametric regression and density estimation have been used in a wide spectrum of applications, ranging from economics [1], system dynamics identification [2,3], and reinforcement learning [4–7]. In recent years, nonparametric density estimation and regression have been dominated by parametric methods such as those based on deep neural networks. These parametric methods have demonstrated an extraordinary capacity in dealing with both high-dimensional data—such as images, sounds, or videos—and large datasets. However, it is difficult to obtain strong guarantees on such complex models, which have been shown easy to fool [8]. Nonparametric techniques have the advantage of being easier to understand, and recent work overcame some of their limitations by, e.g., allowing linear-memory and sub-linear query time for density kernel estimation [9,10]. These methods allowed nonparametric kernel density estimation to be performed on datasets of 10^6 samples and up to 784 input dimension. As such, nonparametric methods are a suitable choice when one is willing to trade performance for statistical guarantees; and the contribution of this paper is to advance the state-of-the-art on such guarantees.

Studying the error of a statistical estimator is important. It can be used, for example, to tune the hyper-parameters by minimizing the estimated error [11–14]. To this end, the estimation error is usually decomposed into an estimation bias and variance. When it is not possible to derive these quantities, one performs an asymptotic behavior analysis or a convergence to a probabilistic distribution of the error. While all aforementioned analyses give interesting insights on the error and allow for hyper-parameter optimization, they do not provide any strong guarantee on the error, i.e., we cannot upper bound it with absolute certainty.

Beyond hyper-parameter optimization, we argue that another critical aspect of error analysis is to provide hard (non-probabilistic) bounds of the error for critical data-driven algorithms. We believe that learning agents taking autonomous, data-driven, decisions will be increasingly present in the near future. These agents will, for example, be autonomous surgeons, self-driving cars or autonomous manipulators. In many critical applications

involving these agents, it is of primary importance to bound the prediction error in order to provide some technical guarantees on the agent's behavior. In this paper we derive a hard upper bound of the estimation bias in non-parametric regression with minimal assumptions on the problem. The bound can be readily applied to a wide range of applications.

Specifically, we consider the Nadaraya–Watson kernel regression [15,16], which can be seen as a conditional kernel density estimate. We derive an upper bound of the estimation bias under weak local Lipschitz assumptions. The reason for our choice of estimator falls in its inherent simplicity compared to more sophisticated techniques. The bias of the Nadaraya–Watson kernel regression has been previously studied by [17], and has been reported in a number of related work [18–21]. The analysis of the bias conducted by Rosenblatt (1969) [17] still remains the main reference for this regression technique. The main assumptions of Rosenblatt's analysis are $h_n \to 0$ (where n is the number of samples) and $nh_n \to \infty$ where h_n is the kernel's bandwidth. Rosenblatt's analysis suffers from an asymptotic error $o(h_n^2)$, which means that for large bandwidths it is not accurate; making it inapplicable to derive a hard upper bound. To the best of our knowledge, the only proposed bound on the bias requires the restrictive assumption that the samples must be placed evenly on a closed interval [22]. In contrast, we derive an upper bound of the bias of the Nadaraya–Watson kernel regression that is valid for a large class of design and for any choice of bandwidth.

We build our analysis on weak Lipschitz assumptions [23], which are milder than the (global) Lipschitz, as we require only $|f(x) - f(y)| \leq L|x - y| \ \forall y \in \mathcal{C}$ given a fixed x, instead of the classic $|f(x) - f(y)| \leq L|x - y| \ \forall y, x \in \mathcal{C}$—where $\mathcal{C}$ is the data domain. Lipschitz assumptions are common in different fields, and usually allow a wide family of admissible functions. This is particularly true when the Lipschitz is require for only a subset of the function's domain (like in our case). Moreover, notice that the classical analysis requires the knowledge of the second derivative of the regression function m, and therefore the continuity of m'. Our Lipschitz assumption is less restrictive, allowing us to obtain a bias upper bound even for functions like $m(x) = |x|$, at points (like $x = 0$) where m'' is undefined. The Rosenblatt analysis builds on a Taylor expansion of the estimator and therefore when the bandwidth h_n is large, Rosenblatt's bias analysis tends to provide wrong estimates of the bias, as observed in the experimental section. We consider multidimensional input space, and we apply the bound to a realistic regression problem.

2. Preliminaries

Consider the problem of estimating $\mathbb{E}[Y|X = \mathbf{x}]$ where $X \sim f_X$ and $Y = m(X) + \epsilon$, with noise ϵ, i.e., $\mathbb{E}[\epsilon] = 0$. The noise can depend on $\mathbf{x}$, but since our analysis is conducted point-wise for a given $\mathbf{x}$, $\epsilon_{\mathbf{x}}$ will be simply denoted by ϵ. Let $m : \mathbb{R}^d \to \mathbb{R}$ be the regression function and f_X a probability distribution on X called design. In our analysis we consider $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. The Nadaraya–Watson kernel estimate of $\mathbb{E}[Y|X = \mathbf{x}]$ is

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n} K_{\mathbf{h}}(\mathbf{x} - \mathbf{x}_i) y_i}{\sum_{j=1}^{n} K_{\mathbf{h}}(\mathbf{x} - \mathbf{x}_j)}, \tag{1}$$

where $K_{\mathbf{h}}$ is a kernel function with bandwidth-vector $\mathbf{h}$, the $\mathbf{x}_i$ are drawn from the design f_X and y_i from $m(\mathbf{x}_i) + \epsilon$. Note that both the numerator and the denominator are proportional to Parzen-Rosenblatt density kernel estimates [24,25]. We are interested in the point-wise bias of such estimate $\mathbb{E}[\hat{m}(\mathbf{x})] - m(\mathbf{x})$. In the prior analysis of [17], knowledge of m', m'', f_X, f_X' is required and f, m' must be continuous in a neighborhood of x. In addition, and as discussed in the introduction, the analysis is limited to a one-dimensional design and an infinitesimal bandwidth. We briefly present the classical bias analysis of [17] before introducing our results for clarity of exposition.

Theorem 1. Classic Bias Estimation [17]. *Let $m : \mathbb{R} \to \mathbb{R}$ be twice differentiable. Assume a set $\{x_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i$ are i.i.d. samples from a distribution with non-zero differentiable density f_X.*

Assume $y_i = m(\mathbf{x}_i) + \epsilon_i$, with noise $\epsilon_i \sim \varepsilon(\mathbf{x}_i)$. The bias of the Nadaraya–Watson kernel in the limit of infinite samples and for $h \to 0$ and $nh_n \to \infty$ is

$$\begin{aligned}\mathbb{E}\Big[\lim_{n\to\infty} \hat{m}_n(x)\Big] - m(x) &= h_n^2 \left(\frac{1}{2} m''(x) + \frac{m'(x) f_X'(x)}{f_X(x)}\right) \int u^2 K(u)\, \mathrm{d}u + o_P\left(h_n^2\right) \\ &\approx h_n^2 \left(\frac{1}{2} m''(x) + \frac{m'(x) f_X'(x)}{f_X(x)}\right) \int u^2 K(u)\, \mathrm{d}u. \end{aligned} \tag{2}$$

Note that Equation (2) must be normalized with $\int_{-\infty}^{\infty} k(u)\, \mathrm{d}u$ when the kernel function does not integrate to one. The o_P term denotes the asymptotic behavior w.r.t. the bandwidth. Therefore, for a larger value of the bandwidth, the bias estimation becomes worse, as is illustrated in Figure 1.

3. Main Result

In this section, we present two bounds on the bias of the Nadaraya–Watson estimator. The first one considers a bounded regression function m (i.e., $|m(x)| \leq M$), and allows for weak Lipschitz conditions on a subset of the design's support. Instead, the second bound does not require the regression function to be bounded but only the weak Lipschitz continuity to hold on all of its support. The definition of "weak" Lipschitz continuity will be given below.

To develop our bound on the bias for multidimensional inputs is essential to define some subset of the $\mathbb{R}^d$ space. In more detail, we consider an open n-dimensional interval in $\mathbb{R}^d$ which is defined as $\Omega(\boldsymbol{\tau}^-, \boldsymbol{\tau}^+) \equiv (\tau_1^-, \tau_1^+) \times \cdots \times (\tau_d^-, \tau_d^+)$ where $\tau^-, \tau^+ \in \overline{\mathbb{R}}^d$. We now formalize what is meant by weak (log-)Lipschitz continuity. This will prove useful as we need knowledge of the weak-Lipschitz constants of m and $\log f_X$ in our analysis.

Definition 1. *Weak Lipschitz continuity at $\mathbf{x}$ on the set $\mathcal{C}$ under the L_1-norm.*
Let $\mathcal{C} \subseteq \mathbb{R}^d$ and $f : \mathcal{C} \to \mathbb{R}$. We call f weak Lipschitz continuous at $\mathbf{x} \in \mathcal{C}$ if and only if

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L|\mathbf{x} - \mathbf{y}| \quad \forall \mathbf{y} \in \mathcal{C},$$

where $|\cdot|$ denotes the L_1-norm.

Definition 2. *Weak log-Lipschitz continuity at $\mathbf{x}$ on the set $\mathcal{C}$ under the L_1-norm.*
Let $\mathcal{C} \subseteq \mathbb{R}^d$. We call f weak log-Lipschitz continuous at $\mathbf{x}$ on the set $\mathcal{C}$ if and only if

$$|\log f(\mathbf{x}) - \log f(\mathbf{y})| \leq L|\mathbf{x} - \mathbf{y}| \quad \forall \mathbf{y} \in \mathcal{C}.$$

Note that the set $\mathcal{C}$ can be a subset of the function's domain.

