# Beating Monte Carlo Integration: a Nonasymptotic Study of Kernel Smoothing Methods

**Stephan Clémençon** **François Portier**
Télécom ParisTech, LTCI, Université Paris Saclay

## Abstract

Evaluating integrals is an ubiquitous issue and Monte Carlo methods, exploiting advances in random number generation over the last decades, offer a popular and powerful alternative to integration deterministic techniques, unsuited in particular when the domain of integration is complex. This paper is devoted to the study of a kernel smoothing based competitor built from a sequence of $n \geq 1$ i.i.d random vectors with arbitrary continuous probability distribution $f(x)dx$, originally proposed in [7], from a nonasymptotic perspective. We establish a probability bound showing that the method under study, though biased, produces an estimate approximating the target integral $\int_{x\in\mathbb{R}^d} \varphi(x)dx$ with an error bound of order $o(1/\sqrt{n})$ uniformly over a class $\Phi$ of functions $\varphi$, under weak complexity/smoothness assumptions related to the class $\Phi$, outperforming Monte-Carlo procedures. This striking result is shown to derive from an appropriate decomposition of the maximal deviation between the target integrals and their estimates, highlighting the remarkable benefit to averaging strongly dependent terms regarding statistical accuracy in this situation. The theoretical analysis then rests on sharp probability inequalities for degenerate $U$-statistics. It is illustrated by numerical results in the context of covariate shift regression, providing empirical evidence of the relevance of the approach.

Proceedings of the $21^{st}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

## 1 INTRODUCTION

For over two thousands years, numerical integration has been the subject of intense research activity, starting with Babylonian mathematics and the elaboration of quadrature rules for measuring areas and volumes. It lead to the development of a very wide variety of algorithms for calculating approximately the numerical value of a given (well-defined) integral with a controlled error, ranging from (possibly adaptive) methods based on interpolating functions to (quasi/advanced) Monte Carlo techniques. One may refer to *e.g.* [6] for an excellent account of deterministic techniques for numerical integration and to [13] for an introduction to Monte Carlo integration. Probabilistic approaches have been proved quite useful in high-dimensional cases to circumvent the curse of dimensionality phenomenon with the advent of computer technology and significant advances in pseudo-random number generation. Error bounds achieved by Monte Carlo integration methods based on a simulated sample of size $n \geq 1$ are typically of order $1/\sqrt{n}$, the rate of the classical CLT. Recently, a competitor based on kernel smoothing has been proposed in [7]. The resulting integral estimates can be interpreted as *biased* importance sampling (IS, in abbreviated form) estimates, where the (true) importance function is replaced by leave-one-out kernel estimators. Provided that the instrumental density used in this integral estimation procedure is smooth enough, it has been proved that the asymptotic rate of convergence can be faster than $1/\sqrt{n}$ for an appropriate choice of the kernel bandwidth (see also [2] for a similar study in the Markovian context). It is the goal of this paper to investigate this striking phenomenon much further, from both a nonasymptotic and functional perspective and establish confidence upper bounds holding true for finite samples, uniformly over classes of functions of controlled complexity. The main argument relies on an adequate decomposition of the integral estimates obtained by means of this method, in which degenerate $U$-statistics appear in particular, and on recent concentration inequalities for such functionals, gener-

ally used in the context of asymptotic study of (variable bandwidth) kernel density estimation methods, see *e.g.* [10]. Incidentally, attention should be paid to the fact that the analysis carried out in this article sheds light on a striking phenomenon: whereas it has been shown that the dependence structure among averaged identically distributed r.v.'s may significantly deteriorate the convergence rates in a wide variety of situations (*e.g.* for long-memory processes or weakly dependent sequences with non geometrically decaying mixing coefficients, see [8] for instance, in cross-validation procedures), it is proved here that the dependence between the components averaged to produce the kernel smoothing-based integral estimates is in contrast of great benefit to statistical accuracy.

The article is structured as follows. Basics in Monte-Carlo integral approximation are briefly recalled in section 2, together with the alternative method originally proposed in [7]. The main results of the paper are then stated in section 3 and an illustrative application is presented in section 4. Finally, some concluding remarks are collected in section 5. Technical details and additional remarks are deferred to the Supplementary Material.

# 2 BACKGROUND

Here and throughout, $(X_n)_{n\geq 1}$ is a sequence of continuous independent and identically distributed random vectors, taking their values in $\mathbb{R}^d$, $d \geq 1$, with density $f(x)$ w.r.t. Lebesgue measure $\mu_{Leb}$, $\Phi$ is a given class of Borelian functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ and $K : \mathbb{R}^d \to \mathbb{R}$ is a symmetric *kernel function* of order $l \geq 1$, *i.e.* a Borelian function, integrable w.r.t. Lebesgue measure such that $\int K(x)dx = 1$, $K(x) = K(-x)$ for all $x \in \mathbb{R}^d$. We set $||z||_\Phi := \sup_{\varphi\in\Phi} |z(\varphi)|$ for any real valued sequence $z = \{z(\varphi)\}_{\varphi\in\Phi}$. Denote by $\mathbb{I}\{\mathcal{E}\}$ the indicator variable of any event $\mathcal{E}$. For any Borelian function $g : \mathbb{R}^d \to \mathbb{R}$, the closure of the set $\{x \in \mathbb{R}^d : g(x) \neq 0\}$ is denoted by $Supp(g)$ and by $||g||_\infty$ is meant the essential supremum of $g$ when it is bounded almost everywhere. For any $h > 0$ and $x \in \mathbb{R}^d$, we set $K_h(x) = K(h^{-1}x)/h^d$. When well-defined, the convolution product between two real-valued Borelian functions $g(x)$ and $w(x)$ is denoted by $g * w(x) = \int_{x'\in\mathbb{R}^d} g(x - x')w(x')dx'$. For any $\beta > 0$, we set $\lfloor\beta\rfloor = \max\{n \in \mathbb{N} : n < \beta\}$. Let $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, we set $|\alpha| = \sum_{i=1}^d \alpha_i$ and mean by $\partial_\alpha$ the differential operator $\partial^{|\alpha|}/\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}$. For $m \in \mathbb{N}$, whenever $\Omega$ is an open subset of $\mathbb{R}^d$, the space of real-valued functions on $\Omega$ that are differentiable up to order $m$ is denoted by $\mathcal{C}^m(\Omega)$ and, for any $\beta > 0$, $L > 0$, we denote by $\mathcal{H}_{\beta,L}(\Omega)$ the space of functions $g$ in $\mathcal{C}^{\lfloor\beta\rfloor}(\Omega)$ with all derivatives up to order $\lfloor\beta\rfloor$ bounded by $L$ and such that, for any multi-index $\alpha \in \mathbb{N}^d$ with $|\alpha| \leq \lfloor\beta\rfloor$:

$$\forall (x, y) \in \Omega^2, \ \ |\partial_\alpha f(x) - \partial_\alpha f(y)| \leq L||x - y||^{\beta-|\alpha|},$$

denoting by $||.||$ the usual Euclidean norm on $\mathbb{R}^d$.

## 2.1 Integral(s) Approximation

It is the goal of this paper to analyze the performance of statistical techniques to approximate accurately the integral

$$\mathcal{I}(\varphi) = \int_{x\in\mathbb{R}^d} \varphi(x)dx, \tag{1}$$

based on the observation of the i.i.d. sample $X_1, \ldots, X_n$, $n \geq 1$. When the support $\mathcal{K}$ of $\varphi$, *i.e.* the domain of integration related to (1), is compact, a basic Monte-Carlo method would consist in generating independent random vectors $U_1, \ldots, U_n$ uniformly distributed over a domain $\mathcal{H} \supset \mathcal{K}$ (a union of hypercubes typically, for computational simplicity) and compute the natural (unbiased) Monte-Carlo estimate:

$$\widehat{\mathcal{I}}_n(\varphi) = \frac{1}{n}\sum_{i=1}^n \varphi(U_i). \tag{2}$$

Beyond classic limit theorems (SLLN, CLT, LIL, *etc.*), the accuracy of estimate (2) can be evaluated for a fixed sample size $n \geq 1$. For simplicity, suppose that $\varphi$ is bounded almost-everywhere. In absence of any smoothness assumption for the integrand $\varphi$, a straightforward application of Hoeffding's inequality (see [12]) shows that, for all $n \geq 1$, for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$:

$$\left|\widehat{\mathcal{I}}_n(\varphi) - \mathcal{I}(\varphi)\right| \leq ||\varphi||_\infty \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Maximal deviations over a class $\Phi$ of functions $\varphi$ s.t. $||\varphi||_\infty \leq M_\Phi < +\infty$ can be obtained by means of concentration inequalities under appropriate complexity assumptions on $\Phi$. Indeed, by virtue of McDiarmid's inequality (see [18]) combined with classical symmetrization and randomization arguments, for all $n \geq 1$, for any $\delta \in (0, 1)$, we have with probability larger than $1 - \delta$:

$$\left|\left|\widehat{\mathcal{I}}_n - \mathcal{I}\right|\right|_\Phi \leq 2\mathbb{E}\left[\mathcal{R}_n(\Phi)\right] + M_\Phi\sqrt{\frac{2\log(2/\delta)}{n}}, \tag{3}$$

where the Rademacher average associated to the set $\{(\varphi(U_1), \ldots, \varphi(U_n)) : \varphi \in \Phi\}$ is denoted by $\mathcal{R}_n(\Phi) = \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\sup_{\varphi\in\Phi} \frac{1}{n}\left|\sum_{i=1}^n \epsilon_i\varphi(U_i)\right|\right]$ and $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher variables, independent from the $U_i$'s. The expected Rademacher average measures the richness of the class $\Phi$, see *e.g.* [15]. Classically [24], if $\Phi$ is a Vapnik-Chervonenkis VC major class of functions of finite VC dimension $V < \infty$ (*i.e.* if the collection of sub-level sets

$\{\{\varphi(x) \geq t\} : (\varphi,\ t) \in \Phi \times \mathbb{R}\}$ is of finite VC dimension $V < +\infty$), we have $\mathbb{E}\left[\mathcal{R}_n(\Phi)\right] \leq C\sqrt{V/n}$, where $C < +\infty$ is a universal constant and the basic Monte Carlo procedure permits to approximate the integrals (1) uniformly over the class $\Phi$ at the rate $1/\sqrt{n}$ in this case. Except in pathologic situations, a basic CLT argument can be used to prove that this rate bound cannot be improved. Whereas many refinements of the bounds stated (involving the variance of the $\varphi(U_i)$'s or other measures of complexity for classes of functions, such as metric entropies) can be considered, focus is here on an alternative method, significantly improving upon Monte Carlo integration in terms of order of the (nonasymptotic) rate bound achieved.