It is important to note that, in contrast to the global Lipschitz continuity, which requires $|f(\mathbf{y}) - f(\mathbf{z})| \leq L|\mathbf{y} - \mathbf{z}|\ \forall \mathbf{y}, \mathbf{z} \in \mathcal{C}$, the weak Lipschitz continuity is defined at a specific point $\mathbf{x}$ and therefore allows the function to be discontinuous elsewhere. The Lipschitz assumptions are not very restrictive, and in practice require a bounded gradient. They have been widely used in various fields. Note that when the Lipschitz constants are not known, they can be estimated from the dataset [26]. In the following we list the set of assumptions that we use in our theorems.

A1. f_X and m are defined on $\mathrm{Y} \equiv \Omega(\mathbf{x} - v^-, \mathbf{x} + v^+)$ and $v^-, v^+ \in \overline{\mathbb{R}}_+^d$;
A2. f_X is log weak Lipschitz with constant L_f at $\mathbf{x}$ on the set $\mathcal{D} \equiv \Omega(\mathbf{x} - \boldsymbol{\delta}^-, \mathbf{x} + \boldsymbol{\delta}^-) \subseteq \mathrm{Y}$ and $f_X(\mathbf{x}) \geq f_X(\mathbf{z})\ \forall \mathbf{z} \in \mathrm{Y} \backslash \mathcal{D}$ with positive defined $\delta^-, \delta^+ \in \overline{\mathbb{R}}_+^d$ (note that this implies $f_X(\mathbf{y}) > 0\ \forall \mathbf{y} \in \mathcal{D}$);
A3. m is weak Lipschitz with constant L_m at $\mathbf{x}$ on a the set $\mathcal{G} \equiv \Omega(\mathbf{x} - \boldsymbol{\gamma}^-, \mathbf{x} + \boldsymbol{\gamma}^+) \subseteq \mathcal{D}$ with positive defined $\boldsymbol{\gamma}^-, \boldsymbol{\gamma}^+ \in \overline{\mathbb{R}}_+^d$.

To work out a bound on the bias valid for a wide class of kernels, we must enumerate some assumption and quantify some integrals with respect to the kernel.

A4. The multidimensional kernel $K_{\mathbf{h}} : \mathbb{R}^d \to \mathbb{R}$ can be decomposed in a product of independent uni-dimensional kernels, i.e., $K_{\mathbf{h}}(\mathbf{x}) = \prod_{i=1}^{d} k_i(x/h_i)$ with $k_i : \mathbb{R} \to \mathbb{R}$;

A5. the kernels are non-negative $k_i(x) \geq 0$ and symmetric $k_i(x) = k_i(-x)$;

A6. for every $a, x \in \mathbb{R}$ and $h \neq 0$, the integrals $\int_0^a k_i(x)\,\mathrm{d}x = \Phi_i(a)$, $\int_0^a k_i(x)e^{-xL_f}\,\mathrm{d}x = B_i(x, L_f)$, $\int_0^a k_i(x)xe^{-xL_f}\,\mathrm{d}x = C_i(x, L_f)$ are finite (i.e., $< +\infty$).

Assumptions A4–A6 are not really restrictive, and includes any kernel with both finite domain and co-domain, or not heavy-tailed (e.g., Gaussian-like). Furthermore, Axiom A4 allows any independent composition of different kernel functions. In Appendix B we detail the integrals of Axiom A6 for different kernels. Note that when the integrals listed in A6 exist in closed form, the computation of the bound is straightforward, and requires negligible computational effort.

In the following, we propose two different bounds of the bias. The first version considers a bounded regression function ($M < +\infty$), this allows both the regression function and the design to be weak Lipschitz on a subset of their domain. In the second version instead, we consider the case of an unbounded regression function ($M = +\infty$) or an unknown bound M. In this case both the regression function and the design must be weak Lipschitz on the entire domain Y.

Theorem 2. *Bound on the Bias with Bounded Regression Function.*

Assuming A1–A3, $\mathbf{h} \in \mathbb{R}_+^d$ a positive defined vector of bandwidths $\mathbf{h} = [h_1, h_2, \ldots, h_n]^\intercal$, K_h the multivariate kernel defined in A4–A6, $\hat{m}_n(\mathbf{x})$ the Nadaraya–Watson kernel estimate using n observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$ with $x_i \sim f_X$, $y_i = m(\mathbf{x}_i) + \epsilon_i$ and with noise $\epsilon_i \sim \varepsilon(\mathbf{x}_i)$ centered in zero ($\mathbb{E}[\varepsilon(\mathbf{x}_i)] = 0$), $n \to \infty$, and furthermore assuming there is a constant $0 \leq M < +\infty$ such that $|m(\mathbf{y}) - m(\mathbf{z})| \leq M\ \forall \mathbf{y}, \mathbf{z} \in \mathrm{Y}$, the considered Nadaraya–Watson kernel regression bias is bounded by

$$\left| \mathbb{E}\left[\lim_{n\to\infty} \hat{m}_n(\mathbf{x}) \right] - m(\mathbf{x}) \right| \leq \frac{L_m \sum_{k=1}^{d} \xi_k(\phi_k^-, \phi_k^+) \prod_{i\neq k}^{d} \zeta(\phi_i^-, \phi_i^+) + M\left(\prod_{i=1}^{d} \zeta(\gamma_i^-, \gamma_i^+) - \prod_{i=1}^{d} \zeta(\phi_i^-, \phi_i^+) + D \right)}{\prod_{i=1}^{d} \psi_i(\delta_i^-, \delta_i^+)}$$

where

$$\psi_i(a,b) = h_i\Big(B_i(b/h_i, L_f h_i) - B_i(-a/h_i, -L_f h_i)\Big), \zeta_i(a,b) = h_i\Big(B_i(b/h_i, -L_f h_i) - B_i(-a/h_i, L_f h_i)\Big),$$

$$\xi_i(a,b) = h_i^2\Big(C_i(b/h_i, -L_f h_i) + C_i(-ah_i, L_f h_i)\Big), D = \lim_{\omega\to+\infty} \prod_{i=1}^{d} h_i \Phi_i\left(\frac{\omega}{h_i}\right) - \prod_{i=1}^{d} h_i \varphi_i,$$

with $\varphi_i = \Phi_i(\gamma_i^+/h_i) + \Phi_i(\gamma_i^-/h_i)$, and $0 < \phi_i^- \leq \gamma_i^-, 0 < \phi_i^+ \leq \gamma_i^+$ can be freely chosen to obtain a tighter bound. We suggest $\phi_i^+ = \min(\gamma_i^+, M/L_m), \phi_i^- = \min(\gamma_i^-, M/L_m)$.

In the case where M is unknown or infinite, we propose the following bound.

Theorem 3. *Bound on the Bias with Unbounded Regression Function.*

Assuming A1–A3, $\mathbf{h} \in \overline{\mathbb{R}}_+^d$ a positive defined vector of bandwidths $\mathbf{h} = [h_1, h_2, \ldots, h_n]^\intercal$, K_h the multivariate kernel defined in A4–A6, $\hat{m}_n(\mathbf{x})$ the Nadaraya–Watson kernel estimate using n observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$ with $x_i \sim f_X$, $y_i = m(\mathbf{x}_i) + \epsilon_i$ and with noise $\epsilon_i \sim \varepsilon(\mathbf{x}_i)$ centered in zero ($\mathbb{E}[\varepsilon(\mathbf{x}_i)] = 0$), $n \to \infty$, and furthermore assuming that $\mathrm{Y} \equiv \mathcal{D} \equiv \mathcal{G}$, the considered Nadaraya–Watson kernel regression bias is bounded by

$$\left| \mathbb{E}\left[\lim_{n\to\infty} \hat{m}_n(\mathbf{x}) \right] - m(\mathbf{x}) \right| \leq \frac{L_m \sum_{k=1}^{d} \xi_k(v_k^-, v_k^+) \prod_{i\neq k}^{d} \zeta_i(v_i^-, v_i^+)}{\prod_{i=1}^{d} \psi_i(v_i^-, v_i^+)},$$

where ξ_k, ζ_i, ψ_i are defined as in Theorem 2.

We detail the proof of both theorems in Appendix A. Note that the conditions required by our theorems are mild, and they allow a wide range of random designs, including and not limited to Gaussian, Cauchy, Pareto, Uniform, and Laplace distributions. In general, every continuously differentiable density distribution is also weak log-Lipschitz in some closed subset of its domain. For example, the Gaussian distribution does not have a finite Lipschitz constant on its entire domain, but there is a finite weak Lipschitz constant on any closed interval. Examples of distributions that are weak log-Lipschitz are presented in Table 1.

Table 1. Examples of parameters to use for different univariate random design.

Distribution	Density	$\mathbf{Y}$	$\mathcal{D}$	L_f
Laplace(μ, λ)	$\frac{1}{2\lambda} \exp\left(-\frac{\lvert x-\mu\rvert}{\lambda}\right)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	λ^{-1}
Cauchy$(\mu; \gamma)$	$\left(\pi\gamma + \pi\frac{(x-\mu)^2}{\gamma}\right)^{-1}$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	$\frac{2(z-\mu)}{\gamma^2+(z-\mu)^2}1$
Uniform(a, b)	$\begin{cases} \frac{1}{b-a} & \text{if} \quad a \le x \le b \\ 0 & \text{otherwise} \end{cases}$	(a, b)	(a, b)	0
Pareto(α)	$\begin{cases} \frac{\alpha}{x^{\alpha+1}} & \text{if} x \ge 1 \\ 0 & \text{otherwise} \end{cases}$	$(1, +\infty)$	$(1, +\infty)$	$1 + \alpha$
Normal(μ, σ)	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$	$(-\infty, +\infty)$	(a, b)	$f_{\mu,\sigma}(a, b)$

4. Analysis

Although the bound applies to different kernel functions, in the following we analyze the most common Gaussian kernel. It worth noting that for $k(x) = e^{-x^2}$,

$$\phi(x) = \frac{\sqrt{\pi}}{2}\operatorname{erf}(a), \quad B(a, c) = \frac{\sqrt{\pi}}{2} e^{\frac{c^2}{4}} \left(\operatorname{erf}\left(a + \frac{c}{2}\right) - \operatorname{erf}\left(\frac{c}{2}\right)\right),$$
$$C(a, c) = \frac{1}{2}\left(1 - e^{-a(a+c)} - cB(a, c)\right).$$

Note that we removed the subscripts from the functions ψ, B, and C, as we consider only the Gaussian kernel. To provide a tight bound, we consider many quantities that describe the design's domain, the Lipschitz constants of the design and of the regression function, the bound of the image of the regression function, and the different bandwidths for each dimension of the space. This complexity results in an effective but poorly readable bound. In this section, we try to simplify the problem and to analyze the behavior of the bound in the limit.

Asymptotic Analysis: Let us consider, for the moment, the case of one-dimension ($d = 1$) and infinite domains and co-domains (M unknown and $v^- = v^+ = \delta- = \delta+ = \gamma^- = \gamma^+ = \infty$). In this particular case, the bound becomes

$$\left|\mathbb{E}\left[\lim_{n\to\infty} \hat{m}_n(\mathbf{x})\right] - m(\mathbf{x})\right| \le L_m h \left(\frac{1}{2\sqrt{\pi}\exp\frac{L_f^2 h^2}{4}} + hL_f\left(\operatorname{erf}\left(\frac{hL_f}{2}\right) + 1\right)\right) = A_1.$$

As expected, for $h \to 0$ or for $L_m = 0$, $B_1 = 0$. This result is in line with (2) (since $L_m = 0 \implies m' = 0, m'' = 0$). A completely flat design, e.g., uniformly distributed, does not imply a zero bias. This can be seen either in Rosenblatt's analysis or by just considering the fact that, notoriously, the Nadaraya–Watson estimator suffers from the boundary bias. When we analyze our bound, we find in fact that $\lim_{L_f\to 0} B_1 \propto L_m h$. It is

also interesting to analyze the asymptotic behavior when these quantities tend to infinity. Similarly to (2), we observe that A_1 grows quadratically w.r.t. the kernel's bandwidth h and it scales linearly w.r.t. the Lipschitz constant of the regression function (which is linked to m'). A further analysis brings us to the consideration that Rosenbatt's analysis is linear w.r.t. $\mathrm{d}/\,\mathrm{d}x \log f_X(x)$ (since $f'_X(x)/f_X(x) = \mathrm{d}/\,\mathrm{d}x \log f_X(x)$). Our bound has a similar implication, as the Log-Lipschitz constant is also related to the derivative of the logarithm of the design function, and $A_1 = \mathcal{O}(L_f)$.

Boundary Bias: The Nadaraya–Watson kernel estimator is affected by the boundary bias. The boundary bias is an additive bias term affecting the estimation in the region close to the boundaries of the design's domains. Since in our framework, we can consider a closed domain of the design, we can also see what is happening close to the border. Let us consider still a one-dimensional regression, but this time $v^- \to 0$, $v^- = \delta^- = \gamma^-$ and $v^+ = \delta^+ = \gamma^+ = \infty$. In this case, we obtain

$$\left| \mathbb{E}\left[\lim_{n\to\infty} \hat{m}_n(\mathbf{x}) \right] - m(\mathbf{x}) \right| \leq \frac{L_m h}{\sqrt{\pi} e^{\frac{L_f^2 h^2}{4}} \left(1 - \operatorname{erf}\left(\frac{hL_f}{2}\right)\right)} + \frac{L_m L_f h^2 \left(1 + \operatorname{erf}\left(\frac{hL_f}{2}\right)\right)}{2 - 2\operatorname{erf}\left(\frac{hL_f}{2}\right)} = A_2.$$

Keeping in account that $\mathrm{d}/\mathrm{d}x \operatorname{erf}(x) \propto e^{-x}$ and using L'Hôpital's rule, we can observe that the bound is now exponential w.r.t. h and L_f, i.e., $A_2 = \mathcal{O}(e^{hL_f})$, which implies that it is more "sensible" to higher bandwidths or less smooth design. Interestingly, instead, the bounds maintains its linear relation w.r.t. L_m.

Dimensionality: Let us now study the multidimensional case, supposing that each dimension has same bandwidth and same values for the boundaries. In this case,

$$\left| \mathbb{E}\left[\lim_{n\to\infty} \hat{m}_n(\mathbf{x}) \right] - m(\mathbf{x}) \right| \leq \frac{d\xi(v^-, v^+)\zeta(v^-, v^+)^{d-1}}{\psi(v^-, v^+)^d} \propto d\left(\frac{\zeta(v^-, v^+)}{\psi(v^-, v^+)}\right)^{d-1}.$$

Therefore, the bound scales exponentially w.r.t. the dimension. We observe an exponential behavior when $\mathbf{x}$ is close to the boundary of the design's domain. In these regions, in fact the ratio $\zeta(v^-, v^+)/\psi(v^-, v^+)$ is particularly high. Of course, when the aforementioned ratio tends to one, the linearity w.r.t. d is predominant.

We can conclude the analysis by noticing that our bound has similar limiting behavior with the Rosenblatt's analysis, but it provides a hard bound on the bias.

5. Numerical Simulation

In this section, we provide three numerical analyses of our bounds on the bias (the code of our numerical simulations can be found at http://github.com/SamuelePolimi/UpperboundNWBias). The first analysis of our method is conducted on uni-dimensional input spaces for display purposes and aims to show the properties of our bounds in different scenarios. The second analysis aims instead at testing the behavior of our method on a multidimensional input space. The third analysis emulates a realistic scenario where our bound can be applied.

Uni-dimensional Analysis: We select a set of regression functions with different Lipschitz constants and different bounds,

- $y = \sin(5x)$; $L_m = 5$ and $M = 1$,
- $y = \log x$ which for $\mathcal{G} \equiv \Omega(-1, +\infty)$ has $L_m = 1$ and $M = +\infty$,
- $y = 60^{-1} \log\cosh 60x$ which has $L_m = 1$, is unbuounded, and has a particularly high second derivative in $x = 0$, with $m''(0) = 60$,
- $y = \sqrt{x^2 + 1}$ which has $L_m = 1$ and is unbounded.

A zero-mean Gaussian noise with standard deviation $\sigma = 0.05$ has been added to the output y. Our theory applies to a wide family of kernels. In this analysis we consider a

Gaussian kernel, with $k(x) = e^{-x^2}$, a box kernel, with $k(x) = I(x)$, and a triangle kernel, with $k(x) = I(x)(1 - |x|)$ with

$$I(x) = \begin{cases} 1 & \text{if} |x| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

We further analyze the aforementioned kernels in Appendix B. In order to provide as many different scenarios as possible we also used the distributions from Table 1, using therefore both infinite domain distributions, such as Cauchy and Laplace, and finite domain such as Uniform. In order to numerically estimate the bias, we approximate $E[\hat{m}_n(x)]$ with an ensemble of estimates $N^{-1} \sum_{j=1}^{N} \hat{m}_{n,j}(x)$ where each estimate $\hat{m}_{n,j}$ is built on a different dataset (drawn from the same distribution f_X). In order to "simulate" $n \to \infty$ we used $n = 10^5$ samples, and to obtain high confidence of the bias' estimate, we used $N = 100$ models.

In this section we provide some simulations of our bound presented in Theorems 2 and 3, and for the Rosenblatt's case we use

$$\left| h_n^2 \left(\frac{1}{2} m''(x) + \frac{m'(x) f_X'(x)}{f_X(x)} \right) \int u^2 K(u) \, \mathrm{d}u \right|.$$

Since the Rosenblatt's bias estimate is not an upper bound, it can happen that the true bias is higher (as well as lower) than this estimate, as it is possible to see in Figure 1. We presented different scenarios, both with bounded and unbounded functions, infinite and finite design domains, and a larger or smaller bandwidth choice. It is possible to observe that, thanks to the knowledge of f, f', m', m'' the Rosenblatt's estimation of the bias tends to be more accurate than our bound. However, it can happen that it largely overestimates the bias, like in the case of $m(x) = 60^{-1} \log \cosh(60x)$ in $x = 0$ or to underestimate it, most often in boundary regions. In contrast, our bound always overestimates the true bias, and despite its lack of knowledge of f, f', m', m'', it is most often tight. Moreover, when the bandwidth is small, both our method and Rosenblatt's deliver an accurate estimation of the bias. In general, Rosenblatt tends to deliver a better estimate of the bias, but it does not behave as a bound, and in some situations, it also can deliver larger mispredictions. In detail, the plot (a) in Figure 1 shows that with a tight bandwidth, both our method and Rosenblatt's method achieve good approximations of the bias, but only our method correctly upper bounds the bias. When increasing the bandwidth, we obtain both a larger bias and subsequent larger estimates of the bias. Our method consistently upper bounds the bias, while in many cases, Rosenblatt's method underestimates it, especially in proximity of boundaries (subplots b, d, e). An interesting case can be observed in subplot (c), where we test the function $m(x) = 60^{-1} \log \cosh(60x)$, which has a high second-order derivative in $x = 0$: in this case, Rosenblatt's method largely overestimates the bias.

The figure shows that our bound can deal with different functions and random designs, being reasonably tight, if compared to Rosenblatt's estimation, which requires the knowledge of the regression function and the design, and respective derivatives.