### 2.2 A Kernel Smoothing Alternative

We now describe at length the integral estimation procedure promoted in this paper. As an alternative to (2), it is proposed in [7] to consider the estimate

$$\widetilde{\mathcal{I}}_n(\varphi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}_{n,i}(X_i)}, \tag{4}$$

denoting by

$$\widehat{f}_{n,i}(x) = \frac{1}{n-1}\sum_{1\leq j\leq n,\ j\neq i} K_h(x - X_j) \tag{5}$$

the smoothed *leave-one-out* estimator of $f(x)$ based on kernel $K$ and bandwidth $0 < h \leq h_0$, computed with all the $X_j$'s except $X_i$, for $1 \leq i \leq n$. The expectation of the estimate (5) is equal to the convolution product $K_h * f(x)$. Assume in addition that the kernel function $K$ is of order $\lfloor\beta\rfloor$ with $\beta > 0$, meaning that $x \in \mathbb{R}^d \mapsto ||x||^l |K(x)|$ is integrable for all $l \leq \lfloor\beta\rfloor$ and

$$\int_{x\in\mathbb{R}^d} \prod_{i=1}^{d} x_i^{\alpha_j} K(x)dx = 0$$

for all $\alpha \in \mathbb{N}^d$ such that $|\alpha| \leq \lfloor\beta\rfloor$. Provided that $f$ belongs to the Hölder space $\mathcal{H}_{\beta,L}(\mathbb{R}^d)$ for some $L > 0$, the deviation $|K_h * f(x) - f(x)|$ is of order $O(h^\beta)$, see Lemma 5 in the supplementary file. As shall be seen at length in the next subsection, though biased and complex (the quantities involved in the average (4) exhibit a strong dependence structure), the estimator (4) is significantly more accurate, under specific hypotheses (on the decay rate of $h$ as $n \to +\infty$ and the smoothness of $f$ in particular), than the (unbiased) IS Monte Carlo estimate

$$\bar{\mathcal{I}}_n(\varphi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\varphi(X_i)}{f(X_i)},$$

obtained when replacing the leave-one-out estimators $\widehat{f}_{n,i}(X_i)$ by the true values $f(X_i)$ of the instrumental density in (4). Although the smoothing stage induces a bias in the estimation of (1), it may drastically accelerate the convergence, as claimed in limit results proved in [7]. Before investigating the accuracy of (4), uniformly over a class of functions of controlled complexity (in a sense that shall be specified later) from a nonasymptotic angle, a few remarks are in order.

**Remark 1.** (Multivariate kernel) *Many univariate kernels have been proposed in the literature:* $u \mapsto (1/2)\mathbb{I}\{-1 \leq u \leq +1\}$ *(rectangular),* $u \mapsto (1 - |u|)\mathbb{I}\{-1 \leq u \leq +1\}$ *(triangular),* $u \mapsto (1/\sqrt{2\pi})\exp(-u^2)/2$ *(Gaussian) or* $u \mapsto (3/4)(1 - u^2)\mathbb{I}\{-1 \leq u \leq +1\}$ *(Epanechnikov). Extensions to the multivariate framework is straightforward by tensorization: for any univariate kernel* $K(u)$ *of order* $m \in \mathbb{N}$*, the product kernel defined below is a multivariate kernel of same order:* $\forall d \geq 1$, $u = (u_1,\ \ldots,\ u_d) \in \mathbb{R}^d \mapsto \prod_{i=1}^{d} K(u_i)$.

**Remark 2.** (On computational complexity) *The fact that the theoretical results stated in [7] and in the next section (see Theorem 1 therein) are valid whathever the dimension* $d \geq 1$ *makes the method described above very attractive. Truth be told, the latter is appropriate in low/moderate dimensional settings only. In high dimensions, kernel smoothing methods behave poorly as they face the dimensionality curse, see [21]. In addition, the computational budget of* $\widetilde{\mathcal{I}}_n$ *(in* $n^2$*) is larger than that of* $\bar{\mathcal{I}}_n$ *(in* $n$*). This makes* $\widetilde{\mathcal{I}}_n$ *particularly appropriate in these situations: (i) real dataset when* $f$ *is unknown (see [2]) and (ii) numerical integration when* $\varphi$ *is computationally expensive (see [19]).*

## 3 NON-ASYMPTOTIC BOUNDS

It is the purpose of this section to establish nonasymptotic upper bounds for the maximal deviation

$$||\widetilde{\mathcal{I}}_n - \mathcal{I}||_\Phi = \sup_{\varphi\in\Phi} \left|\widetilde{\mathcal{I}}_n(\varphi) - \mathcal{I}(\varphi)\right| \tag{6}$$

of estimated integrals based on kernel smoothing from their true values. As previously mentioned, the variables $\varphi(X_i)/\widehat{f}_{n,i}(X_i)$ averaged in (4) are identically distributed and "close" to the $\varphi(X_i)/f(X_i)$'s but are, in contrast, highly dependent: a same subset of $n-2$ original observations is involved in the computation of any pair of such r.v.'s. However, it is well-known in Statistics that averaging dependent (identically distributed) random variables may considerably refine accuracy: a $U$-statistics, say $U_n$, is a typical example of statistics obtained by averaging strongly dependent terms and providing estimate of the mean $\theta = \mathbb{E}[U_n]$ with minimum variance among all unbiased estimates

of $\theta$, see *e.g.* [16][1]. By means of an appropriate decomposition of (4) (where, incidentally, the appearance of degenerate $U$-statistics plays a crucial role), we shall show that the uniform deviation (6) may be much smaller than the bound (3) in certain situations (for proper choice of the bandwidth $h = h_n$ in particular). The following assumptions are involved in the subsequent analysis.

**$\mathbf{A_1}$** Let $\beta > 0$ and $L > 0$. The density $f$ belongs to the Hölder class $\mathcal{H}_{\beta,L}(\mathbb{R}^d)$.

**$\mathbf{A_2}$** The class $\Phi$ is countable, uniformly bounded, *i.e.* $M_\Phi := \sup_{\varphi\in\Phi} ||\varphi||_\infty < +\infty$, and of VC type [9] (w.r.t. the constant envelope $M_\Phi$), meaning that there exist nonnegative constants $A$ and $v$ s.t. for all probability measures $Q$ on $\mathbb{R}^d$ and any $0 < \epsilon < 1$: $\mathcal{N}(\Phi, L_2(Q), \epsilon) \leq (AM_\Phi/\epsilon)^v$, where $\mathcal{N}(\Phi, L_2(Q), \epsilon)$ denotes the smallest number of $L_2(Q)$-balls of radius less than $\epsilon$ required to cover class $\Phi$ (covering number).