Multidimensional Analysis: We want to study if our bounds work in a multidimensional case and how much it overestimates the true bias (therefore, how tight it is). For this purpose, we took a linear function $m(\mathbf{x}) = \mathbf{1}^\intercal \mathbf{x}$ where $\mathbf{1}$ is a column-vector of d ones. This function, for any dimension d, has a Lipschitz constant $L_m = 1$ and is unbounded ($M = \infty$). We set a Gaussian design with zero mean and unit diagonal covariance. Since in higher dimensions, the estimation's variance grows exponentially [22], we used a large number of samples ($n = 10^6$), and we averaged over $N = 10^5$ independent estimations. In Figure 2 we show how the "true" bias (estimated numerically averaging over a thousand Nadaraya–Watson regressions) and our bound evolve with a growing number of dimensions d. Far from the low-density region $\mathbf{x} = \mathbf{0}$ we notice that the bias tends to have a linear behavior, while close to the boundary the bias tends to be exponential. We can observe that our bound correctly bounds the bias in all the cases.

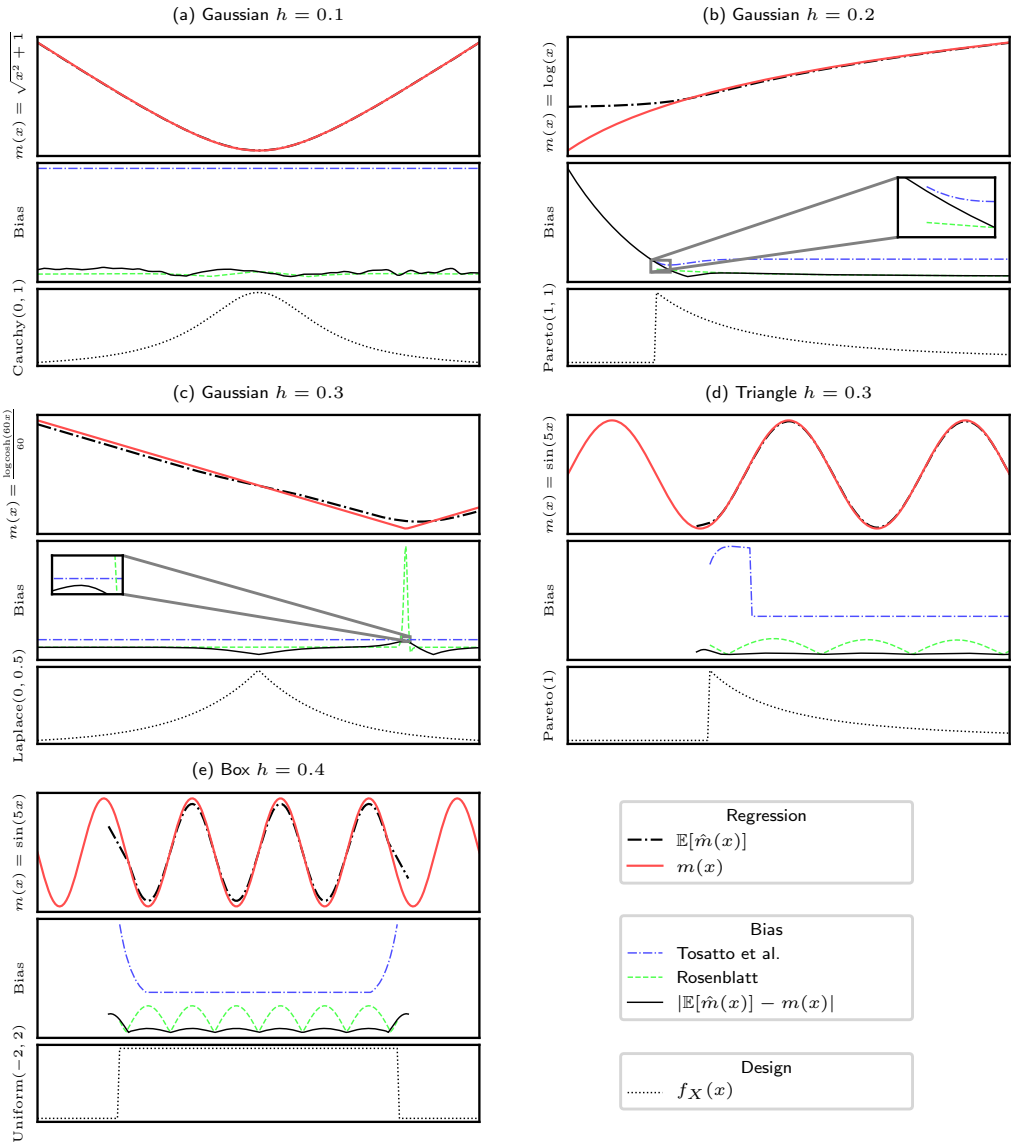

Figure 1. We propose some simulations of Nadaraya–Watson regression with different designs, regression functions, and bandwidths. (**a**)–(**c**) use Gaussian kernels, while (**d**) and (**e**) use Triangle and Box kernels, respectively. The regression function $m(x)$ is represented with a solid line, while the Nadaraya–Watson estimate $\hat{m}(x)$ is represented with a dash-dotted line in the top subplot of each experiment. In the second subplots, it is possible to observe the true bias (solid line), as well as our upper bound (dashed line) and Rosenblatt's estimate (dash-dotted line). The bottom subplots depict the design used. The bandwidth used for the estimation is denoted with h. It is possible to observe that Rosenblatt's estimate often under or overestimates the bias , e.g., subfigures (**b**) and (**c**). In all the different test conditions, our method correctly upper bounds the bias.

Realistic Scenario: Let us consider the regression problem of the dynamics of an under-actuated pendulum of length l and mass m. In particular, the state of the pendulum can be described by its angle α and its angular rotation $\dot{\alpha}$. Furthermore, a force u can be applied to the pendulum. The full system is described by,

$$\ddot{\alpha} = \frac{3}{ml^2}u - \frac{3g}{2l}\sin(\alpha + \pi),$$

where g is the gravitational acceleration. In practice, when this model is discretized in time, the next state is estimated via numerical integration. Fort this reason, the Lipschitz

constant L_m is unknown. Notice that also m' and m'' required by the Rosenblatt's analysis are unknown. We estimated the Lipschitz constant L_m by selecting the highest ratio $|y_i - y_j|/|\mathbf{x}_i - \mathbf{x}_i|$ in the dataset. In our analysis, we want to predict all the states with fixed $\dot{\alpha} = 0$, $u = 0$, but variable $\alpha \in [-\pi, \pi]$. In order to generate the dataset, we use the simulator provided by gym [27]. To train our models, we generate tuples of $\alpha, \dot{\alpha}, u$ by sampling independently each variable from a uniform distribution i.e., $\alpha \sim \text{Uniform}(-\pi, \pi)$, $\dot{\alpha} \sim \text{Uniform}(-8, 8)$ and $u \sim \text{Uniform}(-2, 2)$ (hence, $L_f = 0$). We fit 100 different models with 50,000 samples. We choose a Gaussian kernel with bandwidth $\mathbf{h} = [0.2, 0.2, 0.2]$. Figure 3 depicts our bound and the estimated bias. We notice the bias is low and increases close to the boundaries. Our upper bound is tight, but, as expected, it becomes overly pessimistic in the boundary region.

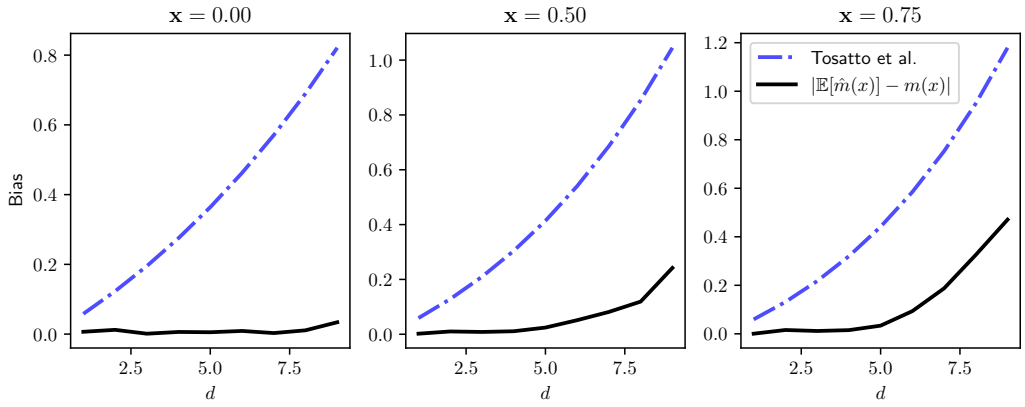

Figure 2. We propose a study on how the bias varies w.r.t. the dimension. While the bias grows almost linearly in the high density region (i.e., $\mathbf{x} = 0$), it tends to grow exponentially in lower densities (i.e., $\mathbf{x} = 0.75$). In both cases, our bound perform correctly.

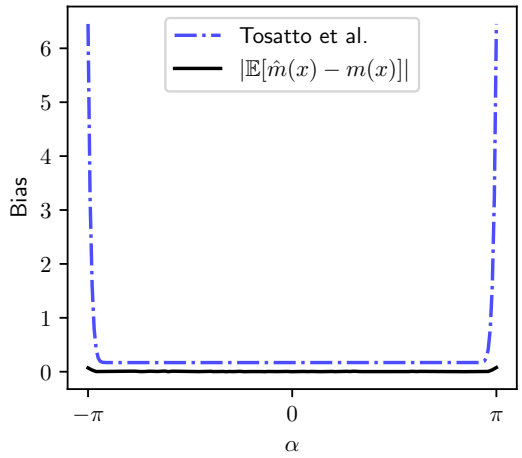

Figure 3. Experiment on the under-actuated pendulum. Our bound gives the possibility to ensure that $\mathbb{E}[m(x) - \hat{m}(x)]$ does not exceed a certain quantity. In this example, it is possible to observe that the bias is correctly bounded.

6. Applications

Our bound can be applied alongside any use of the Nadaraya–Watson regression estimate. In this section, we want to point out an interesting application in reinforcement learning. In reinforcement learning, we aim to approximate, from samples, an optimal behavior of an agent acting in a given environment. To do so, one can find a so-called "value-function" that determines how well the agent behaves. This function can be approximated using dynamic programming combined with functional approximators such as neural networks [28], but using also regression-trees [29], Gaussian processes [4], etc. In our

prior work [30], we used the presented bound to subsequently bound the bias of the value function estimated via Nadaraya–Watson regression.

7. Conclusions

The Nadaraya–Watson kernel regression is one of the most well-known nonparametric estimator, used in a wide range of applications. Its asymptotic bias and variance are well-known in the literature. However, to the best of our knowledge, such an estimator's bias has never been bounded before. In this paper, we proposed a hard bound of the bias that requires mild assumptions. Our proposed bound is numerically tight and accurate for a large class of regression functions, kernels, and random designs. We believe that providing hard, non-probabilistic guarantees on a regression error is an essential step in adopting data-driven algorithms in real-world applications. Our future research will focus on extending the bias analysis to a broader class of kernel functions and with a finite-samples analysis.

Author Contributions: S.T. conceived of the presented idea, developed the theory and performed the numerical simulations. R.A. verified the analytical method. Both S.T. and R.A. wrote the paper. All authors provided critical feedback and helped shape the research, analysis and manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: Authors are grateful to three anonymous referees and the editor for their valuable comments and suggestions, which certainly improved the presentation and quality of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Theorems Derivation

In order to provide proofs of the stated Theorems, we need to introduce some quantities and to state some facts that will be used in our proofs.

Proposition A1. *With $a \in \mathbb{R}$, $h \neq 0$ and $c \in \mathbb{R}$,*

$$\int_0^a k_i\Big(\frac{x}{h}\Big) e^{-xL_f}\,\mathrm{d}x = B_i(a/h, ch). \tag{A1}$$

Proposition A2. *With $a, b \geq 0$, $h_i > 0$ and $L_f \geq 0$,*

$$\int_{-a}^b k_i\left(\frac{x}{h_i}\right) e^{-|x|L_f}\,\mathrm{d}x = h_i B_i(b/h_i, L_f h_i) - h_i B_i(-a/h, -L_f h_i) = \psi_i(a, b). \tag{A2}$$

Proposition A3. *With $a, b \geq 0$, $h_i > 0$ and $L_f \geq 0$,*

$$\int_{-a}^b k_i\left(\frac{x}{h_i}\right) e^{|x|L_f}\,\mathrm{d}x = h_i B_i(b/h_i, -L_f h_i L_f) - h_i B_i(-a/h_i, L_f h_i) = \zeta_i(a, b). \tag{A3}$$

Proposition A4. *With $a \in \mathbb{R}$, $h \neq 0$ and $c \in \mathbb{R}$,*

$$\int_0^a k_i\Big(\frac{x}{h}\Big) e^{-xc} x\,\mathrm{d}x = h_i^2 (C_i(a/h, ch)). \tag{A4}$$

Proposition A5. *With $a, b \geq 0$, $h_i > 0$ and $L_f \geq 0$,*

$$\int_{-a}^b k\left(\frac{x}{h_i}\right) e^{|x|L_f} |x|\,\mathrm{d}x = h_i^2 \Big(C_i(b/h, L_f h_i) + C_i(-a/h_i, L_f h_i)\Big) = \xi_i(a, b). \tag{A5}$$

Definition A1. *Integral on a d-interval*

Let $\mathcal{C} \equiv \Omega(\boldsymbol{\tau}^-, \boldsymbol{\tau}^+)$ with $\boldsymbol{\tau}^-, \boldsymbol{\tau}^+ \in \overline{\mathbb{R}}^d$. Let the integral of a function $f : \mathcal{C} \to \mathbb{R}$ defined on $\mathcal{C}$ be defined as

$$\int_{\mathcal{C}} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\tau_1^-}^{\tau_1^+} \int_{\tau_2^-}^{\tau_2^+} \cdots \int_{\tau_d^-}^{\tau_d^+} f([x_1, x_2, \ldots, x_d]^\intercal)\,\mathrm{d}x_d \ldots \mathrm{d}x_2\,\mathrm{d}x_1.$$

Proposition A6. *There is a function $g : \mathrm{Y} \to \mathbb{R}$ such that*

$$f_X(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{\int_{\mathrm{Y}} e^{g(\mathbf{x})}\,\mathrm{d}\mathbf{x}}$$

and, given A2, $|g(\mathbf{x}) - g(\mathbf{y})| \leq L_f|\mathbf{x} - \mathbf{y}| \quad \forall \mathbf{y} \in \mathcal{D}$.

Proposition A7. *Independent Factorization*
Let $\mathcal{C} \equiv \Omega(\boldsymbol{\tau}^-, \boldsymbol{\tau}^+)$ where $\boldsymbol{\tau}^-, \boldsymbol{\tau}^+ \in \mathbb{R}^d$, and $f_i : \mathbb{R} \to \mathbb{R}$,

$$\int_{\mathcal{C}} \prod_{i=1}^{d} f_i(x_i)\,\mathrm{d}\mathbf{x} = \prod_{i=1}^{d} \int_{\mathcal{C}} f_i(x_i)\,\mathrm{d}\mathbf{x}.$$

Proposition A8. *Given $\mathcal{C} \equiv \Omega(\boldsymbol{\tau}^-, \boldsymbol{\tau}^+)$, $p : \mathbb{R} \to \mathbb{R}$, $q : \mathbb{R} \to \mathbb{R}$,*

$$\int_{\mathcal{C}} \left(\prod_{i=1}^{d} p(z_i)\right)\left(\sum_{k=1}^{d} g(z_k)\right)\mathrm{d}\mathbf{z} = \sum_{k=1}^{d} \left(\prod_{i\neq k}^{d} \int_{\tau_i^-}^{\tau_i^+} p(z)\,\mathrm{d}z\right) \int_{\tau_k^-}^{\tau_k^+} p(z)q(z)\,\mathrm{d}z.$$

Proof. Proof of Theorem 2:

$$\begin{aligned}
& \left| \mathbb{E}\left[\lim_{n\to\infty} \hat{f}_n(\mathbf{x})\right] - m(\mathbf{x}) \right| \\
= & \left| \mathbb{E}\left[\lim_{n\to\infty} \frac{\sum_{i=1}^n K_{\mathbf{h}}(\mathbf{x} - \mathbf{x}_i) y_i}{\sum_{j=1}^n K_{\mathbf{h}}(\mathbf{x} - \mathbf{x}_j)}\right] - m(\mathbf{x}) \right| \\
= & \left| \mathbb{E}\left[\lim_{n\to\infty} \frac{n^{-1}\sum_{i=1}^n K_{\mathbf{h}}(\mathbf{x} - \mathbf{x}_i) y_i}{n^{-1}\sum_{j=1}^n K_{\mathbf{h}}(\mathbf{x} - \mathbf{x}_j)}\right] - m(\mathbf{x}) \right| \\
= & \left| \mathbb{E}\left[\frac{\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z})\big(m(\mathbf{z}) - \epsilon(\mathbf{z})\big) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z}}{\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z}}\right] - m(\mathbf{x}) \right| \\
= & \left| \frac{\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) m(\mathbf{z}) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z}}{\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z}} - m(\mathbf{x}) \right| \\
= & \left| \frac{\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z})\big(m(\mathbf{z}) - m(\mathbf{x})\big) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z}}{\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z}} \right| \\
= & \frac{\left| \int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z})\big(m(\mathbf{z}) - m(\mathbf{x})\big) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z} \right|}{\left| \int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z} \right|}.
\end{aligned}$$

We want to obtain an upper bound of the bias. Therefore we want to find an upper bound of the numerator and a lower bound of the denominator.

Lower bound of the Denominator:

The denominator is always positive, so the module can be removed,

$$
\begin{aligned}
&\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z}) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
=&\int_{\mathrm{Y}} f_X(\mathbf{z}) \prod_{i=1}^{d} k\left(\frac{z_i}{h_i}\right) \mathrm{d}\mathbf{z} \\
\geq&\int_{\mathcal{D}} f_X(\mathbf{z}) \prod_{i=1}^{d} k\left(\frac{z_i}{h_i}\right) \mathrm{d}\mathbf{z} \quad \text{(since } \mathcal{D} \subseteq \mathrm{Y} \text{ and the integrand is always non-negative)} \\
=&\frac{e^{g(\mathbf{x})}}{\int_{\mathrm{Y}} e^{g(\mathbf{z})}\,\mathrm{d}\mathbf{z}} \int_{\mathcal{D}} e^{g(\mathbf{z})-g(\mathbf{x})} \prod_{i=1}^{d} k\left(\frac{z_i}{h_i}\right) \mathrm{d}\mathbf{z} \qquad \text{(Prop. A6)} \\
=&f_X(\mathbf{x}) \int_{\overline{\mathcal{D}}} e^{g(\mathbf{x}+\mathbf{l})-g(\mathbf{x})} \prod_{i=1}^{d} k\left(\frac{l_i}{h_i}\right) \mathrm{d}\mathbf{l} \qquad \text{let } \mathbf{l}=\mathbf{z}-\mathbf{x} \text{ and } \overline{\mathcal{D}} \equiv \Omega(-\boldsymbol{\delta}^{-},+\boldsymbol{\delta}^{+}) \\
\geq&f_X(\mathbf{x}) \int_{\overline{\mathcal{D}}} e^{-|\mathbf{l}|L_f} \prod_{i=1}^{d} k\left(\frac{l_i}{h_i}\right) \mathrm{d}\mathbf{l} \qquad \text{(Axiom A2 + Lipschitz Inequality)} \\
=&f_X(\mathbf{x}) \int_{\overline{\mathcal{D}}} \prod_{i=1}^{d} e^{-|l_i|L_f} k\left(\frac{l_i}{h_i}\right) \mathrm{d}\mathbf{l}
\end{aligned}
$$