**$\mathbf{A_3}$** The set $D_\Phi \stackrel{def}{=} \bigcup_{\varphi\in\Phi} Supp(\varphi)$ is compact.

**$\mathbf{A_4}$** The density of the $X_i$'s is bounded by below on the domain $D_\Phi$: $\lambda \stackrel{def}{=} \inf_{x\in D_\Phi} f(x) > 0$.

**$\mathbf{A_5}$** For all $\varphi \in \Phi$, the function $\varphi/f$ belongs to the Hölder class $\mathcal{H}_{\beta,L}(\mathbb{R}^d)$.

The result stated below reveals that, under these hypotheses, the integral approximation method recalled in subsection 2.2, achieves a rate bound faster than $1/\sqrt{n}$ for an appropriate choice of the bandwidth $h_n > 0$.

**Theorem 1.** (Probability rate bounds) *Suppose that assumptions $\mathbf{A_1} - \mathbf{A_5}$ are fulfilled. For all $\delta \in (0,1)$, there exists a set $\mathcal{C}_\delta \subset \mathbb{N}\times\mathbb{R}$ depending on $\delta$, $\Phi$, $K$, $(\beta,\ L)$ and $f$ such that, for all $(n,h) \in \mathcal{C}_\delta$, with probability at least $1-\delta$, we have:*

$$\sup_{\varphi\in\Phi} \left|\widetilde{\mathcal{I}}_n(\varphi) - \mathcal{I}(\varphi)\right| \leq C_\delta \left\{ h^\beta + \frac{|\log\left(h^{d/2}\right)|}{nh^d} \right\}.$$

*where $C_\delta$ is a constant depending on $\delta$, $\Phi$, $K$, $(\beta,\ L)$ and $f$. In particular, choosing $h = h_n$ so that $h_n = o(1/n^{1/(2\beta)})$ and $1/n^{1/(2d)} = o(h_n)$ as $n \to +\infty$, which guarantees that $(n,h) \in \mathcal{C}_\delta$ and is always possible as soon as $\beta > d$, yields a rate bound of order $o_{\mathbb{P}}(1/\sqrt{n})$.*

Before sketching the argument of the theorem above, a few comments are in order.

[1] Let $(S,\mathcal{S})$ be a measurable space. Recall that the $U$-statistic of kernel $\omega : S\times S \to \mathbb{R}$ based on the i.i.d. observations $Z_1, \ldots, Z_n$ valued in $S$ is the quantity $1/n(n-1)\sum_{i\neq j}\omega(Z_i,Z_j)$. One says it is *degenerate* when $\mathbb{E}[\omega(Z,z)] = \mathbb{E}[\omega(z,Z)] = 0$, for all $z \in S$.

**Remark 3.** (On complexity/smoothness assumptions) *It is supposed here that the class $\Phi$ is of* VC *type, cf assumption $\mathbf{A_2}$, meaning that uniform entropy numbers grow at a polynomial rate. We recall for completeness that a uniformly bounded* VC *major class of functions of finite* VC *dimension $V < +\infty$ is of course of* VC *type (constants $A$ and $v$ can be expressed as functions of $V$, see e.g. Theorem 2.6.7 in [24]). The hypothesis that $\Phi$ is countable can be weakened, using the notion of countable approximability, see the definition in [17] on p. 492. In addition, observe that we assumed here that $f$ and the $\varphi/f$'s belong to the same Hölder class for the sake of simplicity only. The analysis can be straightforwardly extended to more general smoothness assumptions, at the price of more complex formulas for the rate bounds.*

*Proof.* The argument is based on the following decomposition of the estimator (4) for an arbitrary element $\varphi$ of class $\Phi$:

$$\begin{aligned}\widetilde{I}_n(\varphi) = \frac{2}{n}\sum_{i=1}^n \frac{\varphi(X_i)}{f(X_i)} \\ - \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} \frac{\varphi(X_i)K_h(X_i-X_j)}{f^2(X_i)} \\ + \frac{1}{n}\sum_{i=1}^n \frac{\varphi(X_i)\left(f(X_i) - \widehat{f}_{n,i}(X_i)\right)^2}{f^2(X_i)\widehat{f}_{n,i}(X_i)}.\end{aligned}$$

The first term is an i.i.d. sample mean providing an unbiased estimate of $2\mathcal{I}(\varphi)$, while the second one is a $U$-statistic $U_n(\varphi)$ of degree two with kernel given by $H(x,x') = \varphi(x)K_h(x-x')/f^2(x)$ for $x,\ x'$ in $\mathbb{R}^d$ and that can be considered as a biased estimate of $-\mathcal{I}(\varphi)$. One may classically write the Hoeffding decomposition (*i.e.* Hajek projection) of $U_n(\varphi)$ [16]:

$$U_n(\varphi) = T_n(\varphi) + S_n(\varphi) + W_n(\varphi) - \mathbb{E}\left[U_n(\varphi)\right],$$

where $W_n(\varphi)$ is a degenerate $U$-statistic with zero mean and kernel $Q_n(x,x') = H(x,x') - \mathbb{E}[H(x,X)] - \mathbb{E}[H(X,x')] + \mathbb{E}[U_n(\varphi)]$ and

$$\begin{aligned} T_n(\varphi) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[H(X_i,X) \mid X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^n \frac{\varphi(X_i)}{f^2(X_i)}\left(K_h * f\right)(X_i), \\ S_n(\varphi) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[H(X,X_i) \mid X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^n \left(K_h * \frac{\varphi}{f}\right)(X_i), \end{aligned}$$

denoting by $X$ a random vector independent from the $X_i$'s, distributed according to $f(x)dx$. Observe incidentally that: $\forall n \geq 1$, $\forall \varphi \in \Phi$,

$$\begin{aligned}\mathbb{E}[U_n(\varphi)] &= \mathbb{E}\left[T_n(\varphi)\right] = \mathbb{E}[S_n(\varphi)] \\ &= \int_{x\in\mathbb{R}^d} \frac{\varphi(x)}{f(x)}(K_h * f)(x)dx.\end{aligned}$$