Now considering Propositions A1 and A7, we obtain

$$\int_{-\infty}^{+\infty} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z}) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z} \geq f_X(\mathbf{x}) \prod_{i=1}^{d} \psi_i(\delta_i^{-},\delta_i^{+}). \tag{A6}$$

Upper bound of the Numerator:

$$
\begin{aligned}
&\left|\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})\big(m(\mathbf{z})-m(\mathbf{x})\big) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z}\right| \\
\leq\ &\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})|m(\mathbf{z})-m(\mathbf{x})| f_X(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
=\ &\int_{\mathcal{G}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})|m(\mathbf{z})-m(\mathbf{x})| f_X(\mathbf{z})\,\mathrm{d}\mathbf{z} + \int_{\mathrm{Y}\setminus\mathcal{G}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})|m(\mathbf{z})-m(\mathbf{x})| f_X(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
\leq\ &\int_{\mathcal{G}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})|m(\mathbf{z})-m(\mathbf{x})| f_X(\mathbf{z})\,\mathrm{d}\mathbf{z} + f_X(\mathbf{x}) M \int_{\mathrm{Y}\setminus\mathcal{G}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})\,\mathrm{d}\mathbf{z} \\
=\ &\frac{e^{g(\mathbf{x})}}{\int_{\mathrm{Y}} e^{g(\mathbf{z})}\,\mathrm{d}\mathbf{z}} \int_{\mathcal{G}} e^{g(\mathbf{z})-g(\mathbf{x})} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})|m(\mathbf{z})-m(\mathbf{x})|\,\mathrm{d}\mathbf{z} + f_X(\mathbf{x}) M \int_{\mathrm{Y}\setminus\mathcal{G}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})\,\mathrm{d}\mathbf{z} \\
\leq\ &f_X(\mathbf{x})\left(\int_{\mathcal{G}} e^{g(\mathbf{z})-g(\mathbf{x})} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})|m(\mathbf{z})-m(\mathbf{x})|\,\mathrm{d}\mathbf{z} + MD\right)
\end{aligned}
$$

where

$$
\begin{aligned}
D = \int_{\mathrm{Y}\setminus\mathcal{G}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})\,\mathrm{d} &= \lim_{\omega\to+\infty} \prod_{i=1}^{d} \int_{x_i-\omega}^{x_i+\omega} k_i\left(\frac{x_i-z}{h_i}\right)\mathrm{d}z - \prod_{i=1}^{d} \int_{x_i-\gamma_i^{-}}^{x_i+\gamma_i^{+}} k_i\left(\frac{x_i-z}{h_i}\right)\mathrm{d}z \\
&= \lim_{\omega\to+\infty} \prod_{i=1}^{d} h_i \Phi_i\left(\frac{\omega}{h_i}\right) - \prod_{i=1}^{d} h_i \Phi_i\left(\frac{\gamma_i^{+}}{h}\right) + h_i \Phi_i\left(\frac{\gamma_i^{-}}{h_i}\right).
\end{aligned} \tag{A7}
$$

Let $\mathcal{F} \equiv \Omega(\mathbf{x} - \boldsymbol{\phi}^-, \mathbf{x} + \boldsymbol{\phi}^+) \subseteq \mathcal{G}$, we will later define at our convenience.

$$
\begin{aligned}
& f_X(\mathbf{x}) \left(\int_{\mathcal{G}} e^{g(\mathbf{z}) - g(\mathbf{x})} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) |m(\mathbf{z}) - m(\mathbf{x})| \, \mathrm{d}\mathbf{z} + MD \right) \\
= & f_X(\mathbf{x}) \left(\int_{\mathcal{F}} e^{g(\mathbf{z}) - g(\mathbf{x})} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) |m(\mathbf{z}) - m(\mathbf{x})| \, \mathrm{d}\mathbf{z} + \right. \\
& \left. \int_{\mathcal{G} \setminus \mathcal{F}} e^{g(\mathbf{z}) - g(\mathbf{x})} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) |m(\mathbf{z}) - m(\mathbf{x})| \, \mathrm{d}\mathbf{z} + MD \right) \\
\leq & f_X(\mathbf{x}) \left(\int_{\mathcal{F}} e^{g(\mathbf{z}) - g(\mathbf{x})} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) |m(\mathbf{z}) - m(\mathbf{x})| \, \mathrm{d}\mathbf{z} \right. \\
& \left. + M \int_{\mathcal{G} \setminus \mathcal{F}} e^{g(\mathbf{z}) - g(\mathbf{x})} K_{\mathbf{h}}(\mathbf{x} - \mathbf{z}) \, \mathrm{d}\mathbf{z} + MD \right) \\
= & f_X(\mathbf{x}) \left(\int_{\overline{\mathcal{F}}} e^{g(\mathbf{x}+\mathbf{l}) - g(\mathbf{x})} K_{\mathbf{h}}(-\mathbf{l}) |m(\mathbf{x}+\mathbf{l}) - m(\mathbf{l})| \, \mathrm{d}\mathbf{l} \right. \\
& \left. + M \int_{\overline{\mathcal{G}} \setminus \overline{\mathcal{F}}} e^{g(\mathbf{x}+\mathbf{l}) - g(\mathbf{x})} K_{\mathbf{h}}(-\mathbf{l}) \, \mathrm{d}\mathbf{l} + MD \right) \\
& \text{with } \mathbf{l} = \mathbf{z} - \mathbf{x}, \, \overline{F} \equiv \Omega(-\boldsymbol{\phi}^-, \boldsymbol{\phi}^+) \text{ and } \overline{G} \equiv \Omega(-\boldsymbol{\gamma}^-, \boldsymbol{\gamma}^+) \\
\leq & f_X(\mathbf{x}) \left(\int_{\overline{\mathcal{F}}} e^{L_f |\mathbf{l}|} K_{\mathbf{h}}(\mathbf{l}) L_m |\mathbf{l}| \, \mathrm{d}\mathbf{l} + M \int_{\overline{\mathcal{G}} \setminus \overline{\mathcal{F}}} e^{L_f |\mathbf{l}|} K_{\mathbf{h}}(\mathbf{l}) \, \mathrm{d}\mathbf{l} + MD \right)
\end{aligned}
$$

(A2, A3 + Lipschitz Inequality)

The first integral instead can be solved with Propositions A3, A5 and A8,

$$
\begin{aligned}
& \int_{\overline{\mathcal{F}}} e^{L_f |\mathbf{l}|} K_{\mathbf{h}}(\mathbf{l}) L_m |\mathbf{l}| \, \mathrm{d}\mathbf{l} \\
= & \int_{\overline{\mathcal{F}}} \left(\prod_{i=1}^{d} k\left(\frac{l}{h_i} \right) e^{|l| L_f} \right) L_m \sum_{i=1}^{d} |l_i| \, \mathrm{d}l \, \mathrm{d}\mathbf{l} \\
= & L_m \sum_{k=1}^{d} \left(\prod_{i \neq k}^{d} \int_{-\phi_i^-}^{\phi_i^+} k\left(\frac{l}{h_i} \right) e^{|l| L_f} \, \mathrm{d}z \right) \int_{-\phi_k^-}^{\phi_k^+} k\left(\frac{l}{h} \right) e^{|l| L_f} |l_i| \, \mathrm{d}l \qquad \text{(Prop. A8)} \\
= & L_m \sum_{k=1}^{d} \xi_k(\phi_k^-, \phi_k^+) \left(\prod_{i \neq k}^{d} \zeta_i(\phi_i^-, \phi_i^+) \right).
\end{aligned}
$$

The second integral can be solved using Propositions A1 andA7,

$$
\begin{aligned}
\int_{\mathcal{G} \setminus \mathcal{F}} e^{L_f |\mathbf{l}|} K_{\mathbf{h}}(\mathbf{l}) \, \mathrm{d}\mathbf{z} &= \int_{\mathcal{G}} e^{L_f |\mathbf{l}|} K_{\mathbf{h}}(\mathbf{l}) \, \mathrm{d}\mathbf{l} - \int_{\mathcal{F}} e^{L_f |\mathbf{l}|} K_{\mathbf{h}}(\mathbf{l}) \, \mathrm{d}\mathbf{l} \\
&= \prod_{i=1}^{d} \zeta_i(\gamma_i^-, \gamma_i^+) - \prod_{i=1}^{d} \zeta_i(\phi_i^-, \phi_i^+).
\end{aligned}
$$