Hence, the deviation between the estimate (4) and the target integral (1) can be decomposed as the sum of four terms:

$$\widetilde{\mathcal{I}}_n(\varphi) - \mathcal{I}(\varphi) = M_n(\varphi) + W_n(\varphi) + B_h(\varphi) + R_n(\varphi),$$

where

$$\begin{aligned}B_h(\varphi) &= \mathcal{I}(\varphi) - \mathbb{E}[U_n(\varphi)] \\ &= \int \varphi(x)\left(1 - \frac{(K_h * f)(x)}{f(x)}\right)dx \quad (7)\end{aligned}$$

is a deterministic term vanishing as $h > 0$ tends to zero under adequate conditions (see Lemma 1),

$$M_n(\varphi) = \frac{2}{n}\sum_{i=1}^{n}\frac{\varphi(X_i)}{f(X_i)} - T_n(\varphi) - S_n(\varphi) - 2B_h(\varphi)$$

is a centered sum of i.i.d. random variables and

$$R_n(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\frac{\varphi(X_i)}{f^2(X_i)\widehat{f}_{n,i}(X_i)}\left(f(X_i) - \widehat{f}_{n,i}(X_i)\right)^2. \quad (8)$$

The proof consists in establishing bounds showing that each of these four terms is of order $o_{\mathbb{P}}(n^{-1/2})$ uniformly over $\Phi$. In contrast to the maximal deviations results used in general to investigate the accuracy of Empirical Risk Minimization in statistical learning (see *e.g.* [3]), one should pay attention to the fact that sharp inequalities (involving bounds for the maximal variance) are considered in the present analysis in order to deal properly with the dependence on $n$ (through the bandwidth $h_n$) of the classes of functions/kernels considered, more commonly needed in the asymptotic study of kernel density estimators, see *e.g.* [10]. Constants involved in the intermediary results below are not necessarily the same at each appearance.

**Bias.** As can be shown by examining the proof of the lemma below, a bound for the deterministic term (7) can be obtained using well-known approximation theoretic arguments under the smoothness hypotheses stipulated for the elements of class $\Phi$.

**Lemma 1.** *Under asssumptions* $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$ *and* $\mathbf{A_4}$, *we have the uniform bound:* $\forall h > 0$,

$$||B_h||_\Phi = \sup_{\varphi\in\Phi}|B_h(\varphi)| \leq \frac{C\mu_{Leb}(D_\Phi)M_\Phi}{\lambda}\cdot h^\beta,$$

*where* $C = \frac{L}{\lfloor\beta\rfloor!}\sum_{\alpha\in\mathbb{N}^d:\,|\alpha|=\lfloor\beta\rfloor}\int_{z\in\mathbb{R}^d}|K(z)|\prod_{i=1}^{d}|z_i|^{\alpha_i}dz$. Its technical proof is deferred to the Supplementary Material.

**Empirical process.** As shown in the Supplementary Material, the maximal deviation result related to the empirical process $\{M_n(\varphi)\}_{\varphi\in\Phi}$ stated below is proved by means of a specific version of an exponential inequality of [23].

**Lemma 2.** *Suppose that assumptions* $\mathbf{A_1}$-$\mathbf{A_5}$ *are fulfilled. For all* $\delta \in (0,1)$, *there exists* $n_\delta \geq 1$ *(defined in (B.2), in the supplementary file), depending on* $\delta$, $K$, $\lambda$ *and* $\Phi$ *only, such that for all* $n \geq n_\delta$, *with probability at least* $1-\delta$, *we have:*

$$||M_n||_\Phi \leq c_\Phi\frac{h^\beta}{\sqrt{n}}\sqrt{2\max\{\log(C_2/\delta)/C_3,\ C_1^2\log(2)\}},$$

*where* $c_\Phi$, $C_1$, $C_2$ *and* $C_3$ *are constants depending on* $\Phi$, $K$, $(\beta,\ L)$ *and* $\lambda$.

**Degenerate $U$-process.** Whereas concentration inequalities for degenerate $U$-processes (*i.e.* collections of $U$-statistics indexed by classes of functions) have been established in various articles such as [1] in the context of VC classes (see also [5] for more general results for instance), the major difficulty here arises from the fact that the class of kernels considered here depends on $n \geq 1$ (through the bandwidth $h_n$ namely). As shown in the Supplementary Material, the following bound can be proved by means of an exponential inequality for degenerate $U$-processes indexed by classes of kernels of VC type, involving a bound for the maximal variance, established in [17].

**Lemma 3.** *Suppose that assumptions* $\mathbf{A_2}$, $\mathbf{A_3}$ *and* $\mathbf{A}_4$ *are fulfilled. Then, for all* $\delta \in (0,1)$, *there exists* $\mathcal{C}_{\delta,1} \subset \mathbb{N}\times\mathbb{R}$ *(defined in (C.2), in the supplementary file) depending on* $\delta$, $K$, $\Phi$ *and* $\lambda$ *only, such that, for all* $(n,h) \in \mathcal{C}_{\delta,1}$, *with probability greater than* $1-\delta$, *we have:*

$$\begin{aligned}\|W_n\|_\Phi \leq\ & \frac{\gamma_\Phi}{(n-1)h^{d/2}}\times \\ & \max\left\{\log(C_2/\delta)/C_3,\ C_1\log\left(2G_\Phi/(\gamma_\Phi h^{d/2})\right)\right\},\end{aligned}$$

*where* $\gamma_\Phi$, $C_1$, $C_2$ *and* $C_3$ *are constants depending on* $\Phi$, $\lambda$ *and* $K$.

**Residuals.** We now turn to the residual term (8). The lemma below is established in the Supplementary Material. Its proof is based on the control of the probability that the $\widehat{f}_{n,i}(X_i)$'s get close to zero in particular.

**Lemma 4.** *Suppose that assumptions* $\mathbf{A_1}$-$\mathbf{A_5}$ *are fulfilled. For all* $\delta \in (0,1)$, *there exists* $\mathcal{C}_{\delta,2} \subset \mathbb{N}\times\mathbb{R}$ *(defined in (D.1) and (D.2), in the supplementary file)*

*depending on $\delta$, $K$, $\Phi$ and $\lambda$ only, such that, for all $(n,h) \in \mathcal{C}_{\delta,2}$, with probability at least $1-\delta$, we have:*

$$||R_n||_\Phi \leq \widetilde{\gamma}_\Phi \left( \frac{\max\left\{ \frac{\log\left(\frac{C_2}{\delta}\right)}{C_3},\ C_1 \log\left( \frac{2\|K\|_\infty}{c_{K,f} h^{d/2}} \right) \right\}}{nh^d} + h^{2\beta} \right),$$

*where $\widetilde{\gamma}_\Phi$, $c_{K,f}$, $C_1$, $C_2$ and $C_3$ are constants depending on $\Phi$, $K$, $(\beta,\ L)$ and $\lambda$.*

**Derivation of the stated bound.** The bound in Theorem 1 results from those stated in Lemmas 1-4 by taking $\mathcal{C}_\delta$ as the intersection of $\mathcal{C}_{\delta,1}$ (in Lemma 3), $\mathcal{C}_{\delta,2}$ (Lemma 4), $n \geq n_\delta$ (in Lemma 2), and also values of $(n, h \leq h_0)$ such that the bound for the bias in Lemma 1 (resp. for the residuals in Lemma 4) is larger than the bound for the empirical process term in Lemma 2 (resp. the $U$-process term in Lemma 3). □

# 4 APPLICATION TO SHIFT COVARIATE IN REGRESSION

We now present an application of the method analysed in the previous section in order to illustrate its performance in practice. After a brief presentation of the framework of covariate shift regression, a diagnostic tool for evaluating the quality of the prediction in a given covariate region is introduced (see *e.g.* [4]) and implemented by means of the method promoted based on toy data.

**Covariate shift regression.** Let $(x_i, y_i)_{i=1,\ldots,n}$ denote a training dataset of size $n \geq 1$ where, for each $i \in \{1,\ldots,n\}$, $y_i \in \mathcal{Y}$ stands for the output and $x_i \in \mathcal{X}$ is the covariate/input vector. The regression task consists in (i) learning a predictor $g$ from the training data in order to (ii) predict unobservable $y_{te}$ with $g(x_{te})$ for a so called test covariate $x_{te}$. Classical regression is concerned with a test covariate $x_{te} \in \mathcal{X}$ that is similarly distributed as the training covariates. In contrast, *covariate shift regression* considers situations where $x_{te} \in \mathcal{X}$ is not distributed in the same way as the training covariates. That is, when learning $g$, the main risk is to focus too much on regions containing the $x_i$'s but faraway from $x_{te}$. Under covariate shift and mispecification, it is known [20] that standard regression techniques such as maximum likelihood estimation does not provide accurate estimate. The most popular approach to the covariate shift regression problem is based on a re-weighting strategy (see [22], [14] and the references therein). Suppose for simplicity that the training dataset forms an i.i.d. sequence distributed according to $(Y,X)$. The conditional risk of the predictor $g$ given $X=x$ is denoted by $R(g|x) = \mathbb{E}[(Y-g)^2|X=x]$. The marginal distribution of $x_i$ is denoted by $f_X^{tr}$. If $f_X^{te}$ denotes the test distribution, *i.e.* the distribution of $x_{te}$, then the underlying risk can be expressed as $R_{te}(g) = \int R(g|x) f_X^{te}(x)dx$. A natural estimate of this risk is then given by

$$\widehat{R}_{te}(g) = n^{-1} \sum_{i=1}^n (y_i - g(x_i))^2 w_i, \tag{9}$$

with $w_i = f_X^{te}(x_i)/f_X^{tr}(x_i)$. As the weights $w_i$ are unknown in practice, one should estimate them based on the training sample and, when available, the test sample. The naive strategy (subject to the curse of dimensionality) is to estimate $f_X^{te}$ and $f_X^{tr}$ using kernel smoothing estimates and then to replace the unknown weights in (9) by the estimates. More sophisticated methods relying on the Kullback-Leibler divergence and on the least-squares distance are proposed in [22] and [14], respectively.

**Diagnostic tool for prediction quality in covariate regions.** When no test covariate $x_{te}$ is observed (making impossible an estimation of the importance weights $w_i$), an interesting issue is to know whether or not a given region in the covariate space $\mathcal{X}$ has a prediction of good quality. In the following, a region $(\mu,\Gamma)$ is represented by the Gaussian distribution with center $\mu \in \mathcal{X} \subset \mathbb{R}^p$ and a dispersion matrix $\Gamma \in \mathbb{R}^{p\times p}$. The risk related to the region $(\mu,\Gamma)$ is given by $R_{\mu,\Gamma}(g) = \int R(g|x)\phi_{\mu,\Gamma}(x)dx$, where $\phi_{\mu,\Gamma}$ stands for the density of $\mathcal{N}(\mu,\Gamma)$. The empirical "oracle" counter part (because it requires to know $f_X^{tr}$) is $\widehat{R}_{\mu,\Gamma}^{(or)}(g) = n^{-1}\sum_{i=1}^n (\phi_{\mu,\Gamma}(x_i)/f_X^{tr}(x_i))(y_i - g(x_i))^2$, and, the estimator based on the kernel smoothing approach is