A good choice for $\mathcal{F}$ is $\phi_i^- = \min(\gamma_i^-, M/L_f)$ and $\phi_i^+ = \min(\gamma_i^+, M/L_f)$, as in this way we obtain a tighter bound. In last analysis,

$$
\left| \mathbb{E}\left[\lim_{n \to \infty} \hat{m}_n(\mathbf{x}) \right] - m(\mathbf{x}) \right| \leq \frac{L_m \sum_{k=1}^{d} \xi_k(\phi_i^k, \phi_i^k) \prod_{i \neq k}^{d} \zeta(\phi_i^-, \phi_i^+) + M\left(\prod_{i=1}^{d} \zeta(\gamma_i^-, \gamma_i^+) - \prod_{i=1}^{d} \zeta(\phi_i^-, \phi_i^+) + D \right)}{\prod_{i=1}^{d} \psi_i(\delta_i^-, \delta_i^+)}
$$

showing the correctness of Theorem 2. □

In order to prove Theorem 3 we shall note that $\mathrm{Y} \equiv \mathcal{G} \equiv \mathcal{F}$, therefore the lower bound can be bounded by

$$\begin{aligned}\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z}) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z} &= \int_{\mathrm{Y}} f_X(\mathbf{z}) \prod_{i=1}^{d} k_i\left(\frac{x_i - z_i}{h_i}\right)\mathrm{d}\mathbf{z} \\ &\geq f_X(\mathbf{x}) \int_{\overline{\mathrm{Y}}} \prod_{i=1}^{d} k_i\left(\frac{l_i}{h_i}\right) e^{-l_i L_f}\,\mathrm{d}\mathbf{l} \\ &= f_X(\mathbf{x}) \prod_{i=1}^{d} \psi_i(v_i^-, v_i^+) \end{aligned} \tag{A8}$$

for the numerator, instead

$$\begin{aligned}&\left|\int_{\mathrm{Y}} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z})\big(m(\mathbf{z}) - m(\mathbf{x})\big) f_X(\mathbf{z})\,\mathrm{d}\mathbf{z}\right| \\ \leq\ & f_X(\mathbf{x}) \int_{\mathrm{Y}} e^{g(\mathbf{z}) - g(\mathbf{x})} K_{\mathbf{h}}(\mathbf{x}-\mathbf{z}) |m(\mathbf{z}) - m(\mathbf{x})|\,\mathrm{d}\mathbf{z} \\ \leq\ & f_X(\mathbf{x}) \int_{\overline{\mathrm{Y}}} e^{L_f|\mathbf{l}|} K_{\mathbf{h}}(\mathbf{l}) L_m |\mathbf{l}|\,\mathrm{d}\mathbf{l} \qquad \text{where } \overline{\mathrm{Y}} \equiv \Omega(v^-, v^+)\end{aligned}$$

and therefore, for the reasoning already made for Theorem 2,

$$\left|\mathbb{E}\left[\lim_{n\to\infty} \hat{m}_n(\mathbf{x})\right] - m(\mathbf{x})\right| \leq \frac{L_m \sum_{k=1}^{d} \xi_k(v_k^-, v_k^+) \prod_{i\neq k}^{d} \zeta_i(v_i^-, v_i^+)}{\prod_{i=1}^{d} \psi_i(v_i^-, v_i^+)}$$

where $\xi_{A_k}, \zeta, \Psi, \varphi$ are defined as in Theorem 2 and $\phi_i^- = v_i^-, \phi_i^+ = v_i^+$.

Appendix B. Kernels

Appendix B.1. Gaussian Kernel

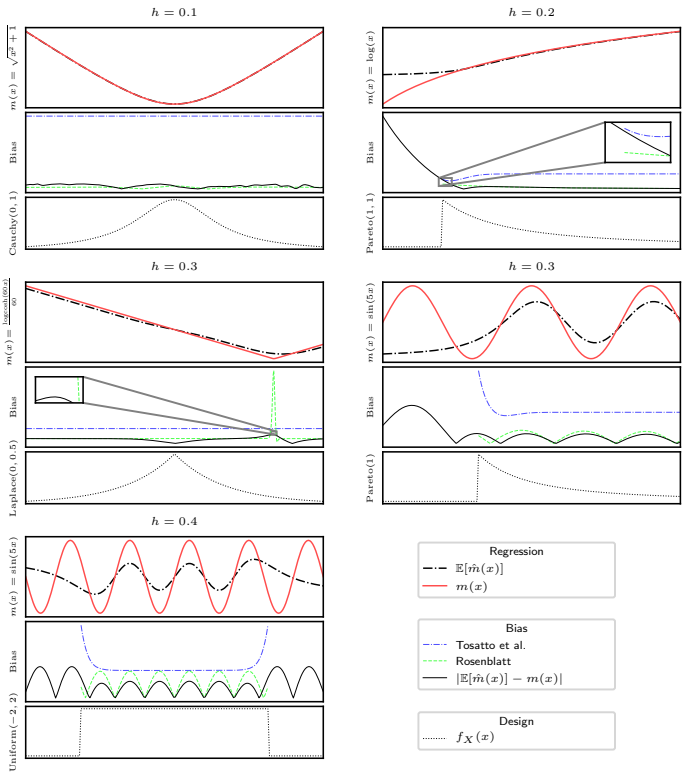

Figure A1. Same experiment as in Figure 1, with Gaussian Kernels.

Appendix B.2. Box Kernel

The box kernel, defined as $k(x) = I(x)$, has $B(a,c) = g_B(a,c) - g_B(0,c)$ and $C(a,c) = g_C(a,c) - g_C(0,c)$, where

$$g_B(x,c) = \int k(x)e^{-xc}\,\mathrm{d}x = \begin{cases} 0 & \text{if} \quad x < -1 \\ \frac{e^{-c}\left(e^{2c}-1\right)}{c} & \text{if} \quad x > 1, \quad (+\text{const}), \\ \frac{e^{-c}\left(e^{cx+c}-1\right)}{c} & \text{otherwise} \end{cases}$$

$$g_C(x,c) = \int k(x)e^{-xc}x\,\mathrm{d}x = \begin{cases} 0 & \text{if} \quad x < -1 \\ \frac{e^{-c}\left(ce^{2c}-e^{2c}+c+1\right)}{c^2} & \text{if} \quad x > 1, \quad (+\text{const}), \\ \frac{e^{-c}\left(cxe^{cx+c}-e^{cx+c}+c+1\right)}{c^2} & \text{otherwise} \end{cases}$$

We show the results of a numerical simulation using the Box kernel in Figure A2.

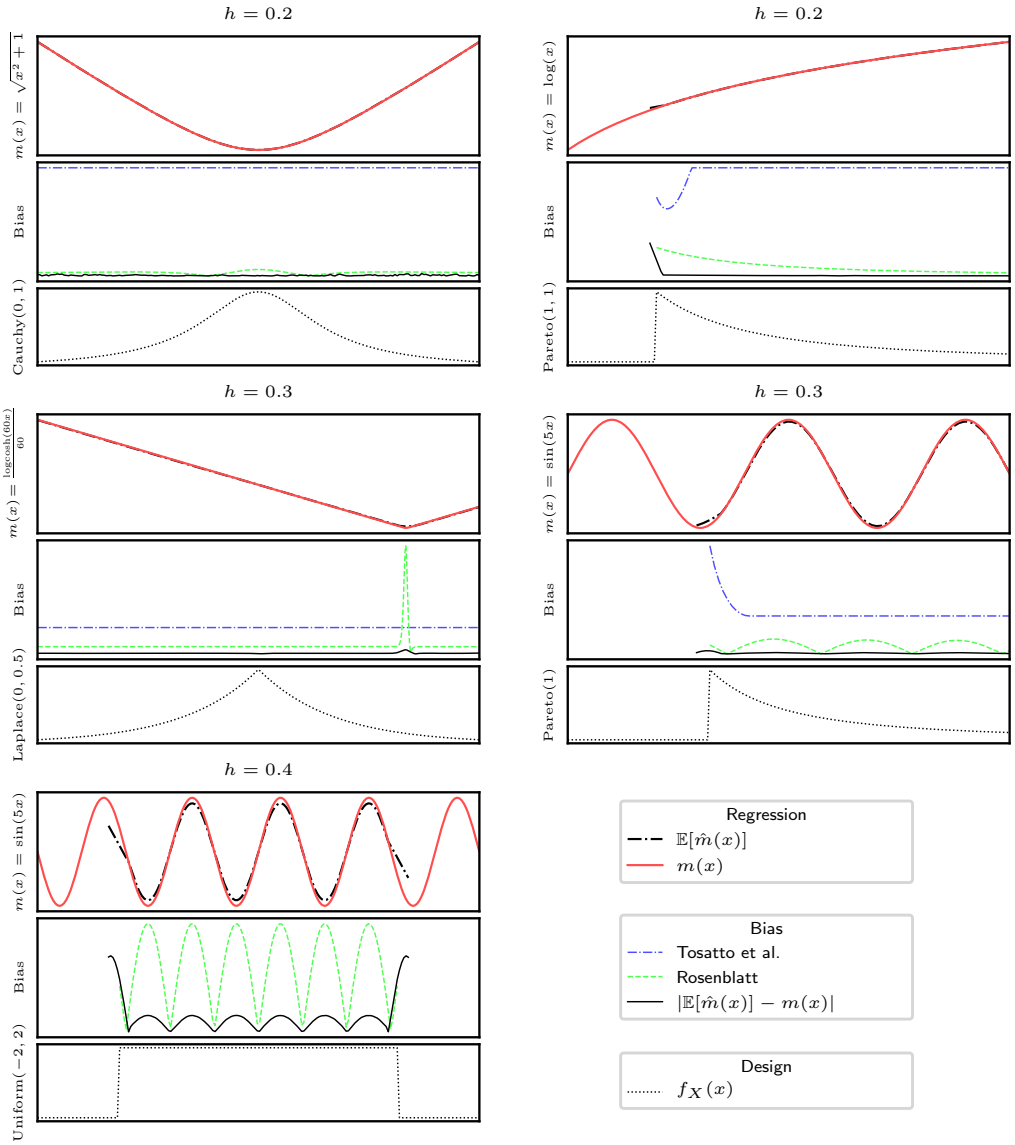

Figure A2. Same experiment as in Figure 1, with Box Kernel.

Appendix B.3. Triangular Kernel

The triangular kernel, defined as $k(x) = I(x)(1 - |x|)$, has $B(a,c) = g_B(a,c) - g_B(0,c)$ and $C(a,c) = g_C(a,c) - g_C(0,c)$, where

$$g_B(x,c) = \int k(x)e^{-xc}\,\mathrm{d}x = \begin{cases} 0 & \text{if} \quad x < -1 \\ \frac{e^{-cx}(e^{cx+c} - c(x+1) - 1)}{c^2} & \text{if} \quad -1 \le x \le 0, \\ \frac{e^{-cx}(c(x-1) - 2e^{cx} + e^{cx+c} + 1)}{c^2} & \text{if} \quad 0 < x \le 1, \\ \frac{e^{-c}(e^c - 1)^2}{c^2} & \text{if} \quad x > 1, \end{cases} (+\text{const}),$$

$$g_C(x,c) = \int k(x)e^{-xc}x\,\mathrm{d}x = \begin{cases} 0 & \text{if} \quad x < -1 \\ \frac{e^{-cx}(-c^2x(x+1) - c(e^{cx+c} + 2x + 1) + 2(e^{cx+c} - 1))}{c^3} & \text{if} \quad -1 \le x \le 0, \\ \frac{e^{-cx}(c^2(x-1)x - c(e^{cx+c} - 2x + 1) + 2(e^{cx-c} + 1 - 2e^{cx}))}{c^3} & \text{if} \quad 0 < x \le 1, \\ -\frac{e^{-c}(e^c - 1)(e^c)c + c - 2e^c + 2}{c^3} & \text{if} \quad x > 1, \end{cases} (+\text{const}),$$

We show the results of a numerical simulation using the Triangular kernel in Figure A3.

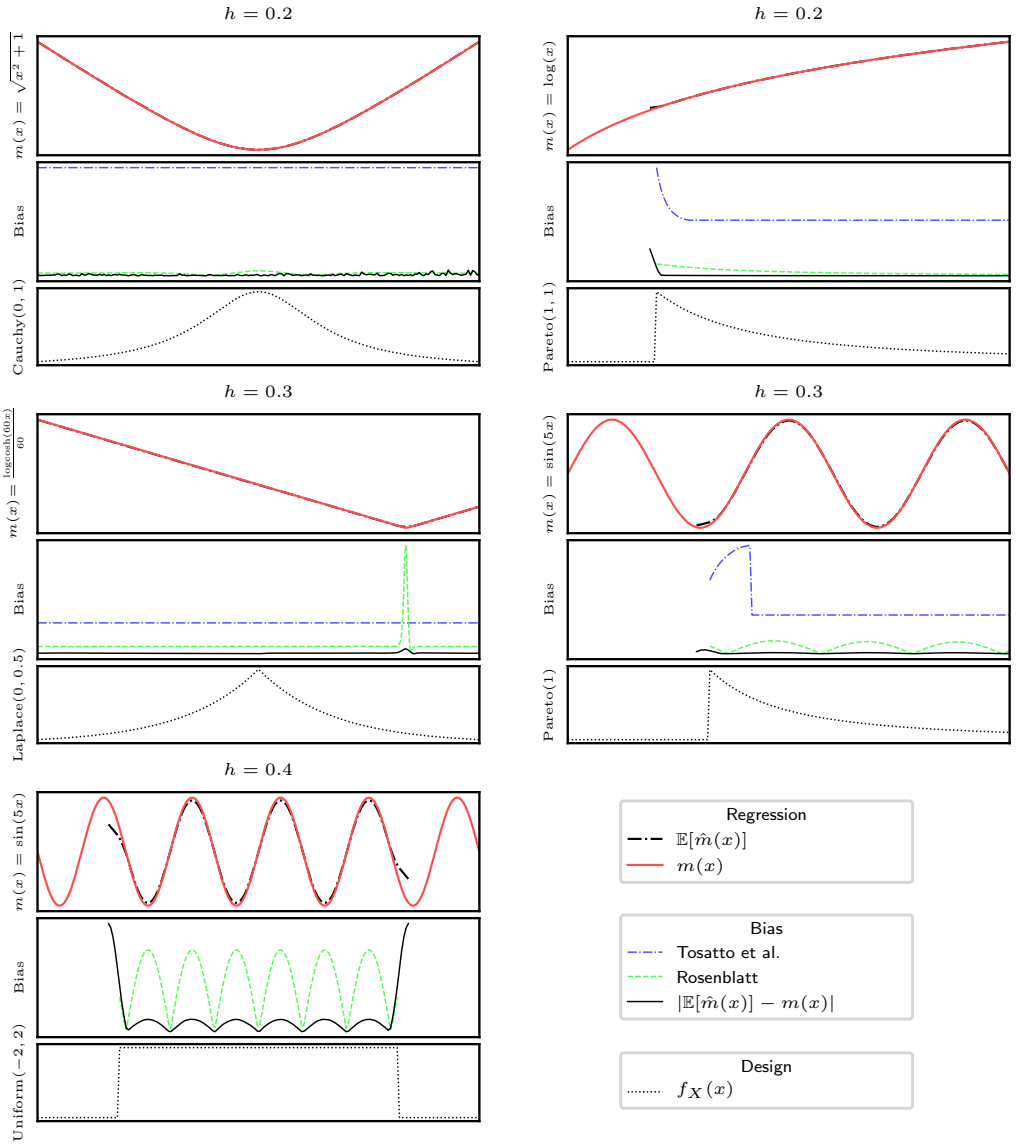

Figure A3. Same experiment as in Figure 1, with Triangular Kernel. Note that this particular kernel, does not have finite bound for $L_f = 0$.

References

1. Bansal, R.; Gallant, A.R.; Hussey, R.; Tauchen, G. Nonparametric Estimation of Structural Models for High-Frequency Currency Market Data. *J. Econom.* **1995**, *66*, 251–287.
2. Nguyen-Tuong, D.; Peters, J. Using Model Knowledge for Learning Inverse Dynamics. In Proceedings of the International Conference on Robotics and Automation, Anchorage, AK, USA, 3–7 May 2010; pp.2677–2682.
3. Wang, J.; Hertzmann, A.; Fleet, D.J. Gaussian Process Dynamical Models. *Adv. Neural Inf. Process. Syst.* **2006**, *27*, 1441–1448.
4. Deisenroth, M.P.; Rasmussen, C.E. PILCO: A Model-based and Data-efficient Approach to Policy Search. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, Omnipress, Bellevue, WA, USA, 28 June–2 July 2011; pp. 465–472.
5. Kroemer, O.; Ugur, E.; Oztop, E.; Peters, J. A Kernel-Based Approach to Direct Action Perception. In Proceedings of the International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 2605–2610.
6. Kroemer, O.B.; Peters, J.R. A Non-Parametric Approach to Dynamic Programming. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Dutchess County, NY, USA, 2011; pp. 1719–1727.
7. Ormoneit, D.; Sen, S. Kernel-Based Reinforcement Learning. *Mach. Learn.* **2002**, *49*, 161–178.
8. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2574–2582.
9. Backurs, A.; Indyk, P.; Wagner, T. Space and Time Efficient Kernel Density Estimation in High Dimensions. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 15799–15808.
10. Charikar, M.; Siminelakis, P. Hashing-Based-Estimators for Kernel Density in High Dimensions. In Proceedings of the 58th Annual Symposium on Foundations of Computer Science (FOCS), Berkeley, CA, USA, 15–17 October 2017; pp. 1032–1043.
11. Herrmann, E.; Gasser, T.; Kneip, A. Choice of Bandwidth for Kernel Regression when Residuals are Correlated. *Biometrika* **1992**, *I*, 783–795.
12. Härdle, W.; Marron, J. Asymptotic Nonequivalence of Some Bandwidth Selectors in Nonparametric Regression. *Biometrika* **1985**, *72*, 481–484.
13. Köhler, M.; Schindler, A.; Sperlich, S. A Review and Comparison of Bandwidth Selection Methods for Kernel Regression. *Int. Stat. Rev.* **2014**, *82*, 243–274.
14. Ray, B.K.; Tsay, R.S. Bandwidth Selection for Kernel Regression with Long-Range Dependent Errors. *Biometrika* **1997**, *84*, 791–802.
15. Nadaraya, E.A. On Estimating Regression. *Theory Probab. Its Appl.* **1964**, *9*, 141–142.
16. Watson, G.S. Smooth Regression Analysis. *Sankhyā Indian J. Stat. Ser. A* **1964**, *28*, 359–372.
17. Rosenblatt, M. Conditional Probability Density and Regression Estimators. *Multivar. Anal. II* **1969**, *25*, 31.
18. Fan, J. Design-Adaptive Nonparametric Regression. *J. Am. Stat. Assoc.* **1992**, *87*, 998–1004.
19. Fan, J.; Gijbels, I. Variable Bandwidth and Local Linear Regression Smoothers. *Annals Stat.* **1992**, *20*, 2008–2036.
20. Mack, Y.; Müller, H.G. Convolution Type Estimators for Nonparametric Regression. *Stat. Probab. Lett.* **1988**, *7*, 229–239.
21. Wasserman, L. *All of Nonparametric Statistics*; Springer: New York, NY, USA, 2006.
22. Györfi, L.; Kohler, M.; Krzyzak, A.; Walk, H. *A Distribution-Free Theory of Nonparametric Regression*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
23. Miculescu, R. A Sufficient Condition for a Function to Satisfy a Weak Lipschitz Condition. Mathematical Reports. 2000. Available online: http://www.imar.ro/journals/Mathematical_Reports/Pdfs/2007/3/miculescu.pdf (accessed on 29 December 2020).
24. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.
25. Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.* **1956**, 832–837.
26. Wood, G.R.; Zhang, B.P. Estimation of the Lipschitz Constant of a Function. *J. Glob. Opt.* **1996**, *8*, 91–103.
27. Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; Zaremba, W. OpenAI Gym. *arXiv* **2016**, arXiv:1606.01540.
28. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-Level Control Through Deep Reinforcement Learning. *Nature* **2020**, *518*,529–533.
29. Ernst, D.; Geurts, P.; Wehenkel, L. Tree-Based Batch Mode Reinforcement Learning. *J. Mach. Learn. Res.* **2005**, *6*, 503–556.
30. Tosatto, S.; Carvalho, J.; Abdulsamad, H.; Peters, J. A Nonparametric Off-Policy Policy Gradient. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Virtual Conference, 26–28 August 2020.