$$\widehat{R}_{\mu,\Gamma}(g) = n^{-1} \sum_{i=1}^n \frac{\phi_{\mu,\Gamma}(x_i)}{\hat{f}_{n,i}(x_i)} (y_i - g(x_i))^2, \tag{10}$$

where $\widehat{f}_{n,i}$ is the leave-one-out estimator, defined in (5), associated to $x_i$. The estimation error associated to $\widehat{R}_{\mu,\Gamma}(g)$ has two component: one is related to the error between $g(x)$ and $\mathbb{E}[Y|X=x]$ and one associated to the noise $Y - \mathbb{E}[Y|X=x]$. Theorem 1 can be used to handle the first component in the error decomposition, i.e., the function $\varphi$ in Theorem 1 is taken equal to $x \mapsto \phi_{\mu,\Gamma}(x)(g(x) - \mathbb{E}[Y|X=x])^2$ which in many cases verifies each of our assumptions except the compact support assumption on $\varphi$ stated in $\mathbf{A_3}$ and (consequently) the lower bound assumption on $f_X$ stated in $\mathbf{A_4}$. This problem can be solved in practice by considering a trimming version (as proposed for instance in [11]), i.e., ignoring the terms with a too small $\hat{f}_{n,i}(x_i)$ in (10). Addressing these technicalities is beyond the scope of the paper.

**Ordinary least squares with misspecification.** To illustrate our proposal we consider a toy model according to which we generate the training dataset $(x_i, y_i)_{i=1,\ldots,n}$ for $n = 500$. It is given by $y_i = x_{i,1}^2 \mathbb{I}\{x_{i,1} > 0\} + \epsilon_i$, where $x_i = (x_{i,1}, \ldots, x_{i,p}) \sim \mathcal{N}((a, 0, 0, \ldots, 0)^T, I_p)$ , $\epsilon_i \sim \mathcal{N}(0, s^2)$, $s$ is chosen such that the signal-noise quotient is 0.5 and $p = 10$. The parameter $a$ is either $-1$ or 2 in order to highlight different situations. Estimation of $g$ is made through ordinary least squares, with $g(x_i) = \widehat{\alpha} + \widehat{\beta}^T x_i$, where $(\widehat{\alpha}, \widehat{\beta}) \in \operatorname{argmin}_{\alpha,\beta} \sum_{i=1}^n (y_i - \alpha - \beta^T x_i)^2$. To avoid the dimensionality curse, we consider regions associated to one specific covariate $x_{i,1}$, that is, the distribution $\phi_{\mu,\Gamma}$ is Gaussian. We set $\Gamma = 1/2$. We are interested in the performance of $\hat{R}_{\mu,\Gamma}^{(or)}(g)$ (ORACLE) and $\widehat{R}_{\mu,\Gamma}(g)$, when $\mu$ varies in in the range of $x_{i,1}$ estimating $R_{\mu,\Gamma}(g)$ (TRUE). For $\widehat{R}_{\mu,\Gamma}(g)$, $\hat{f}_{n,i}$ is either the classical kernel density estimator based on $x_{i,1}$ (KDE), or the leave-one-out estimator based on $x_{i,1}$(KDE-LOO). The parameter $h$ for the density estimate in (10) is picked via the "rule of thumb" in [21], giving $h = 1.06\widehat{\sigma}^2 n^{-1/5}$, where $\widehat{\sigma}^2$ is the empirical estimator of the variance of $x_{i,1}$, $i = 1, \ldots, n$. Fig. 1 provides an illustration for one particular dataset. The estimation accuracy (reflected by small values of $R_{\mu,\Gamma}(g)$) is not homogeneous. When $a = -1$, $g$ is not sharp in the right tail of $x_{i,1}$ whereas when $a = 2$, $g$ performs poorly in both the left and the right tails. For each value of $a$, KDE-LOO recovers this trend pretty well. Fig. 2 confirms that estimating $R_{\mu,\Gamma}(g)$ is more difficult when only few points $x_{i,1}$ are lying around $\mu$. Notice that KDE-LOO over performs KDE for any value of $\mu$. The ORACLE presents less bias, but a larger variance than KDE.

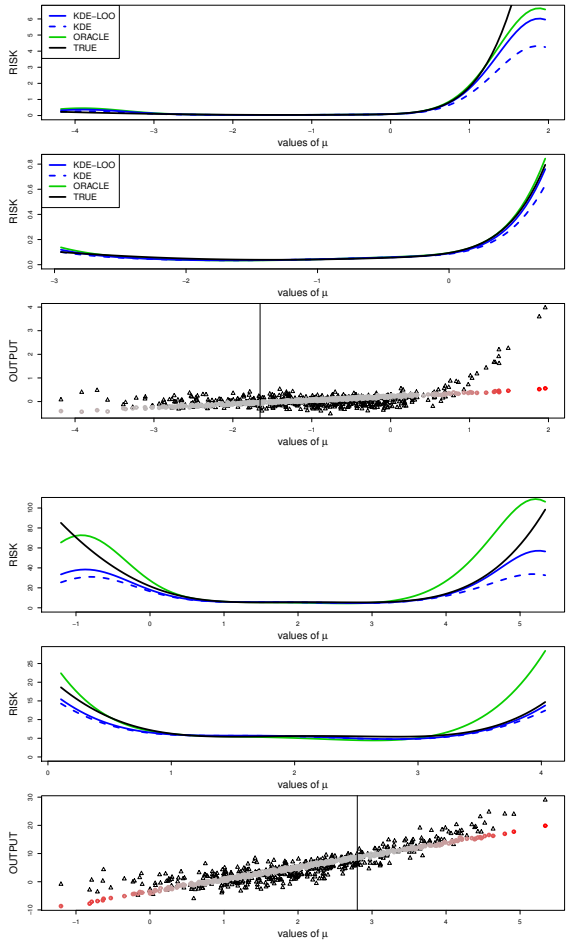


Figure 1: Top/middle : graph of $R_{\mu,\Gamma}(g)$ for TRUE, KDE-LOO, KDE and ORACLE, when $\mu$ lives in the whole range of $x_{i,1}$ (top) and zooming around the mean of $x_{i,1}$. Bottom : outputs $y_i$ and predicted values $g(x_{i,1})$ versus $x_{i,1}$. Red and grey colors reflects large and small values of KDE-LOO. In the right, $a = -1$. In the left $a = 2$. The signal-noise quotient is 0.5.

## 5 CONCLUSION

We provided a sound nonasymptotic analysis of the performance of a kernel smoothing integral estimation method that can be used as an alternative to the Monte-Carlo technique and compares favourably with it under certain assumptions. Precisely, though biased and involving highly dependent averaged components, the integral estimates thus produced achieve rate bounds that surpass those attained by traditional Monte Carlo methods (of order $O_{\mathbb{P}}(1/\sqrt{n})$) provided the instrumental density is sufficiently smooth and the kernel/bandwidth used are picked appropriately, uniformly over a class of functions of controlled complexity. The main tools exploited for establishing this striking result are an appropriate decomposition of the deviation between the target integrals and their estimates plus sharp concentration inequalities involving the variance of the functionals thus considered. Beyond theoretical results, a numerical example illustrates the practical performance of our method pro-

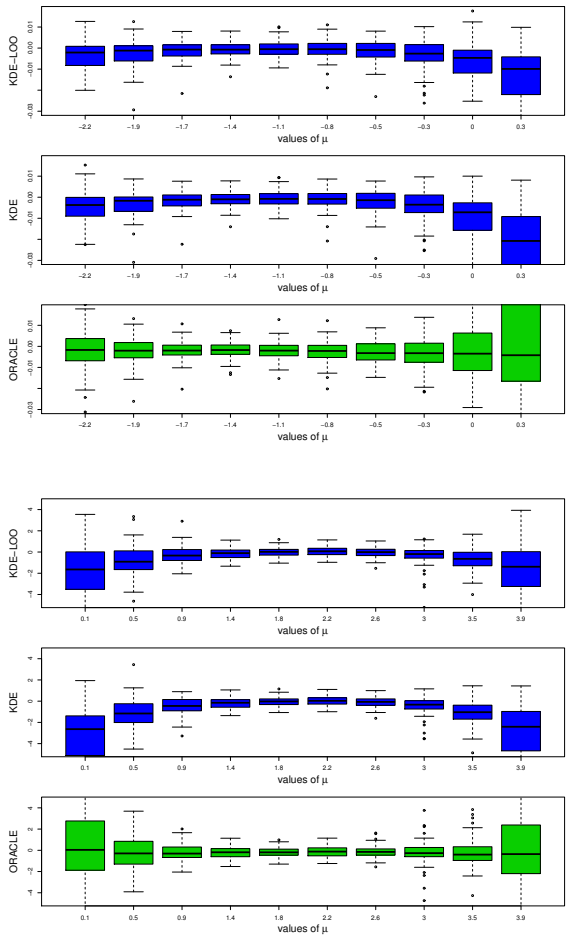


Figure 2: Boxplot (based on 100 replications) of the error for KDE-LOO, KDE and ORACLE when estimating $R_{\mu,\Gamma}(g)$ for different values of $\mu$ in the range of $x_{i,1}$. In the right, $a = -1$. In the left $a = 2$. The signal-noise quotient is 0.5.

moted here.

**Acknowledgements**

We thank the industrial Chair Machine Learning for Big Data of Télécom ParisTech for partly funding this research.

## References


[1] M. A. Arcones and E. Giné. U-processes indexed by VC classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters. *SPA*, 52:1738, 1994.

[2] R. Azaïs, B. Delyon, and F. Portier. Integral estimation based on Markovian design. *Submitted, available at https://arxiv.org/abs/1609.01165*, 2017.

[3] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[4] X. Chen, M. Monfort, A. Liu, and B. Ziebart. Robust covariate shift regression. In *Proceedings of AISTATS*, 2016.

[5] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *Ann. Statist.*, 36(2):844–874, 2008.

[6] P.J. Davis and P. Rabinovitz. *Methods of Numerical Integration*. Second edition. Dover, 2007.

[7] B. Delyon and F. Portier. Integral approximation by kernel smoothing. *Bernoulli*, 22(4):2177–2208, 2016.

[8] P. Doukhan. *Mixing. Properties and examples*. Springer-Verlag, New York, 1994.

[9] E. Giné and A. Guillou. On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. IHP.*, 37(4):503–522, 2001.

[10] E. Giné and H. Sang. Uniform asymptotics for kernel density estimators with variable bandwidths. *J. Nonparametr. Stat.*, 22(5-6):773–795, 2010.

[11] Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*, 84(408):986–995, 1989.

[12] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[13] M.H. Kalos and P.H. Whitlock. *Monte Carlo Methods*. Wiley-Blackwell, 2008.

[14] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.*, 10:1391–1445, 2009.

[15] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *The Annals of Statistics*, 34:2593–2706, 2006.

[16] A. J. Lee. *U-statistics: Theory and Practice*. CRC Press, 1990.

[17] P. Major. An estimate on the supremum of a nice class of stochastic integrals and $U$-statistics. *Probab. Theory Related Fields*, 134(3):489–537, 2006.

[18] C. McDiarmid. Concentration. In Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, volume 16 of *Algorithms and Combinatorics*, pages 195–248. Springer Berlin Heidelberg, 1998.

[19] C. Oates, M. Girolami, and N. Chopin. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

[20] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference*, 90(2):227–244, 2000.

[21] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman, 1986.

[22] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Ann. ISM.*, 60(4):699–746, 2008.

[23] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.

[24] A. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, New York, 1996.